B3.2 GEOMETRY OF SURFACES

Mathematical Institute, Oxford.
Prof. Alexander F. Ritter.

Comments and corrections are welcome: ritter@maths.ox.ac.uk

Contents

B3.2 Course policy: *It is essential that you read your notes after each lecture, otherwise you may feel lost. In the third and fourth year courses, the majority of courses will not cover all the material in lectures. You are expected to read the lecture notes to complement the lectures. Geometry of surfaces is a difficult and vast course, and I will do my best to make it digestible. But this will not happen by itself: it requires effort on your part, thinking on your own about the notes, the examples, the exercises.*

B3.2 Homework policy: *Homeworks are typically much harder than the exams: the aim of the homeworks is to make you better mathematicians, to stretch you and to inspire you. The homeworks are not designed to assess your basic understanding of the course (unlike exams). So homework marks are not aiming to predict your exam marks.*

Date: This version of the notes was created on September 19, 2018.

1. Examples

1.1 Four classes of surfaces

Our goal is to study and relate three classes of surfaces:

(1) Topological surfaces (*topological 2-manifolds*),
(2) Smooth surfaces (*smooth real 2-manifolds*),
   (a) embedded in $\mathbb{R}^3$,
   (b) abstractly (i.e. possibly without a choice of embedding into $\mathbb{R}^N$),
(3) Riemann surfaces (*complex 1-manifolds*).

We will postpone the precise definition to later. For now, the rough idea is that a surface locally looks like a 2-dimensional disc. Whether it looks like the disc continuously, smoothly or holomorphically distinguishes the cases (1), (2), and (3) respectively. The reason for studying (2a) before (2b) is that you already know what it means for functions on $\mathbb{R}^3$ to be smooth (infinitely differentiable), whereas in (2b) the definition is a little more difficult because you need to first define local smooth coordinates on the surface. Some surfaces are part of all four classes (such as a torus), others only of some, but all our surfaces belong to class (1).

Relation to future Part B and Part C courses.
You can study Riemann surfaces (and more generally algebraic curves) using tools from algebra, in **B3.3 Algebraic Curves**. The word *manifold* is the generalization of surface to higher dimensions, so n-manifold means your space locally looks like $\mathbb{R}^n$ (or rather, like a ball in $\mathbb{R}^n$). **C3.3 Differentiable Manifolds**, **C3.5 Lie Groups** (manifolds which are also groups) and **C7.5/7.6 General Relativity** all study manifolds from various perspectives. Complex manifolds locally look like $\mathbb{C}^n$, and you can study them (in greater generality) using tools from algebra in **C3.4 Algebraic Geometry** and in **C3.7 Elliptic Curves**, both of which build upon B3.3. The best tools to study topological manifolds (and more general topological spaces) come from topology and algebra, and this is done in **C3.1: Algebraic Topology**.

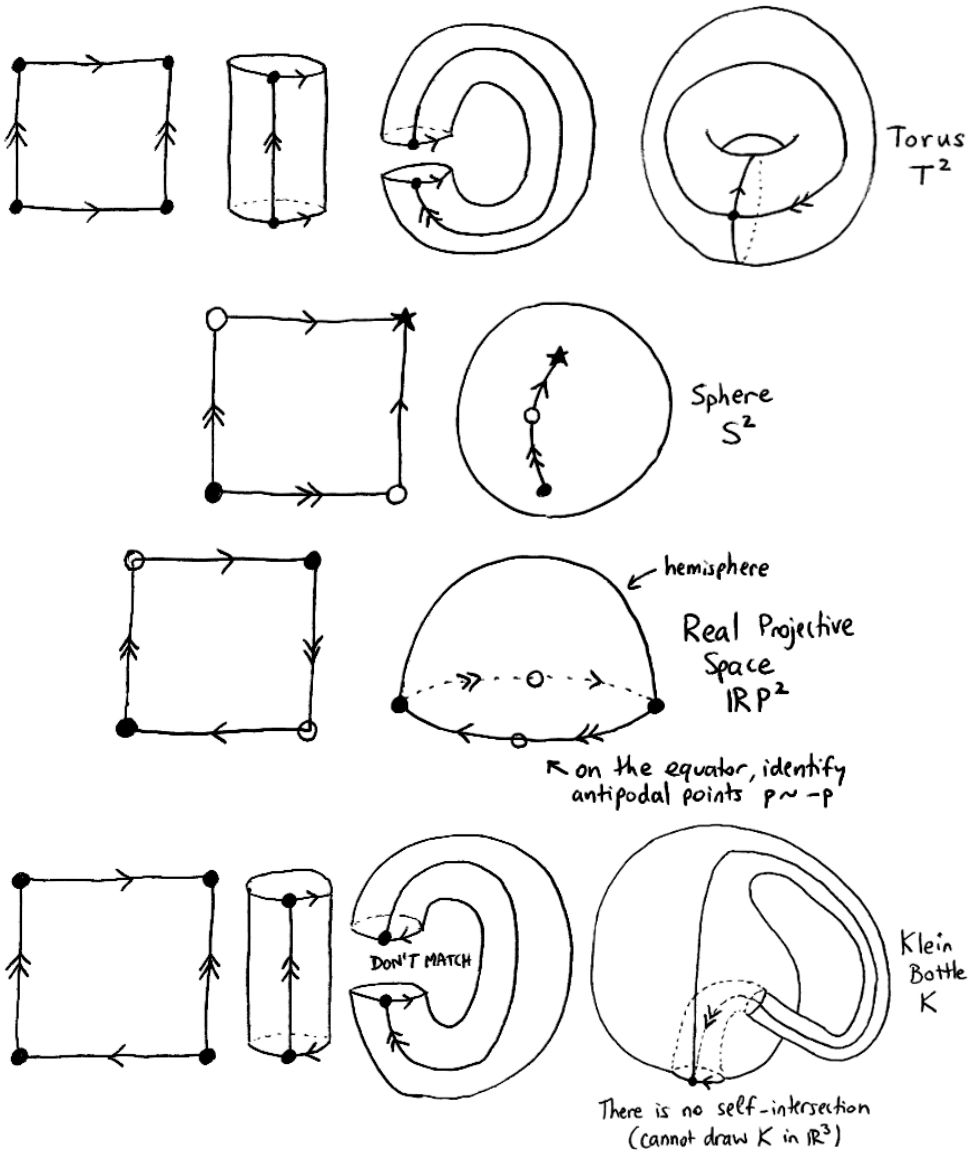
References for this course:
The notes from 2013 by Prof. Nigel Hitchin (see Course page online).
The notes from 1986 by Prof. Graeme Segal (see Course page online).
Manfredo P. do Carmo, *Differential Geometry of Curves and Surfaces*.
Pelham M. H. Wilson, *Curved Spaces*.
Several good references are suggested in the syllabus. In particular, the Course page online has a link to the notes by Prof. Richard Earl of a former second year course that partly overlaps with some of this course, called *Geometry: The Local Theory of Curves and Surfaces*.

Analysis and topology dictionary:
On the online course page you will also find the handout:
Analysis and Topology Dictionary – Handout
which summarises various useful terminology (e.g. topological space, Hausdorff, connected, compact, continuous, homeomorphism, smooth, diffeomorphism, holomorphic, etc.)

1.2 Examples in each of the three classes

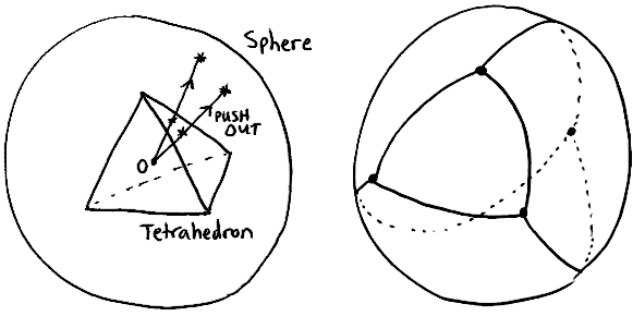
(1) **TOPOLOGICAL SURFACES:**

• Gluing edges of a square yields various surfaces:



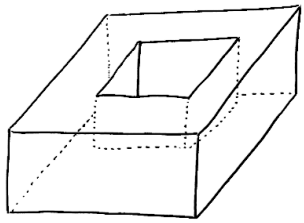
• A **cube** is a topological surface, and so are the other regular polyhedra (**tetrahedron, octahedron, icosahedron, dodecahedron**). But in fact, they are topologically the same (i.e. homeomorphic) to the sphere. Indeed, thinking of the polyhedra as sitting inside $\mathbb{R}^3$, pick a huge sphere which contains the polyhedron and then simply "continuously push" each point of the polyhedron radially outwards until the point reaches the sphere. This defines a homeomorphism (convince yourself of this!)[1]

[1]It is useful to get an acquaintance for spotting whether something is a homeomorphism or not, without the painstaking effort of writing down an explicit formula (if one is likely to make a mistake in spotting homeomorphisms, then one is probably also likely to write down an incorrect formula!). However, in this case



Remark. The image of the edges/vertices of the polyhedron divide the sphere into curved polygons (e.g. on the right we got a **triangulation** of the sphere).

• The **torus** (up to homeomorphism) is also a polyhedron (non-regular, non-convex):

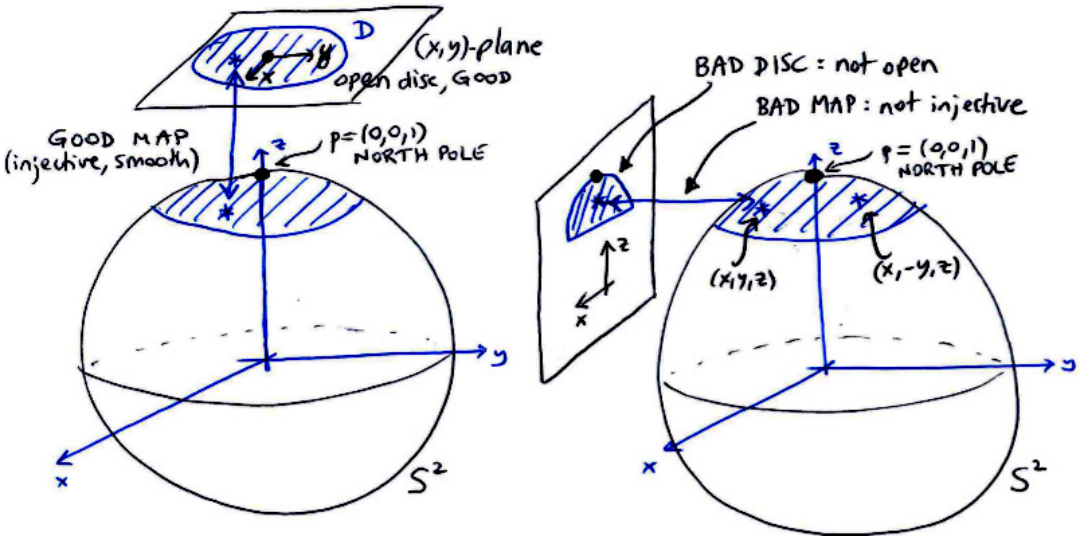


(2a) SMOOTH SURFACES IN $\mathbb{R}^3$:

• The $\mathbb{R}^2$ plane inside $\mathbb{R}^3$, so $\mathbb{R}^2 = \{(x, y, z) \in \mathbb{R}^3 : z = 0\}$.

• More generally a **plane** through $q \in \mathbb{R}^3$ with unit normal $n \in \mathbb{R}^3$:
$\{p \in \mathbb{R}^3 : (p - q) \cdot n = 0\} = \{(x, y, z) \in \mathbb{R}^3 : n_1x + n_2y + n_3z = n_1q_1 + n_2q_2 + n_3q_3\}$.

• The unit **sphere** in $\mathbb{R}^3$: $S^2 = \{p \in \mathbb{R}^3 : \|p\| = 1\} = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1\}$.



it is quite easy: if 0 lies on the inside of the polyhedron, then the map $x \mapsto \frac{x}{\|x\|}$ is an explicit map from the polyhedron to the unit sphere, one then easily checks that it is a continuous bijection, and finally one uses the general theorem that a continuous bijection from a compact space to a Hausdorff space is a homeomorphism.

Notice that near each point p you have two independent local smooth coordinates: at least two of the coordinates x, y, z will work. For example, near the North Pole $p = (0, 0, 1)$ you can use local coordinates x, y to uniquely describe nearby points since (x, y) : (neighbourhood of p) $\to$ (neighbourhood of 0) $\subset \mathbb{R}^2$ is a smooth homeomorphism. However near p you cannot use x, z as there are points of the sphere (x, y, z), $(x, -y, z)$, near the North Pole (so $z \approx 1$), which we cannot tell apart using just x, z (so they are no good as coordinates).

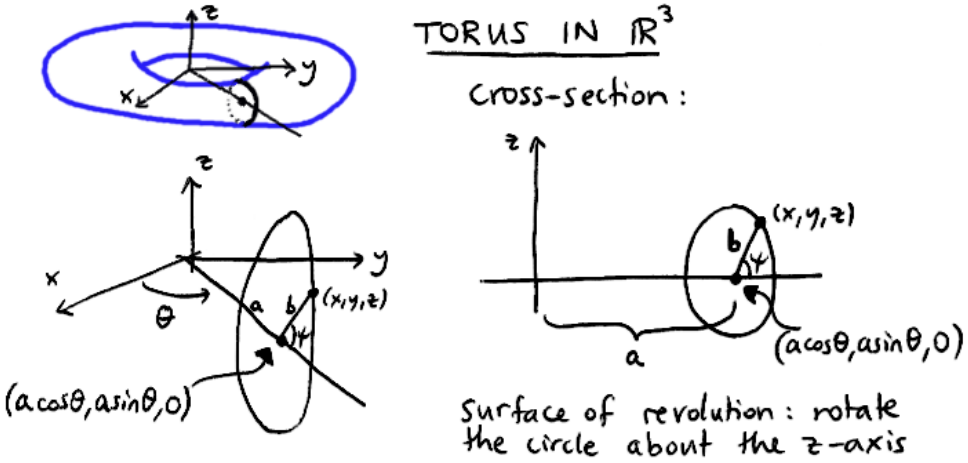
• Generalizing the above quadratic equation yields **ellipsoids** and **hyperboloids**:

$$\begin{aligned} \text{Ellipsoid:} \quad & \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 \\ \text{Hyperboloid with one sheet:} \quad & \frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1 \\ \text{Hyperboloid with two sheets:} \quad & \frac{x^2}{a^2} - \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1, \end{aligned}$$

where a, b, c are fixed constants. Many other surfaces defined by quadratic polynomials will reduce to one of these examples after changing coordinates (by diagonalising).

• A **torus** in $\mathbb{R}^3$: fix constants $a > b > 0$, then we can describe a torus by

$$T^2 \quad = \quad \{((a + b\cos\psi)\cos\theta,\ (a + b\cos\psi)\sin\theta,\ b\sin\psi) : \text{ all } \theta, \psi \in [0, 2\pi]\}$$



• Many more examples arise as **surfaces of revolution**, in which a curve in the (x, z)-plane gets rotated about the z-axis. For example, take:

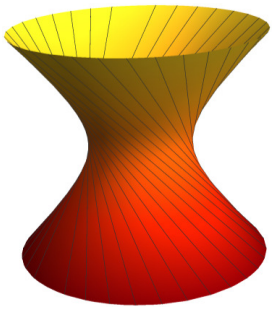
- ◇ Curve = vertical line, then rotating around the z-axis gives a **cylinder**.
- ◇ Curve = straight line which is neither vertical nor horizontal, then rotating around the z axis gives two opposite **cones** touching at a vertex. At the vertex, the surface is not smooth, so we need to remove the vertex.
- ◇ Curve = an ellipse $\frac{x^2}{a^2} + \frac{z^2}{b^2} = 1$, a parabola $z = x^2$, or a hyperbola $xz = 1$, then rotating gives an ellipsoid, paraboloid, hyperboloid. For example the **paraboloid** would be $\{(x, y, z) : z = x^2 + y^2\}$.

• **Ruled surfaces**: these are surfaces swept out my a moving straight line,

$$\{p(t) + s\, n(t) : s, t \in \mathbb{R}\}$$

so the straight line at time t is $p(t) + \mathbb{R}n(t)$ (a straight line through the point $p(t) \in \mathbb{R}^3$ which is parallel to the unit vector $n(t) \in \mathbb{R}^3$). However, some care is needed: not all

choices of $p(t), n(t)$ will give a smooth surface. For example, for $p(t) = (\cos t, \sin t, 0)$ and $n(t) = (\sin t, -\cos t, 1)/\sqrt{2}$, we get the one-sheeted hyperboloid $x^2 + y^2 - z^2 = 1$:



• **Surfaces cut out by one equation**:

$$\{(x, y, z) \in \mathbb{R}^3 : f(x, y, z) = 0\}.$$

Many, but not all, choices of smooth $f : \mathbb{R}^3 \to \mathbb{R}$ ensure this is a smooth surface.

(2b) ABSTRACT SURFACES:

The question "when is a topological surface smooth?" does not quite make sense. For example, the standard sphere $S^2 \subset \mathbb{R}^3$ is a smooth surface, but the cube $\subset \mathbb{R}^3$ is not smooth due to the corners, yet both are homeomorphic topological surfaces. The correct question is "when can we define a smooth structure on a given topological surface?".[1] We now explain how we can give a cube in $\mathbb{R}^3$ a "smooth structure". Near a vertex, the coordinates x, y, z do not vary smoothly.[2] However, we can define smooth local coordinates by composing the homeomorphism cube $\to S^2 \subset \mathbb{R}^3$ with smooth local coordinates for S^2. Such local coordinates near the vertex of the cube will not be smooth functions of the original x, y, z coordinates. Notice that what we have really done is define a smooth structure on the cube by requiring the homeomorphism cube $\to S^2 \subset \mathbb{R}^3$ to be smooth!

The reason the above may at first seem perplexing, is that $\mathbb{R}^3$ is causing unnecessary confusion: there is no reason for considering surfaces as already sitting smoothly inside $\mathbb{R}^3$. Indeed some smooth surfaces cannot be **embedded** inside $\mathbb{R}^3$ (here embedded roughly means[3] a smooth injective map). For example, the Klein bottle is smooth (locally it looks like a square, so we have smooth local x, y coordinates from the square). However, it cannot be embedded inside $\mathbb{R}^3$ (without self-intersections).

• The torus can be viewed as the quotient of $\mathbb{R}^2$ by the group of integral translations parallel to the two axes:

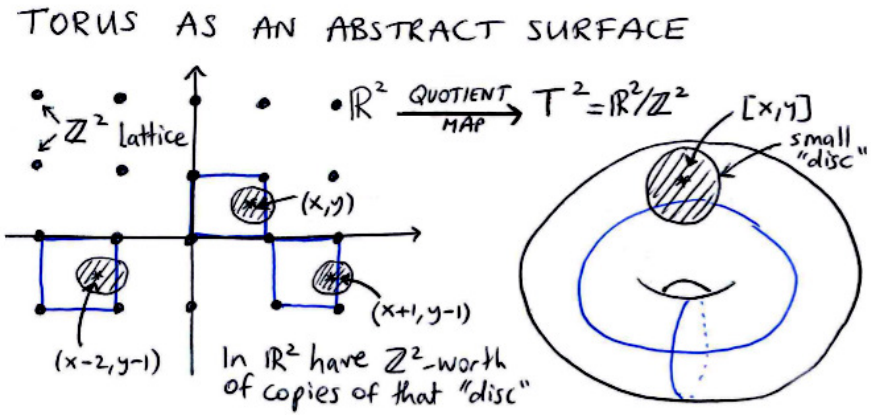
$$T^2 = \mathbb{R}^2/\mathbb{Z}^2 = \{[x, y] : (x, y) \in \mathbb{R}^2, [x, y] = [x + n, y + m] \text{ for all } n, m \in \mathbb{Z}\}.$$

---

[1] *Non-examinable:* The answer is, we can always endow a topological surface with the structure of an abstract smooth surface by choosing clever local coordinates. However, in higher dimensions (for manifolds) it can happen that a topological manifold does not admit a smooth structure. This is a very difficult problem (see "Relationship with topological manifolds" at http://en.wikipedia.org/wiki/Differentiable_manifold).

[2] convince yourself of this, using that each face is given by setting one of the coordinates to a constant.

[3] The precise definition of embedding is: a homeomorphism onto the image.



We define local coordinates near a point $[x_0, y_0] \in T^2$ by simply using the x, y coordinates of $\mathbb{R}^2$ near some pre-image point $(x_0, y_0) \in \mathbb{R}^2$ of $[x_0, y_0]$. There is a $\mathbb{Z}^2$-worth of choices of pre-image points, and any two such choices of local coordinates will differ by a smooth map (an integral translation).

Notice this torus is not sitting inside $\mathbb{R}^3$, and this construction of the torus is much simpler to work with than the above formula for the torus inside $\mathbb{R}^3$. For example, we have a natural notion of smooth function $f : T^2 \to \mathbb{R}$, namely it just[1] means a smooth function $\widetilde{f} : \mathbb{R}^2 \to \mathbb{R}$ which is translation-invariant under the above group, so

$$\widetilde{f}(x + n, y + m) = \widetilde{f}(x, y) \quad \text{for } n, m \in \mathbb{Z}.$$

• Thinking of surfaces as embedded inside $\mathbb{R}^3$ also makes it harder to notice which surfaces are actually the "same". For example, by cutting the torus, deforming, and regluing, we obtain the following **knotted torus**:

This knotted torus is smoothly homeomorphic to the original torus (indeed, think about how you would set up a smooth bijection between them). So the abstract surfaces are the "same", whereas the "knottedness" is extraneous information having to do with how we chose to embed the surface inside $\mathbb{R}^3$. It turns out for example, that you cannot continuously deform (inside $\mathbb{R}^3$) the torus into the knotted torus without creating self-intersections.[2]

• The effort of defining abstract surfaces is worth the trouble, and it is the modern viewpoint in geometry. The sinful secret of geometry is that any smooth n-manifold can[3] be smoothly embedded inside $\mathbb{R}^N$ for large enough N (indeed $N = 2n$ works). So one could in principle only study surfaces embedded in $\mathbb{R}^4$. For example, for the Klein bottle you can remove the self-intersection that you see in $\mathbb{R}^3$ by "lifting" one patch of the two intersecting sheets into

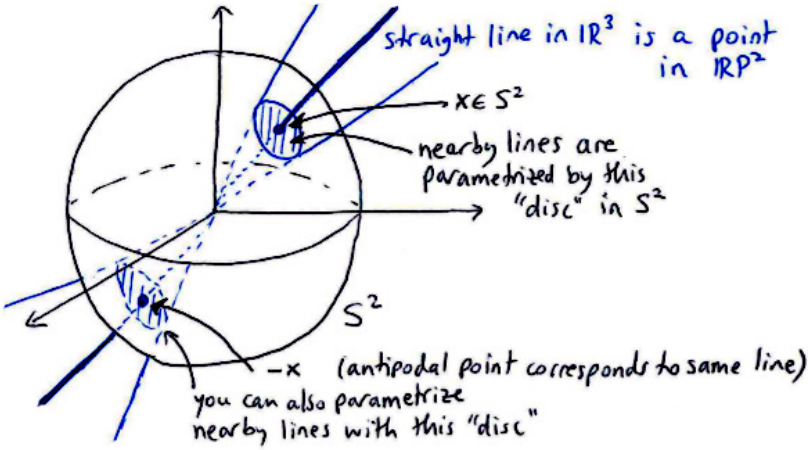
[1] Compare this with the 1-dimensional case: what does it mean to have a smooth function on the circle, $f : S^1 \to \mathbb{R}$? It just means a smooth function $\widetilde{f} : \mathbb{R} \to \mathbb{R}$ which is 1-periodic: $\widetilde{f}(x + 1) = \widetilde{f}(x)$ (or, 2π-periodic $\widetilde{f}(x+2\pi) = \widetilde{f}(x)$ if you think of the circle as parametrized by e^{it} for $t \in [0, 2\pi]$, instead of $e^{2\pi it}$ with $t \in [0, 1]$).

[2] This is a tricky exercise for you. If you need a hint, search for "trefoil knot".

[3] This is the **Whitney embedding theorem**, and is well beyond the scope of this course.

the fourth dimension.[1]

• **Real projective space** $\mathbb{R}P^2$ (which we already mentioned in (1) above). As a set, $\mathbb{R}P^2$ can be defined as the collection of all straight lines in $\mathbb{R}^3$ through the origin.



Such a straight line is determined by a non-zero point $(x, y, z) \in \mathbb{R}^3 \setminus \{0\}$, but rescaling such a point by any $\lambda \neq 0 \in \mathbb{R}$ will yield the same line. Thus $\mathbb{R}P^2$ arises as a quotient of $\mathbb{R}^3 \setminus \{0\}$, whose equivalence classes represent straight lines:

$$\mathbb{R}P^2 = \{[x, y, z] : (x, y, z) \in \mathbb{R}^3 \setminus \{0\}, [x, y, z] = [\lambda x, \lambda y, \lambda z] \text{ for any } \lambda \neq 0 \in \mathbb{R}\}.$$

These coordinates $[x, y, z]$ are called the **homogeneous coordinates** for $\mathbb{R}P^2$ (they are only defined up to rescaling all of them by a non-zero real number).

More explicitly, such a straight line is determined by the antipodal intersection points $\{x, -x\}$ in S^2. So we can locally parametrize this space by a disc, just like for S^2, since nearby straight lines are parametrized by points in S^2 close to $x \in S^2$ (the nearby lines form a double cone with vertex at the origin).

Therefore we can view $\mathbb{R}P^2$ as the quotient of S^2 by the group $\{\mathrm{Id}, A\}$ generated by the antipodal map $A : S^2 \to S^2$, $A(x) = -x$:

$$\mathbb{R}P^2 = \{x \in S^2\}/(x \sim -x).$$

Notice this also recovers the definition of $\mathbb{R}P^2$ in (1) above, since we only need the upper hemisphere to find a representative of each point.

(3) RIEMANN SURFACES:

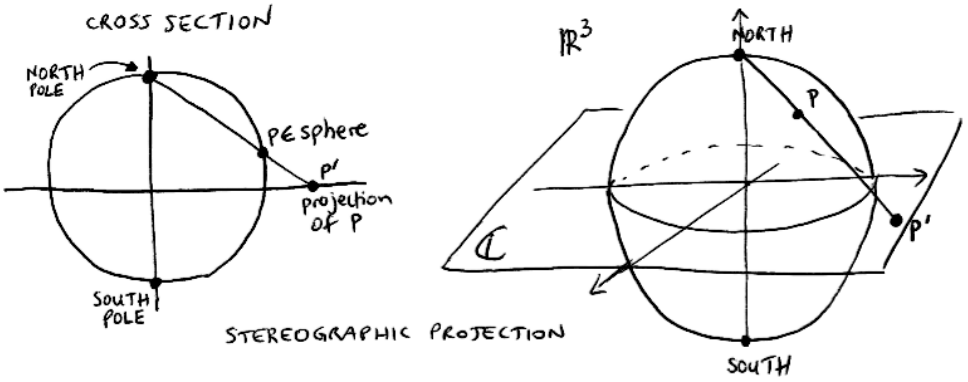
These are surfaces which "holomorphically" locally look like

$$D = \{z \in \mathbb{C} : |z| < 1\},$$

so locally there is a complex coordinate z. This is roughly the same as having two real coordinates x, y with a notion of "rotation by 90°" (multiplication by i), so that $z = x + iy$.

• **Complex projective space** $\mathbb{C}P^1$ is the sphere S^2 as a topological surface, we now define local holomorphic coordinates. One local complex coordinate z, defined everywhere except at the North Pole, is obtained by using the stereographic projection from the North Pole:

[1] As an analogy: for a figure 8 loop (two loops joined at a point) in the (x, y)-plane $z = 0$ in $\mathbb{R}^3$, at the crossing you have two lines intersecting: you can remove the self-intersection by slightly lifting vertically one of the two lines.



Explicitly, $S^2 \setminus (\text{North Pole})$ is identified with $\mathbb{C}$ via

$$S^2 \setminus (\text{North Pole}) \ni (X, Y, Z) \mapsto \frac{X}{1-Z} + i\frac{Y}{1-Z} = z \in \mathbb{C},$$

where $X^2 + Y^2 + Z^2 = 1$. We can also define a local complex coordinate w by taking the conjugate of the stereographic projection projecting from the South Pole.[1] So $S^2 \setminus (\text{South Pole})$ is identified with $\mathbb{C}$.
Notice we have identified $S^2 \setminus (\text{North} \cup \text{South})$ in two different ways with $\mathbb{C} \setminus \{0\}$, corresponding to the two coordinates z, w.

Exercise. Show that the two coordinates are related by

$$w = 1/z.$$

Notice that this **change of coordinates**, $\mathbb{C} \setminus 0 \to \mathbb{C} \setminus 0$, $z \mapsto \frac{1}{z}$ is a holomorphic map (i.e. complex differentiable). In fact, we will see that part of the definition of Riemann surfaces is that coordinate changes must be holomorphic (for smooth surfaces they must be smooth).

• A more general way to think of $\mathbb{C}P^1$, in analogy with $\mathbb{R}P^2$ above, is as the set of complex lines[2] through 0 in $\mathbb{C}^2$. Thus, again introducing **homogeneous coordinates**,

$$\mathbb{C}P^1 = \{[z_0 : z_1] : (z_0, z_1) \in \mathbb{C}^2 \setminus \{0\}, [z_0 : z_1] = [\lambda z_0 : \lambda z_1] \text{ for any } \lambda \neq 0 \in \mathbb{C}\}.$$

Then the region $z_0 \neq 0$ corresponds to $S^2 \setminus (\text{North Pole})$ above, and you can rescale so that $[z_0 : z_1] = [1 : z]$, so you obtain the above local coordinate

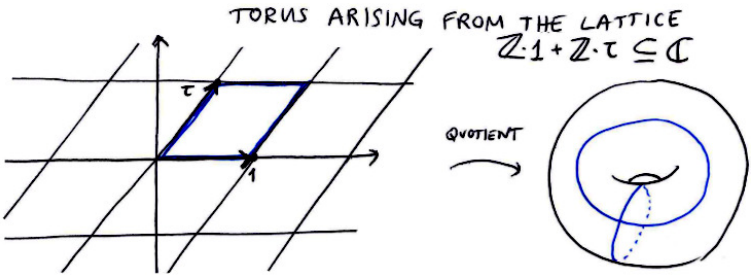
$$z = z_1/z_0$$

with $z = 0$ being the South Pole. Whereas $z_1 \neq 0$ corresponds to $S^2 \setminus (\text{South Pole})$, $[z_0 : z_1] = [w : 1]$, so you obtain the local coordinate $w = z_0/z_1$, and $w = 0$ is the North Pole. On the overlap $(z_0 \neq 0, z_1 \neq 0)$, $[1 : z] = [w : 1]$ recovers the above change of coordinates: $w = 1/z$. Since z parametrizes a copy of $\mathbb{C}$, you can think of $\mathbb{C}P^1$ as the compactification $\mathbb{C} \cup \{\infty\}$ where we add the extra point $\infty = [0 : 1] = (\text{North Pole})$.

• **Elliptic curves** (over $\mathbb{C}$) are the tori you get by quotienting $\mathbb{C}$ by a lattice. Above we used the lattice $\mathbb{Z}^2 \subset \mathbb{R}^2 \equiv \mathbb{C}$, but we could more generally use any $\mathbb{R}$-linearly independent vectors $\omega_1, \omega_2 \in \mathbb{R}^2$ and define the lattice $\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2 \subset \mathbb{R}^2$. By rescaling, and relabelling if necessary, we may as well assume that $\omega_1 = 1 \in \mathbb{C}$ and that $\omega_2 = \tau \in \mathbb{C}$ lies in the upper half-plane $\mathbb{H} = \{z \in \mathbb{C} : \text{Im}(z) > 0\}$:

[1] Exercise: $w = \frac{X}{1+Z} - i\frac{Y}{1+Z}$.

[2] A complex line through 0 in $\mathbb{C}^2$ is a complex vector subspace $V \subset \mathbb{C}^2$ of $\dim_{\mathbb{C}} V = 1$. So $V = \mathbb{C} \cdot (z_0, z_1) \subset \mathbb{C}^2$ for some $(z_0, z_1) \neq 0 \in \mathbb{C}^2$, and notice rescaling does not affect $V = \mathbb{C} \cdot (\lambda z_0, \lambda z_1)$ for any $\lambda \neq 0 \in \mathbb{C}$.



• **Surfaces cut out by one equation**:

$$\{(z,w) \in \mathbb{C}^2 : f(z,w) = 0\}.$$

Many, but not all, choices of a holomorphic function $f : \mathbb{C}^2 \to \mathbb{C}$ ensure this is a Riemann surface. Example: $f =$ complex polynomial in the two variables z, w. We will show that

$$\{(z,w) \in \mathbb{C}^2 : w^2 = 4(z-a)(z-b)(z-c)\},$$

for distinct constants $a, b, c \in \mathbb{C}$, is a torus with a point removed (there is a natural way to add the point $(z,w) = (\infty,\infty)$ to get the whole torus).

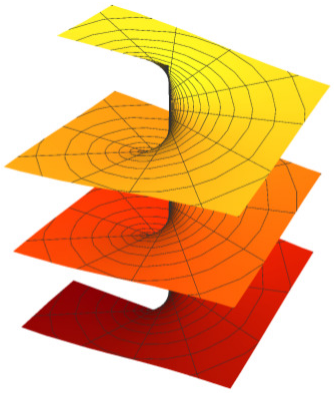
• Historically, Riemann surfaces first appeared in complex analysis when trying to deal with the problem of multi-valued functions. For example, the complex logarithm

$$\mathrm{Log}(z) = \log|z| + i\arg(z)$$

for $z \in \mathbb{C}\setminus 0$ has the problem that the argument is only defined up to adding multiples of $2\pi i$, since $e^{2\pi i n} = 1$ for $n \in \mathbb{Z}$. There are two ways of solving this problem: the ad-hoc approach is to make a cut in the complex plane, so we restrict to $\mathrm{Log} : \mathbb{C}\setminus(-\infty, 0] \to \mathbb{C}$ and artificially declare that $-\pi < \arg(z) < \pi$. This is called a **branch** of the Log "function".

Apart from the nuisance of making artificial choices, this has the problem that for a continuous curve such as the circle $\gamma(t) = e^{2\pi i t}$, the function $\mathrm{Log}(\gamma(t))$ is not continuous (it jumps from π to $-\pi$ at, or rather is not defined at, $t = 1/2$). Silly!

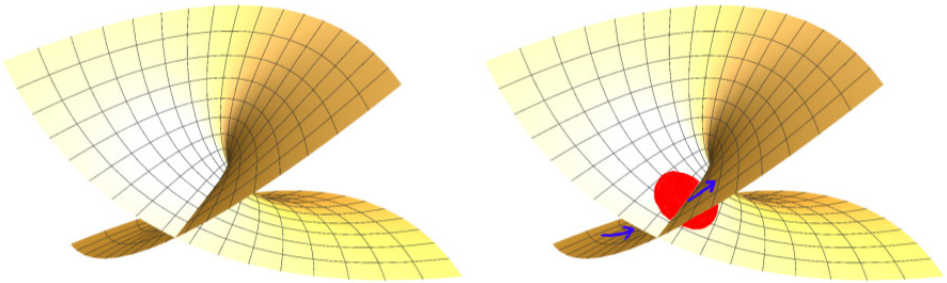
The natural remedy is really to consider all these "cut-domains" (for the various values of arg) as being glued[1] together according to $\arg(z)$-values to form a surface:



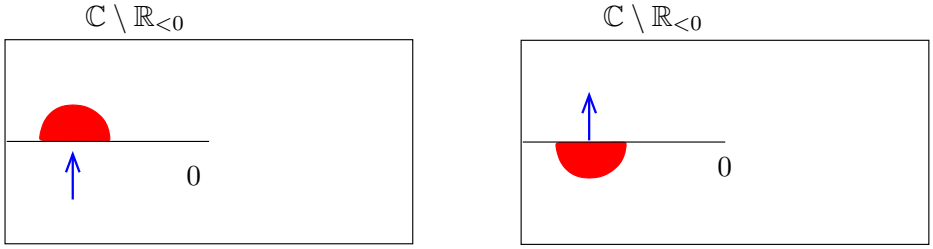
[1]In each cut-domain $\mathbb{C}\setminus(-\infty, 0]$, we re-insert *two* copies of $(-\infty, 0)$ along the two sides of the cut. We glue the cut-domains by identifying the copies of $(-\infty, 0)$ in pairs (so each cut-domain is glued onto two other cut-domains along the two copies of $(-\infty, 0)$). The point "0" does not get re-inserted, we remove it.

Locally the surface looks like $\mathbb{C}$, and in the picture the vertical axis keeps track of the value $\arg(z)$. Notice how the surface is made out of sheets (the cut-domains) and if you move towards the cut from above you go up the staircase to the next level in the sheets (indeed $\arg(z)$ increases), whereas if you approach the cut from underneath you move down in the staircase ($\arg(z)$ decreases). The various branches of $\mathrm{Log}(z)$ thus determine a well-defined complex logarithm function defined on the above Riemann surface (with the vertical axis coordinate telling you which $\arg(z)$ value to use).

• The Riemann surface obtained from a multi-valued holomorphic function by gluing together the "cut-domains" may not always be embeddable inside $\mathbb{R}^3$ (without self-intersections), the case of $\mathrm{Log}\, z$ above was rather special. Consider the square root $z^{1/2} = e^{\frac{1}{2}\mathrm{Log}\, z}$ (you get two distinct solutions $\pm w$ with $(\pm w)^2 = z$, for each $z \neq 0$).[1] Analogously we obtain the surface:



The self-intersection is an illusion caused by wanting to view it in $\mathbb{R}^3$. We can think of this surface S as being obtained by two cut-domains $\mathbb{C} \setminus \mathbb{R}_{<0}$, pictured below, which are glued along the cut.[2] The bottom-cut of one cut-domain is identified with the top-cut of the other domain. The two shaded half-discs glue together to form a disc in S. Walking in the direction of the arrow in the left cut-domain will make us pop out where the arrow points in right cut-domain (and this short walk does not intersect the shaded disc in the actual surface).



Consider the holomorphic map $\varphi : S \to \mathbb{C}$, $w \mapsto w^2 = z$. The preimage of a small disc in $\mathbb{C}$ centred at $z \neq 0$ consists of two disjoint small discs in S centred at $\pm\sqrt{z}$. The exception is when $z = 0$, in that case the preimage of a small disc is just one disc in S: try drawing it in inside the $\mathbb{R}^3$-picture above. What does this correspond to in the two cut-domains?

An equivalent way to describe S is as the "graph" $S = \{(z, w) \in \mathbb{C}^2 : z - w^2 = 0\}$ of the square root function. In this case, $S \to \mathbb{C}$, $(z, w) \mapsto w$ is the square root function, and $(z, w) \mapsto z$ is φ. We can use either z or w as a local holomorphic coordinate for S near (z, w)

[1] More explicitly: the two branches of the square root are $re^{i\theta} \mapsto \sqrt{r}e^{i\theta/2}$ and $re^{i\theta} = re^{2\pi i+i\theta} \mapsto re^{(2\pi i+i\theta)/2} = \sqrt{r}e^{i\pi+i\theta/2} = -\sqrt{r}e^{i\theta/2}$. The surface is like an Escher staircase: if you go up two flights of stairs (i.e. θ increases by 4π) then we will be back to where we started.

[2] To clarify: each cut $\mathbb{R}_{<0} = (-\infty, 0)$ gets replaced by *two* copies of $\mathbb{R}_{<0}$ that we re-insert onto $\mathbb{C}\setminus\mathbb{R}_{<0}$. The new boundary of the cut-domain consists of two half-lines that intersect at 0, call them *bottom-cut*, *top-cut*. We modify the topology near the cuts: the shaded disc above, obtained by gluing two half-discs, is a typical open neighbourhood of a point of one of the two copies of $\mathbb{R}_{<0}$. What does a neighbourhood of 0 look like?

when $z \neq 0$, but near $(0,0)$ we must use w as our local coordinate, not z. Can you see why?

• In general, given a holomorphic function defined on some region of $\mathbb{C}$, the aim is to build a Riemann surface by **analytic continuation**. That is, you patch together local Taylor series for the holomorphic function, and you try to build the largest possible surface (which locally looks like $\mathbb{C}$) on which the function can be (uniquely) extended to. For example, the **Riemann hypothesis** is a conjecture about the zeros of a function, the Riemann zeta function, which is constructed by analytic continuation.

1.3 Non-examples: spaces which are not surfaces

• The disc $D = \{z \in \mathbb{C} : |z| < 1\}$ together with a line segment: $D \cup [1,2)$ is not a surface: at points of the line segment $[1,2)$ the surface is not locally homeomorphic to a disc. Such non-examples are easy to spot since surfaces have to be "locally 2-dimensional".

• The **double cone** (two cones sharing the vertex) with angle θ,

$$\{(x,y,z) \in \mathbb{R}^3 : z^2 \tan^2\theta = x^2 + y^2\},$$

is not a surface (not even topological): at the vertex it is not locally homeomorphic to a disc.[1]

• The closed disc

$$\mathbb{D} = \{z \in \mathbb{C} : |z| \leq 1\}$$

is not a surface (not even topological): at any boundary point, it is locally homeomorphic to a half disc

$$D^+ = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 < 1, y \geq 0\}.$$

One could define **surfaces with boundary**, by requiring that the surface is locally homeomorphic to D^+ at points of the boundary. Then the closed disc $\mathbb{D}$ would be a surface (indeed Riemann surface) with boundary. In this course, we will not study surfaces with boundary.

• The **plane with two origins** is obtained as the quotient of two copies of $\mathbb{R}^2$:

$$(\mathbb{R}^2 \times \{1\}) \sqcup (\mathbb{R}^2 \times \{2\})/((x,y,1) \sim (x,y,2) : \text{ for all } (x,y) \text{ except } (0,0)).$$

This space is not Hausdorff: any open set around $(0,0,1)$ will intersect any open set around $(0,0,2)$, so you cannot separate the two origins $(0,0,1) \neq (0,0,2)$. Locally the space is nevertheless homeomorphic to a disc (for example near $(0,0,1)$ it looks like $D \times \{1\}$).

By convention, we prohibit surfaces from being non-Hausdorff. One reason is we want limits to be unique: $(\frac{1}{n}, 0, 1) \sim (\frac{1}{n}, 0, 2)$ converges to two distinct points: $(0,0,1) \neq (0,0,2)$. Physically, it would be unrealistic to have a common path $(t-1,0,1) \sim (t-1,0,2)$ of a particle at time $t \in [0,1]$ which at time $t = 1$ can be in two different places.

• Quotient spaces can often be non-Hausdorff. Recall that a quotient space is Hausdorff if and only if the equivalence relation $\{(x,x') \in X \times X : x \sim x'\} \subset X \times X$ is a closed set.

2. Definition of surface

2.1 Topological surfaces

Definition 2.1. *A **topological surface** is a Hausdorff topological space S such that each point $p \in S$ has a neighbourhood homeomorphic to an open subset of $\mathbb{R}^2$.*

Remark 2.2 (Locally you are a disc)**.** *If the above homeomorphism is*

$$f : (\text{neighbourhood } U \subset S \text{ of } p) \to V = f(U) \subset \mathbb{R}^2$$

[1] Exercise. Check this. *Hint. what happens to connectedness properties if you remove the vertex?*

then we can shrink U by replacing it with[1] $\widetilde{U} = f^{-1}(D_r(f(p))$. *Then compose f with the map $\mathbb{R}^2 \to \mathbb{R}^2, x \mapsto \frac{1}{r}(x - f(p))$ which rescales and translates $D_r(f(p))$ to our favourite unit disc $D = D_1(0)$. The new map $\widetilde{f} : \widetilde{U} \to \mathbb{R}^2$ shows that S is locally homeomorphic to D near p.*

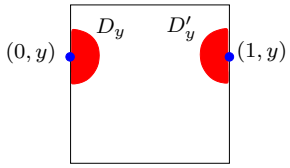
Example. The torus as a topological surface. Recall the quotient of the square:

$$T^2 = [0,1] \times [0,1] \,/\, ((0,y) \sim (1,y) \text{ and } (x,0) \sim (x,1) \text{ for all } x, y \in [0,1])$$

For an interior point $(x,y) \in (0,1) \times (0,1)$ of the square, we can simply pick a small disc around it also lying in the interior, then the homeomorphism f is just the identity. Slightly harder, for a point $(0,y)$ on the left edge of the square, with $y \neq 0$, consider the half-disc

$$D_y = \{(X,Y) \in [0,1]^2 : X^2 + (Y-y)^2 < \varepsilon\}$$

with centre $(0,y)$, radius $\varepsilon < \min\{y, 1-y\}$.



Notice $D'_y = \{(1-X,Y) \in [0,1]^2 : (X,Y) \in D_y\}$ is a half-disc with centre $(1,y)$, radius ε. The two half-discs glue together in the quotient along the common boundary edge $(0,Y) \sim (1,Y)$ to form a disc $D_y \cup D'_y \subset T^2$. Explicitly, we get a homeomorphism

$$f : \big(U = D_y \cup D'_y \subset T^2\big) \to \big(V = \{(X,Y) \in \mathbb{R}^2 : X^2 + (Y-y)^2 < \varepsilon\} \subset \mathbb{R}^2\big)$$

with $f(X,Y) = (X,Y)$ on D_y, and $f(X,Y) = (X-1,Y)$ on D'_y. *Exercise: run a similar argument for the vertex $(0,0)$ of the square (glue four quarter-discs).* I hope this example convinces you that (often) writing tedious formulas does not make an argument more rigorous than drawing pictures.

Remark 2.3 (Topological manifolds)**.** *An n-**manifold** is a Hausdorff topological space S such that each point p has a neighbourhood homeomorphic to an open subset of $\mathbb{R}^n$. As above, one can always find a local homeomorphism onto the ball $B_1(0) = \{z \in \mathbb{R}^n : \|z\| < 1\}$.*

Remark 2.4 (Why not metric spaces?)**.** *Most topological surfaces arise as metric spaces.*[2] *So it's easy to describe the topology: each open set is a union of open balls. So why not start off with a metric space? The metric is extra data which we do not care about: we think of two topological surfaces as being the same if there is a homeomorphism between them, but these rarely ever preserve distances. One could study topological surfaces together with a choice of metric, and require homeomorphisms to be **isometries** (so distance-preserving).*

2.2 Local coordinates and frames of reference

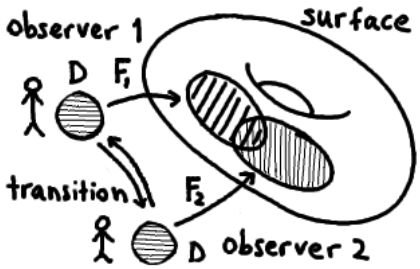
The homeomorphism f above, defined near p, determines **continuous local coordinates** on the surface near p, by declaring $\boxed{q \in S \text{ near } p \text{ has local coordinates } (x,y) = f(q) \in \mathbb{R}^2.}$ The above f is called a **chart**, and the inverse $F = f^{-1}$ is called a **local parametrization**,

$$F = f^{-1} : (V \subset \mathbb{R}^2) \to S.$$

Physicists like to call this a **frame of reference**.

[1] choose a small radius $r > 0$ for the open disc $D_r(f(p)) = \{q \in \mathbb{R}^2 : \|q - f(p)\| < r\}$ so that $D_r(f(p)) \subset V$.

[2] *Non-examinable*: one usually requires surfaces and manifolds to be **second-countable**, and this implies (by Uryshon's metrization theorem) that the topology of a surface or a manifold is always induced by some metric. *Second-countable* means that the topology has a countable base (i.e. there is a countable family of open sets U_i, such that every open set is a union of some subfamily of such U_i).



When observing a particle moving on a surface S you describe it in terms of some coordinates $(x(t), y(t))$ depending on time t. It is important that two observers, using different frames of reference, agree on whether or not the particle is moving continuously. Let's compare two frames of reference F_1, F_2 on the overlap of their images. The particle's local coordinates are $(x(t), y(t))$ and $(\widetilde{x}(t), \widetilde{y}(t))$ for the two observers. They observe the same particle, so

$$F_1(x(t), y(t)) = F_2(\widetilde{x}(t), \widetilde{y}(t)) = (\text{position of particle in } S \text{ at time } t).$$

Therefore, the **change of coordinates** from observer 1 to observer 2 is:

$$(\widetilde{x}(t), \widetilde{y}(t)) = (F_2^{-1} \circ F_1)(x(t), y(t)).$$

Since F_1, F_2 are homeomorphisms, the **transition map** $F_2^{-1} \circ F_1$ is continuous wherever it is defined. Thus:

Corollary 2.5. *For a topological surface, the transition maps (changes of coordinates between two local parametrizations) are always continuous.*

2.3 Smooth surfaces in $\mathbb{R}^3$

Definition 2.6 (Smooth maps)**.** *Given two open subsets $U \subset \mathbb{R}^n$, $V \subset \mathbb{R}^m$, a map $f : U \to V$ is **smooth** if it is infinitely differentiable.*[1]
*Given two arbitrary subsets $X \subset \mathbb{R}^n$, $Y \subset \mathbb{R}^m$, a map $f : X \to Y$ is **smooth** if locally*[2] *it is the restriction of a smooth function $\mathbb{R}^n \to \mathbb{R}^m$.*

Definition 2.7 (Diffeomorphism)**.** *Given two arbitrary subsets $X \subset \mathbb{R}^n$, $Y \subset \mathbb{R}^m$, a map $f : X \to Y$ is called **diffeomorphism** if f is a homeomorphism and f, f^{-1} are both smooth.*

Definition 2.8. *A **smooth surface in** $\mathbb{R}^3$ is a subset $S \subset \mathbb{R}^3$ such that each point $p \in S$ has a neighbourhood diffeomorphic to an open subset of $\mathbb{R}^2$.*

Remarks.

- ◇ As before, we can always arrange to have diffeomorphisms $f : U \to D$ to the disc.
- ◇ Smooth surfaces in $\mathbb{R}^3$ are also topological surfaces, because diffeomorphisms are homeomorphisms, and subspaces of a Hausdorff space such as $\mathbb{R}^3$ are Hausdorff.
- ◇ A local diffeomorphism $f : S \to \mathbb{R}^2$ defined near p determines **smooth local coordinates**: a point q near p has coordinates $f(q) = (x, y) \in \mathbb{R}^2$.
- ◇ The inverse $F = f^{-1}$, which maps some open set of $\mathbb{R}^2$ to some open neighbourhood of $p \in S$, is called a **local parametrization near** p.
- ◇ To define **smooth n-manifolds in** $\mathbb{R}^m$ you simply replace $\mathbb{R}^2$ by $\mathbb{R}^n$ above.

[1] Meaning: all partial derivatives of f of all orders exist (it follows that all partial derivatives are continuous, so this definition is the same as requiring that f has derivative maps of all orders).

[2] Explicitly: for each $p \in X$ there is an open neighbourhood U around p and a smooth map $F : U \to \mathbb{R}^m$, such that $F = f$ on $U \cap X$. We need to extend f to an open set, because in order to take the limit $\partial_{x_i} f(p) = \lim_{t \to 0} \frac{1}{t}(f(p + te_i) - f(p))$ we need f to be defined along the ray $p + te_i$ for small t, and for all we know this ray may not belong to the given set X.

Corollary 2.9. *For a smooth surface in $\mathbb{R}^3$, the transition maps (changes of coordinates between two local parametrizations) are always smooth.*

Proof. As in Section 2.2, the transition map $F_2^{-1} \circ F_1$ is defined between two open sets of $\mathbb{R}^2$. Since F_1, F_2^{-1} are diffeomorphisms, $F_2^{-1} \circ F_1$ is a diffeomorphism (by the chain rule). □

It is easy to check whether a map $\mathbb{R}^2 \to S$ is smooth: you view it as a map $\mathbb{R}^2 \to \mathbb{R}^3$ and check that it is infinitely differentiable. But checking whether a map $S \to \mathbb{R}^2$ is smooth is a nuisance, because by the above definition you would need to first extend the map to a neighbourhood of S, at least locally. We will later prove (by the implicit function theorem) that this nuisance is easily avoided using linear algebra:

Theorem 2.10. *A smooth injective map $F : V \to S$, defined on an open subset $V \subset \mathbb{R}^2$, is a smooth local parametrization $\iff$ $\partial_x F, \partial_y F$ are linearly independent at each point of V.*

We will do one example explicitly, but I hope you agree that we do not want to carry out such calculations every time we encounter a smooth surface. Your time is better spent at developing the ability to spot instinctively whether or not a surface may fail to be smooth.

Example. The torus as a smooth surface in $\mathbb{R}^3$. Recall:

$$T^2 = \{((a + b\cos\psi)\cos\theta,\ (a + b\cos\psi)\sin\theta,\ b\sin\psi) \in \mathbb{R}^3 :\ \text{all } \theta, \psi \in [0, 2\pi]\}.$$

Let's check this is a smooth surface near the point $p = (a + b, 0, 0)$ (taking $\theta = \psi = 0$). Perhaps unsurprisingly, we will try the local parametrization

$$F : \mathbb{R}^2 \supset (-\pi, \pi) \times (-\pi, \pi) \to T^2, (x, y) \mapsto ((a + b\cos y)\cos x,\ (a + b\cos y)\sin x,\ b\sin y).$$

This is manifestly smooth (since $\cos, \sin$ are smooth). It is not so hard to check that it is injective (check this, using that $a > b > 0$). So, by the Theorem, we reduce to linear algebra: we need $\partial_x F$, $\partial_y F$ to be linearly independent in $\mathbb{R}^3$.
So we need to find a non-zero 2×2 subdeterminant of the matrix

$$\begin{pmatrix} \partial_x F & \partial_y F \end{pmatrix} = \begin{pmatrix} -(a + b\cos y)\sin x & b\sin y\cos x \\ (a + b\cos y)\cos x & -b\sin y\sin x \\ 0 & b\cos y \end{pmatrix}.$$

The bottom two rows give subdeterminant $(a + b\cos y)b\cos x\cos y$. Since $a > b$, the first bracket is non-zero, so this subdeterminant is non-zero except when $x, y \in \{\pm\pi/2\}$. But in that case, the top two rows give subdeterminant $(\textit{non-zero}) \cdot \sin^2 x$ with $\sin^2 x = 1$, again non-zero.

What does it mean to have a **smooth map** $f : S_1 \to S_2$ between smooth surfaces in $\mathbb{R}^3$? By Definition 2.6 it means f is a continuous map such that locally near each point of $p \in S_1$, f can be extended to a smooth function $\mathbb{R}^3 \to \mathbb{R}^3$ defined in a neighbourhood of $p \in S_1 \subset \mathbb{R}^3$.

2.4 Abstract smooth surfaces

Definition 2.11 (Abstract smooth surfaces)**.** *A **smooth surface** is a Hausdorff topological space S, together with a family of homeomorphisms, called **local parametrizations**,*

$$F_i : (\textit{open subset } V_i \subset \mathbb{R}^2) \to (\textit{open subset } U_i \subset S),$$

*such that the U_i cover S, so $S = \cup U_i$, and on overlaps the **transition maps** are smooth:*

$$F_j^{-1} \circ F_i : \mathbb{R}^2 \to \mathbb{R}^2$$

is smooth wherever defined.[1]

[1] Explicitly, $F_j^{-1} \circ F_i$ is defined on $F_i^{-1}(U_i \cap U_j) \subset V_i \subset \mathbb{R}^2$.

$\diamond$ Notice that S is automatically a topological surface.
$\diamond$ As usual, each F_i determines **smooth local coordinates**: $q \in S$ near p has local coordinates $(x, y) = F_i^{-1}(q) \in \mathbb{R}^2$ in the parametrization F_i.

Remark 2.12 (Smooth manifolds)**.** *To define* ***smooth*** n***-manifolds****, replace* $\mathbb{R}^2$ *by* $\mathbb{R}^n$ *above.*

Example. The torus as an abstract smooth surface. The quotient $T^2 = \mathbb{R}^2/\mathbb{Z}^2$ is a topological space. It is Hausdorff as the equivalence relation $\{((x, y), (x + n, y + m)) : (n, m) \in \mathbb{Z}^2, (x, y) \in \mathbb{R}^2\} \subset \mathbb{R}^2 \times \mathbb{R}^2$ is a closed subset. For any point $p \in T^2$, pick a representative point $\widetilde{p} = (x, y) \in \mathbb{R}^2$, meaning $p = [x, y]$. Consider

$$D_{\widetilde{p}} = \{(X, Y) \in \mathbb{R}^2 : (X - x)^2 + (Y - y)^2 < \varepsilon\},$$

the disc with centre (x, y) and radius $\varepsilon = 1/100$ (overkill: $\varepsilon \leq 1/2$ would work). Notice that no two points in $D_{\widetilde{p}}$ differ by $\mathbb{Z}^2$, therefore the quotient map

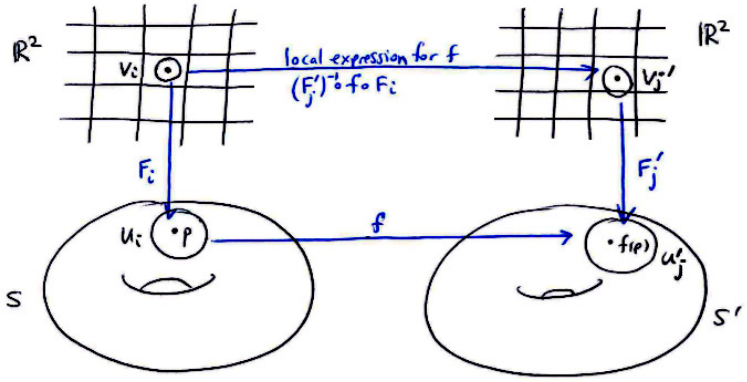
$$F_{\widetilde{p}} : \left(V_{\widetilde{p}} = D_{\widetilde{p}} \subset \mathbb{R}^2\right) \to \left(U_{\widetilde{p}} = \{[X, Y] \in T^2 : (X, Y) \in D_{\widetilde{p}}\} \subset T^2\right), \, F_{\widetilde{p}}(X, Y) = [X, Y]$$

is a homeomorphism. Since $p \in U_{\widetilde{p}}$, we obviously get $T^2 = \cup U_{\widetilde{p}}$ (taking the union over all choices of $\widetilde{p} \in \mathbb{R}^2$ for all $p \in T^2$). Now consider transition maps. Suppose $w \in U_{\widetilde{p}} \cap U_{\widetilde{q}} \subset T^2$ is in an overlap. Say w has local coordinates $(X, Y) \in V_{\widetilde{p}}$ and $(X + n, Y + m) \in V_{\widetilde{q}}$ respectively. By definition of T^2 the n, m are integers. The transition map is smooth since it is the translation

$$F_{\widetilde{q}}^{-1} \circ F_{\widetilde{p}} : \mathbb{R}^2 \to \mathbb{R}^2, (X, Y) \mapsto (X + n, Y + m).$$

What does it mean to have a **smooth map** $f : S \to S'$ between smooth surfaces? We want:

(1) f is continuous as a map of topological spaces,
(2) f is smooth in local coordinates.



The idea in (2) is you pick local coordinates (x, y) near p, and $(\widetilde{x}, \widetilde{y})$ near $f(p)$. Then $f(x, y) = (\widetilde{x}(x, y), \widetilde{y}(x, y))$ and you want $\widetilde{x}(x, y), \widetilde{y}(x, y)$ to be smooth functions of x, y.

More abstractly: for each $p \in S$, we require that for some coordinate patches

$$V_i \xrightarrow{F_i} U_i \subset S \qquad\qquad V_j' \xrightarrow{F_j'} U_j' \subset S'$$

with $p \in U_i$, $f(p) \in U_j'$, the map f written in local coordinates:

$$(F_j')^{-1} \circ f \circ F_i : \mathbb{R}^2 \supset V_i \xrightarrow{F_i} U_i \xrightarrow{f} U_j' \xrightarrow{(F_j')^{-1}} V_j' \subset \mathbb{R}^2$$

is smooth wherever it is defined.[1] A tedious exercise is to show that if it holds for some V_i, V_j' as above, then it must hold for all V_i, V_j' as above (using that transition maps are smooth).

[1]Explicitly, it is defined on $F_i^{-1}(f^{-1}(U_j'))$.

Remark. *Any abstract surface that "lives inside" $\mathbb{R}^3$ is a smooth surface in $\mathbb{R}^3$ in the sense of Section 2.3. We first clarify what "lives inside" means: you have an abstract smooth surface S and a* ***smooth embedding*** *$f : S \to \mathbb{R}^3$ (meaning a homeomorphism onto the image, such that the map and its inverse are smooth). Then $f(S) \subset \mathbb{R}^3$ is a smooth surface in $\mathbb{R}^3$: the local parametrizations for $f(S)$ are given by $f \circ F_i$ using the F_i of Definition 2.11.*

2.5 Riemann surfaces

In Definition 2.11, replacing $\mathbb{R}^2$ by $\mathbb{C}$, and "smooth" by "holomorphic" we obtain:

Definition 2.13. *A* ***Riemann surface*** *is a Hausdorff topological space S, together with a family of homeomorphisms, called* ***local parametrizations***,

$$F_i : (\text{open subset } V_i \subset \mathbb{C}) \to (\text{open subset } U_i \subset S),$$

such that the U_i cover S, so $S = \cup U_i$, and on overlaps the ***transition maps*** *are holomorphic:*

$$F_j^{-1} \circ F_i : \mathbb{C} \to \mathbb{C}$$

is holomorphic wherever defined.[1]

$\diamond$ Notice that S is automatically a topological surface.
$\diamond$ Notice that S is automatically a smooth surface.
$\diamond$ Each F_i determines one **holomorphic local coordinate**: $q \in S$ near p corresponds to $z = F_i^{-1}(q) \in \mathbb{C}$ in the parametrization F_i. As there is just one, we could call it $f_i = F_i^{-1} \in \mathbb{C}$.

Remark 2.14 (Smooth manifolds)**.** *To define* ***complex n-manifolds****, replace $\mathbb{C}$ by $\mathbb{C}^n$ above.*

Example. The torus as a Riemann surface. Consider $T^2 = \mathbb{C}/\Lambda$ for a lattice

$$\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2 \subset \mathbb{C},$$

where $\omega_1, \omega_2 \in \mathbb{C}$ are $\mathbb{R}$-linearly independent (i.e. not real multiples of each other). The quotient $\mathbb{C}/\Lambda$ is a topological space. It is Hausdorff because the equivalence relation $\{(z, z+\lambda) : \lambda \in \Lambda, z \in \mathbb{C}\} \subset \mathbb{C} \times \mathbb{C}$ is a closed subset. For any point $p \in T^2$, pick a representative point $\widetilde{p} \in \mathbb{C}$, meaning $p = [\widetilde{p}]$ in the quotient. Consider

$$D_{\widetilde{p}} = \{z \in \mathbb{C} : |z - \widetilde{p}| < \varepsilon\},$$

the disc with centre $\widetilde{p}$, radius $\varepsilon = \min\{|\omega_1|, |\omega_2|\}/100$. No two points in $D_{\widetilde{p}}$ differ by Λ, so

$$F_{\widetilde{p}} : (V_{\widetilde{p}} = D_{\widetilde{p}} \subset \mathbb{C}) \to \left(U_{\widetilde{p}} = \{[z] \in T^2 : z \in D_{\widetilde{p}}\} \subset T^2\right), \, F_{\widetilde{p}}(z) = [z]$$

is a homeomorphism. Since $p \in U_{\widetilde{p}}$, we get $T^2 = \cup U_{\widetilde{p}}$ (the union over all choices of $\widetilde{p} \in \mathbb{C}$ for all $p \in T^2$). Finally, suppose $[w] \in U_{\widetilde{p}} \cap U_{\widetilde{q}} \subset T^2$ is in an overlap. Say $[w]$ has local coordinates $w \in V_{\widetilde{p}}$ and $w + n\omega_1 + m\omega_2 \in V_{\widetilde{q}}$ respectively, where $n, m \in \mathbb{Z}$ are integers. Then the transition map is holomorphic since it is a translation:

$$F_{\widetilde{q}}^{-1} \circ F_{\widetilde{p}} : \mathbb{C} \to \mathbb{C}, \, z \mapsto z + n\omega_1 + m\omega_2.$$

What is a **holomorphic map** $f : S \to S'$ between Riemann surfaces? We want:

(1) f is continuous as a map of topological spaces,
(2) f is holomorphic in local coordinates.

The meaning of (2) is just as in Section 2.4, replacing $\mathbb{R}^2$ by $\mathbb{C}$, "smooth" by "holomorphic".

Example. Viewing $\mathbb{C}P^1 = \mathbb{C} \cup \{\infty\}$, we show $f : \mathbb{C}P^1 \to \mathbb{C}P^1$, $z \mapsto \frac{1}{z}$ is holomorphic (with $f(0) = \infty$, $f(\infty) = 0$). Notice f is continuous (in particular, the preimage of a neighbourhood of ∞ is a neighbourhood of 0). Write Z for the usual coordinate of $\mathbb{C}$ on the codomain. For $z \neq 0$,

[1] Explicitly, $F_j^{-1} \circ F_i$ is defined on $F_i^{-1}(U_i \cap U_j) \subset V_i \subset \mathbb{C}$.

the local expression of the map is $Z(z) = \frac{1}{z}$ which is holomorphic in z (complex-differentiable). For z close to 0, we need to use the local coordinate $W = 1/Z$ on the codomain because $Z = f(z)$ is close to ∞. The local expression becomes $W(z) = 1/f(z) = z$, which is holomorphic in z. Near $z = \infty$, we use the local coordinate $w = 1/z$ on the domain, and the local expression becomes $Z(w) = f(z) = f(1/w) = w$, which is holomorphic in w.

3. When are two surfaces different?

3.1 Homeomorphisms, diffeomorphisms, biholomorphisms

The class of surfaces affects when we want to view two surfaces as being the same or different. You have seen this in mathematics before: we think of two sets as "being the same" (**isomorphic**) if there is a bijection : $S_1 \to S_2$ between them, whereas for vector spaces (sets with additional structures called addition and rescaling) we want the bijection to preserve the additional structures, so we want a bijection f with f, f^{-1} both *linear*. We define:

(1) Two topological surfaces are isomorphic if they are **homeomorphic**.
Explicitly: $f : S_1 \to S_2$ is a bijection, and f, f^{-1} are continuous.
(2) Two smooth surfaces in $\mathbb{R}^3$ are isomorphic if they are **diffeomorphic**.
Explicitly: $f : S_1 \to S_2$ is a bijection, and f, f^{-1} are smooth.[1]
(3) Two abstract smooth surfaces are isomorphic if they are **diffeomorphic**.
(4) Two Riemann surfaces are isomorphic if they are **biholomorphic**,
Explicitly: $f : S_1 \to S_2$ is a bijection, and f, f^{-1} are holomorphic.

Notice that a diffeomorphism/biholomorphism is also a homeomorphism, so the underlying topological surfaces are the same. However, for all we know, there may be several ways to turn a topological surface into a smooth surface or a Riemann surface, and these may not be related by a diffeomorphism/biholomorphism even though the surfaces are homeomorphic.

Example. In previous courses you have seen Möbius maps: they are biholomorphisms

$$\mathbb{C} \cup \{\infty\} = \mathbb{C}P^1 \to \mathbb{C}P^1 = \mathbb{C} \cup \{\infty\},\ z \mapsto \frac{az+b}{cz+d},$$

where $a, b, c, d \in \mathbb{C}$ with $ad - bc \neq 0$. These maps don't change if you rescale all a, b, c, d by the same non-zero complex number, so you may arrange that $ad - bc = 1$ (which leaves the freedom of rescaling all by ± 1). So such maps are parametrized by $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in SL(2, \mathbb{C})/(\pm I) = PSL(2, \mathbb{C})$. As the maps compose according to matrix multiplication, the group of Möbius maps is isomorphic to $PSL(2, \mathbb{C})$. For example, the inverse of a Möbius map can now be calculated by finding the inverse matrix: $\frac{z-i}{z+i}$ is inverse to $\frac{iz+i}{-z+1}$ since $\left(\begin{smallmatrix} 1 & -i \\ 1 & i \end{smallmatrix}\right)$ is inverse to $\frac{1}{2i}\left(\begin{smallmatrix} i & i \\ -1 & 1 \end{smallmatrix}\right)$. Möbius maps have nice properties: circles/lines map to circles/lines, and angles are preserved (see Section 21.1).

Exercise. Recall we defined homogeneous coordinates for $\mathbb{C}P^1$, so $z \in \mathbb{C} \subset \mathbb{C} \cup \{\infty\}$ becomes the point $[1 : z]$. Show that Möbius maps become $\mathbb{C}P^1 \to \mathbb{C}P^1$, $[z_0 : z_1] \mapsto [cz_0 + dz_1 : az_0 + bz_1]$ in homogeneous coordinates. This is why they are also called *projective linear transformations*.

Exercise. Check that Möbius maps are holomorphic.

Example. We check that the upper half-plane $\mathbb{H} = \{z \in \mathbb{C} : \operatorname{Im} z > 0\}$ is biholomorphic to the open unit disc $D = \{z \in \mathbb{C} : |z| < 1\}$ via $f(z) = \frac{z-i}{z+i}$. We know $f : \mathbb{C}P^1 \to \mathbb{C}P^1$ is a biholomorphism (and $f^{-1}(z) = \frac{iz+i}{-z+1}$), so it remains to check $f(\mathbb{H}) = D$. As $0, i, \infty \in \mathbb{C} \cup \{\infty\}$ map to $-1, 0, 1 \in \mathbb{C} \cup \{\infty\}$, the circle/line $\mathbb{R} \cup \{\infty\}$ must map to the circle/line $S^1 = \partial D$. Since f sends open sets to open sets (as f^{-1} is continuous), the connected component $\mathbb{H}$ of

[1] *Non-examinable:* Since surfaces in $\mathbb{R}^3$ are an abstract surface S together with the additional structure of an embedding $S \to \mathbb{R}^3$, a more appropriate notion of isomorphic is actually **isotopy**: a smooth family of embeddings. Explicitly: a smooth map $H : S \times [0, 1] \to \mathbb{R}^3$ such that $H(\cdot, 0) : S \to \mathbb{R}^3$, $H(\cdot, 1) : S \to \mathbb{R}^3$ are the two embedded surfaces, and we want $H(\cdot, t) : S \to \mathbb{R}^3$ to be an embedding for each $t \in [0, 1]$.

$(\mathbb{C}\cup\{\infty\})\setminus(\mathbb{R}\cup\{\infty\})$ must map bijectively onto one of the two connected components of $(\mathbb{C}\cup\{\infty\})\setminus S^1$. As $f(i)=0$, it follows that $f(\mathbb{H})=D$. (*Alternatively, notice that the distance of $z\in\mathbb{H}$ from i is less than the distance from $-i$, so $|\frac{z-i}{z+i}|<1$, so the map lands inside D.*)

Exercise. Show that a Möbius map sends $\mathbb{H}\to\mathbb{H}$ precisely when $a,b,c,d\in\mathbb{R}$ are real and $ad-bc>0$. Thus, these Möbius maps determine the subgroup $PSL(2,\mathbb{R})\subset PSL(2,\mathbb{C})$. (*Hint: first impose that $\mathbb{R}\to\mathbb{R}$, then you just need to ensure the maps don't flip $\mathbb{H}$ to $-\mathbb{H}$.*)

Cultural Remark. An **automorphism** of S is a biholomorphism $S\to S$. We will prove in the course that the automorphisms of $\mathbb{C}P^1$ are precisely the Möbius maps $PSL(2,\mathbb{C})$. The automorphisms of $\mathbb{H}$ and D are also precisely the Möbius maps, respectively $PSL(2,\mathbb{R})$ and $PSU(1,1)$ (see Section 21.1). The result for $\mathbb{C}P^1$ does not immediately imply the other two because it is not easy to show that a given automorphism extends continuously to the boundary (once we know this for $f:\mathbb{H}\to\mathbb{H}$, a *reflection principle* trick $f(\overline{z})=\overline{f(z)}$ gives us an automorphism of $\mathbb{C}P^1$). A simpler approach is to first prove the result for D, then the result for $\mathbb{H}$ follows by using the above biholomorphism $\mathbb{H}\to D$ (can you see why?). For D the result follows by Schwartz's lemma: see the last two Lemmas in the Appendix.

3.2 Classification of tori

Definition 3.1 (Torus)**.** *A **torus** is any topological space X which is homeomorphic to $S^1\times S^1$ (using the product topology).*

There are several (homeomorphic) ways to describe S^1 as a topological space:

(1) $S^1=\{z\in\mathbb{C}:|z|=1\}\subset\mathbb{C}$ with the subspace topology,
(2) $S^1=\mathbb{R}/\mathbb{Z}$ with the quotient topology, where we identify $x\sim x+n$ any $n\in\mathbb{Z}$,
(3) $S^1=[0,1]/(0\sim 1)$ with the quotient topology.

For example, a homeomorphism from (2) to (1) is $x\mapsto e^{2\pi ix}$.

The torus viewed as the square by identifying parallel sides arises from description (3) of S^1; the torus as a quotient $\mathbb{R}^2/\mathbb{Z}^2$ by the translation group $\mathbb{Z}^2$ arises from description (2); and the torus $S^1\times S^1\subset\mathbb{C}\times\mathbb{C}=\mathbb{R}^4$ arises naturally from description (1).

As a topological surface, all tori are the same: homeomorphic to $S^1\times S^1$. But how many different smooth surfaces are homeomorphic to $S^1\times S^1$, and how many different Riemann surfaces are homeomorphic to $S^1\times S^1$? We will not prove the following hard theorem (a consequence of the classification of compact surfaces):

Theorem 3.2. *Any smooth surface which is topologically a torus is diffeomorphic to $S^1\times S^1$.*

This turns out to be false for Riemann surfaces: there are many non-biholomorphic Riemann surfaces which are tori. Again, we will not prove the following hard theorem:

Theorem 3.3 (Elliptic curves over $\mathbb{C}$)**.** *Any Riemann surface which is topologically a torus is biholomorphic to $\mathbb{C}/\Lambda$ for some lattice Λ which we may rescale so that $\Lambda=\mathbb{Z}\cdot 1+\mathbb{Z}\cdot\tau$ with $\tau\in\mathbb{H}=\{z\in\mathbb{C}:\mathrm{Im}(z)>0\}$. These are called the **elliptic curves**.*

We will now show explicitly why two such tori $\mathbb{C}/\Lambda_1,\mathbb{C}/\Lambda_2$ are diffeomorphic, and we will show that they are not always biholomorphic.

To show that they are diffeomorphic, we might as well show that all quotients $\mathbb{C}/\Lambda$ are diffeomorphic to $\mathbb{R}^2/\mathbb{Z}^2$ (which is the case $\tau=i$), then $\mathbb{C}/\Lambda_1\cong\mathbb{R}^2/\mathbb{Z}^2\cong\mathbb{C}/\Lambda_2$ are diffeomorphic, as required. Identifying $\mathbb{C}=\mathbb{R}^2$, $z\equiv x+iy$, write $\tau=a+ib$. Matrix multiplication

$$\left(\begin{smallmatrix}1&a\\0&b\end{smallmatrix}\right):\mathbb{R}^2\to\mathbb{R}^2$$

is of course a smooth map (it is linear!) and it is bijective (the determinant $b = \mathrm{Im}(\tau) > 0$ is non-zero), and it maps $\mathbb{Z}^2 \to \Lambda$ bijectively (the columns are the images of the standard basis). Thus it defines a diffeomorphism $\mathbb{R}^2/\mathbb{Z}^2 \to \mathbb{C}/\Lambda$. The map is however not holomorphic.[1]

More generally, suppose we are given a biholomorphism

$$f : \mathbb{C}/\Lambda \to \mathbb{C}/\Lambda' \quad \text{where } \Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2, \quad \Lambda' = \mathbb{Z}\omega_1' + \mathbb{Z}\omega_2'.$$

This means that[2] it arises from quotienting a holomorphic map

$$\widetilde{f} : \mathbb{C} \to \mathbb{C} \text{ with } \widetilde{f}(\Lambda) \subset \Lambda'$$

which is injective on any $\mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$-translate of the unit square $(0,1) \times i(0,1) \subset \mathbb{C}$. It follows that $\widetilde{f}$ maps Λ bijectively to Λ'. It easily follows that $\widetilde{f}$ grows linearly in $z \in \mathbb{C}$. So the Taylor series of $\widetilde{f}$ does not contain order z^2 or higher terms. So $\widetilde{f}(z) = Az + B$ for some $A \in \mathbb{C} \setminus \{0\}, B \in \mathbb{C}$. Taking $z = 0$ shows that $B \in \Lambda'$, so by composing $\widetilde{f}$ with the translation $z \mapsto z - B$ we may as well assume that $B = 0$. So $\widetilde{f}(z) = Az$ is linear. Thus the problem reduces to classifying lattices $\Lambda \subset \mathbb{C}$ up to $\mathbb{C}$-linear bijections!

Since $A\omega_1, A\omega_2$ is required to be a $\mathbb{Z}$-linear basis for Λ', the two bases $A\omega_1, A\omega_2$ and ω_1', ω_2' of Λ' differ by a $\mathbb{Z}$-linear bijection (think of row-reduction but working over $\mathbb{Z}$). So

$$\omega_1' = aA\omega_1 + bA\omega_2 \qquad \omega_2' = cA\omega_1 + dA\omega_2.$$

for some invertible integer-valued matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in GL(2,\mathbb{Z})$. Thus, using the convention that $\tau = \pm\frac{\omega_1}{\omega_2}$ adjusting the sign so that $\tau \in \mathbb{H} = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$,

$$\tau' = \pm\frac{\omega_1'}{\omega_2'} = \pm\frac{aA\omega_1 + bA\omega_2}{cA\omega_1 + dA\omega_2} = \pm\frac{\pm a\tau + b}{\pm c\tau + d}.$$

So the matrix acts on the τ parameter like a Möbius map. By properties of Möbius maps, since $\tau, \tau' \in \mathbb{H}$, we deduce that $\tau' = M \cdot \tau$ where $M = \pm\left(\begin{smallmatrix} \pm a & b \\ \pm c & d \end{smallmatrix}\right) \in PSL(2,\mathbb{Z})$.

Corollary 3.4. *$\mathbb{C}/(\mathbb{Z}1 + \mathbb{Z}\tau) \cong \mathbb{C}/(\mathbb{Z}1 + \mathbb{Z}\tau')$ are biholomorphic if and only if $\tau, \tau' \in \mathbb{H} = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$ lie in the same orbit of the $PSL(2,\mathbb{Z})$-action on $\mathbb{H}$ by Möbius maps.*

Corollary 3.5. *Riemann surfaces which are topologically a torus are classified up to biholomorphism by $[\tau] \in \mathbb{H}/PSL(2,\mathbb{Z})$.*

Cultural remark: this **moduli space**, $\mathbb{H}/PSL(2,\mathbb{Z})$, of modular parameters $[\tau]$ is in fact itself a Riemann surface biholomorphic to $\mathbb{C}$.

4. The Euler characteristic

4.1 Euler characteristic of regular polyhedra

Notice the following pattern in the number of vertices, edges, faces of the Platonic solids:

Regular polyhedron	Face type	V	E	F	$\chi = V - E + F$
Tetrahedron	Triangle	4	6	4	2
Cube	Square	8	12	6	2
Octahedron	Triangle	6	12	8	2
Dodecahedron	Pentagon	20	30	12	2
Icosahedron	Triangle	12	30	20	2

The alternating difference $\chi = V - E + F$ is called the **Euler characteristic**. Why is it

[1] $f(x+iy) = (x+ay) + iby$ has: $\partial_x f = 1$, $\partial_y f = a + ib$. The Cauchy-Riemann equations $\partial_x f = -i\,\partial_y f$ fail.

[2] Strictly speaking, we only know this locally, as we used the quotient $\mathbb{C} \to \mathbb{C}/\text{Lattice}$ to define the holomorphic local parametrizations. However, by the **Identity theorem** from complex analysis you know that you can patch together the local Taylor series uniquely to obtain a global holomorphic map defined on $\mathbb{C}$.

always 2 for Platonic solids? It turns out χ is a **topological invariant** of topological surfaces, meaning it is a quantity which is the same for any two surfaces which are homeomorphic. Platonic solids are homeomorphic to the sphere, so $\chi = \chi(S^2) = 2$. The homeomorphism between the Platonic solid and the sphere defines a **cellular decomposition** of the sphere: a subdivision of the sphere into regions homeomorphic to discs (in this case, curved polygons).
Example. Section 1.2.(1) shows a **triangulation** induced by a tetrahedron (get curved triangles).

4.2 Cellular decomposition

Definition 4.1 (Cellular decomposition)**.** *A **cellular decomposition** of a topological surface S is a collection of continuous maps, called **cells**,*

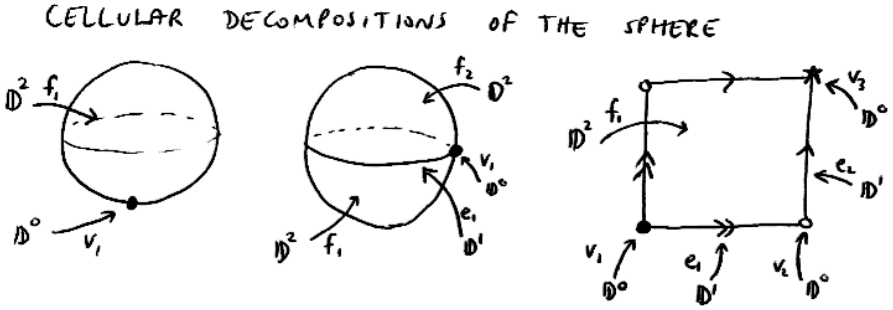
$$v_i : \mathbb{D}^0 \to S \qquad e_j : \mathbb{D}^1 \to S \qquad f_k : \mathbb{D}^2 \to S$$

respectively called 0-cells, 1-cells, 2-cells, where[1] $\mathbb{D}^n = \{p \in \mathbb{R}^n : \|p\| \leq 1\}$ *is the n-dimensional unit disc, and we require that:*

(1) *each map restricted to the interior of the disc is a homeomorphism onto the image,*[2]
(2) *the boundary of the disc is mapped into the image of the lower-dimensional cells,*[3]
(3) *S is partitioned by the interiors of the cells.*[4]

Remarks.
◇ The maps restricted to the boundary $\partial\mathbb{D}^n$ are called **attaching maps** (can be non-injective).
◇ Notice you are building the space inductively by dimension: you start with a bunch of points $X^0 = \bigsqcup v_i(\mathbb{D}^0)$, then you attach line segments $X^1 = X^0 \cup \bigcup e_j(\mathbb{D}^1)$ where the attaching maps $e_j|_{\{0,1\}}$ land inside X^0, and finally you attach 2-discs $X^2 = X^1 \cup \bigcup f_k(\mathbb{D}^2) = S$ where the attaching maps $f_k|_{S^1}$ land in X^1. The subspace $X^i \subset S$ is called the i-**skeleton**.
◇ Condition (3) ensures there are no redundancies or silly overlaps: the vertices are distinct, no vertices touch the interior of an edge or face, no edge touches the interior of a face.
◇ The above definition works more generally for any n-manifold M, in which case you can have cells $c_i : \mathbb{D}^{d_i} \to M$ of any dimension $d_i \in \{0, 1, 2, \ldots, n\}$.
◇ A **triangulation** is a cellular decomposition, where the faces are identified (via homeomorphisms) with triangles and the attaching maps are all injective, so that the boundary edges of the triangles are precisely the 1-cells, and the boundary vertices of the edges are precisely the 0-cells. These conditions are quite harsh, so it is usually very messy[5] to triangulate a surface.

Examples. Here are three examples of cellular decompositions of S^2:



[1] So $\mathbb{D}^0 = \{\text{point}\}$, $\mathbb{D}^1 = [0,1] \subset \mathbb{R}$, $\mathbb{D}^2 = \mathbb{D} = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\} \subset \mathbb{R}^2$. Interiors: $\mathrm{Int}(\mathbb{D}^0) = \mathbb{D}^0$, $\mathrm{Int}[0,1] = (0,1)$, $\mathrm{Int}(\mathbb{D}^2) = D = \{z \in \mathbb{R}^2 : \|z\| < 1\}$. Boundaries: $\partial\mathbb{D}^0 = \emptyset$, $\partial\mathbb{D}^1 = \{0\} \cup \{1\}$, $\partial\mathbb{D}^2 = S^1$.
[2] $e_j : (0,1) \to e_j((0,1)) \subset S$, $f_k : D \to f_k(D) \subset S$ are homeomorphisms (no condition on v_i as $\mathrm{Int}\,\mathbb{D}^0 = \emptyset$).
[3] $e_j(0), e_j(1) \in \bigcup v_i(\mathbb{D}^0)$ and $f_k(S^1) \subset \bigcup v_i(\mathbb{D}^0) \cup \bigcup e_j(\mathbb{D}^1)$.
[4] $S = \bigsqcup v_i(\mathbb{D}^0) \sqcup \bigsqcup e_j(\mathrm{Int}\,\mathbb{D}^1) \sqcup \bigsqcup f_k(\mathrm{Int}\,\mathbb{D}^2)$ is a disjoint union of subsets.
[5] try to triangulate the torus, viewed as a square with parallel sides identified.

Here $f_1 : \mathbb{D}^2 \cong$ (square)$\to$ (square with identifications), and this map is a homeomorphism on the interior, but on the boundary it is not injective. Notice that (just as for the Platonic solids, which also yield cellular decompositions of S^2) the alternating sum of the numbers of cells is always 2:

$$1 - 0 + 1 = 2 \qquad 1 - 1 + 2 = 2 \qquad 3 - 2 + 1 = 2.$$

A very general (hard) fact from algebraic topology is:

Theorem 4.2. *Any compact topological manifold M (e.g. a compact topological surface) "admits"*[1] *a cellular decomposition and the alternating sum of the numbers of cells*

$$\chi(M) = (\#0\text{-cells}) - (\#1\text{-cells}) + (\#2\text{-cells}) - (\#3\text{-cells}) + \cdots$$

is the same for any cellular decomposition. It is called the ***Euler characteristic*** *of M.*

Corollary 4.3. *If M, N are homeomorphic topological manifolds then $\chi(M) = \chi(N)$. So the Euler characteristic is a topological invariant.*

Proof. If $f : M \to N$ is a homeomorphism, then a cellular decomposition $c_i : \mathbb{D}^{d_i} \to M$ of M determines a cellular decomposition $f \circ c_i : \mathbb{D}^{d_i} \to N$ of N. So $\chi(M) = \sum(-1)^{d_i} = \chi(N)$. □

Example. We obtained the torus from a square by identifying the parallel edges. The whole square is a 2-cell $f_2 : \mathbb{D}^2 \to T^2$, the two non-parallel edges are two 1-cells $e_1, e_2 : \mathbb{D}^1 \to T^2$, and the four vertices of the square are identified with one 0-cell $v_1 : \mathbb{D}^0 \to T^2$. Thus

$$\chi(T^2) = 1 - 2 + 1 = 0.$$

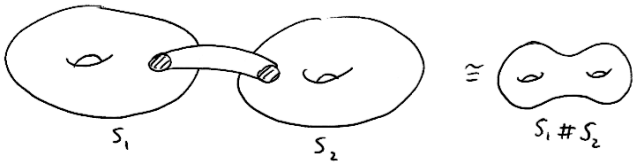
For $\mathbb{R}P^2$ the four vertices instead define two 0-cells, so

$$\chi(\mathbb{R}P^2) = 2 - 2 + 1 = 1.$$

4.3 Connected sum

In general, given two surfaces S_1 and S_2, we can form two new surfaces:

(1) the disjoint union: $S_1 \sqcup S_2$,
(2) the connected sum: $S_1 \# S_2$.



Easy exercise. Show that any connected component of a surface is a surface. Deduce that any surface equals a disjoint union of connected surfaces.

The **connected sum** $S_1 \# S_2$ is obtained by removing a "disc" from each of the two surfaces and identifying the circular boundaries. This identification is the same (up to homeomorphism) as attaching a cylinder by gluing the two boundaries of the cylinder onto the two boundaries of the removed discs.

Exercise. Check that $S_1 \# S_2$ is indeed a topological surface. Convince yourself that if S_1, S_2 are connected then, up to homeomorphism, it does not matter which "discs" you pick.

Exercise. Connected sum with a sphere does nothing.

[1] Non-examinable: the statement as written is known in all dimensions except 4 (and in dimension 2 one can even obtain a triangulation). In reality, one only cares about a cellular decomposition "up to homotopy equivalence": loosely, a type of continuous deformation that is more drastic than just homeomorphisms, and it happens to preserve χ. As an example, you can squash a cylinder to turn it into a circle, both have $\chi = 0$. That, up to homotopy, manifolds have cellular decompositions was proved by John Milnor in his 1959 paper, *On spaces having the homotopy type of a CW-Complex.*

Connected sum with a torus is the same as attaching a handle (Section 4.5).

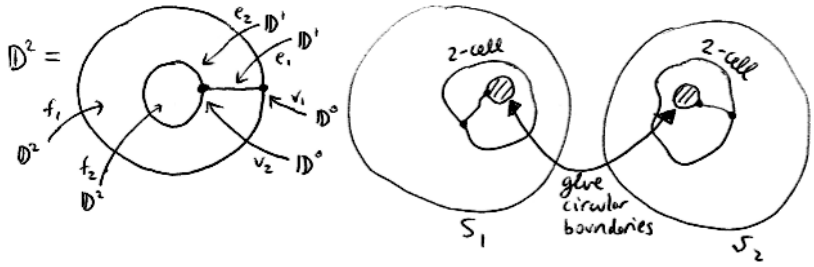
Example. By Exercise sheet 1, $\mathbb{R}P^2$ is obtained by gluing a disc $\mathbb{D}$ onto a Möbius band M along the circular boundary. So $\mathbb{R}P^2 \setminus \mathbb{D} = M$. So connected sum with $\mathbb{R}P^2$ is the same as attaching a Möbius band (Section 4.6).

4.4 Additivity of the Euler characteristic

Lemma 4.4. (1) $\chi(S_1 \sqcup S_2) = \chi(S_1) + \chi(S_2)$,
(2) $\chi(S_1 \# S_2) = \chi(S_1) + \chi(S_2) - 2$.

Proof. [1] Pick a cellular decomposition of S_1, S_2. Then this defines a cellular decomposition of $S_1 \sqcup S_2$ so (1) follows immediately. The idea in (2) is that we remove two faces, which makes χ drop by 2, and we identify the circular boundaries so we lose one copy of S^1, which does not matter for χ since $\chi(S^1) = 0$ (since S^1 is a point with an interval attached to the point, so $\chi = 1 - 1 = 0$). This idea is correct if you triangulate S_1, S_2 and remove two triangular faces. However, if we instead want to work with cellular decompositions (which arise more naturally than triangulations) then the rigorous proof is a little more involved, as follows.

The surface $S_1 \# S_2$ up to homeomorphism only depends on the the choice of connected components in S_1 and in S_2 where you pick the "discs". So we might as well pick each "disc" in the interior of a 2-cell in the cellular decomposition. To do this without destroying the cellular decomposition[2] we subdivide each original 2-cell $f_0 : \mathbb{D}^2 \to S_i$ as in the picture.



The new edges e_1, e_2 map injectively into S_i since the original f_0 is injective on $\mathrm{Int}(\mathbb{D}^2)$, and similarly the new faces f_1, f_2 are injective on $\mathrm{Int}(\mathbb{D}^2)$. However, a comment is required about v_1. If $f_0(S^1)$ already contains a 0-cell, then we can use that for v_1, and χ will not have changed.[3] If $f_0(S^1)$ does not[4] contain a 0-cell then, by the partitioning condition, $f_0(S^1)$ lies inside the image of an edge, so creating a new v_1 means subdividing an edge into two. So we are also creating a new edge. So this new vertex/edge pair does not affect χ: $1 - 1 = 0$. This invariance of χ is a special instance of the very general invariance Theorem 4.2.

Next, we remove the small faces (f_2 in the picture) from S_1, S_2, which makes χ drop by 2. Finally we identify the two boundaries of those faces which means we identify the copies in S_1, S_2 of v_2, e_2 in the above picture, so χ does not change ($+1 - 1 = 0$). So (2) follows. □

Remark. *If a topological space $S = A \cup B$ (such as a surface) is a union of two closed subsets, and suppose*[5] *S admits a cellular decomposition such that it induces cellular decompositions*

[1] **Non-examinable exercise.** *For n-manifolds M, N, explain how one constructs a connected sum $M \# N$ and show that $\chi(M \# N) = \chi(M) + \chi(N) - \chi(S^n)$, where S^n is the n-sphere.*

[2] making a hole inside the "disc" gives an annulus (up to homeomorphism), so it is no longer a 2-cell.

[3] before subdivision, f_0 contributes $+1$, after subdivision v_2, e_1, e_2, f_1, f_2 contribute $1 - 2 + 2 = 1$.

[4] the boundary of the disc must land inside the 1-skeleton, but it may land in the interior of some 1-cell.

[5] *Non-examinable*: More generally, this formula for χ holds whenever S is the union of the interiors of the closed sets A, B, as a consequence of the so-called **Mayer-Vietoris sequence** (see C3.1 Algebraic Topology).

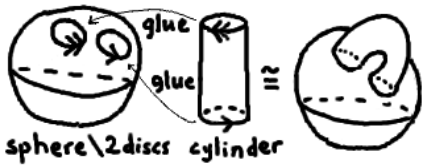
for $A \cap B, A, B$, then by counting you deduce $\chi(S) = \chi(A) + \chi(B) - \chi(A \cap B)$. Can you see how to use this to prove the above formula for $\chi(S_1 \# S_2)$?

4.5 Attaching handles to a sphere

Observe that there is a natural way to orient the boundary of a "disc"[1] in the sphere: we ask that it obeys the **right-hand rule**[2], with the thumb pointing in the normal outward direction (so for very small discs, the boundary is oriented anti-clockwise if you are looking at the sphere $S^2 \subset \mathbb{R}^3$ from far away).

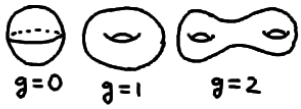
A cylinder is a space homeomorphic to $[0,1] \times S^1$. The boundaries are oriented as follows: $\{1\} \times S^1$ is oriented clockwise, $\{0\} \times S^1$ is oriented anticlockwise.

Attaching a **handle** to S^2 means you remove two disjoint "discs" from S^2, and you glue the two boundary circles of the cylinder $[0,1] \times S^1$ onto the two boundaries of the discs you removed in a way which preserves the above orientations (in practice: draw arrows on the circular boundaries, and glue in a way that respects the arrow directions). The orientation choices ensure that we can think of the handle as attached onto $S^2 \subset \mathbb{R}^3$ from the "outside":



Thus, starting from a sphere, we obtain a sequence of surfaces:

The number g of handles attached to S^2 is called the **genus of the surface**, and corresponds to the number of "doughnut holes":



Exercise. Suppose you instead remove two disjoint discs from the surface, and identify the two boundary circles (in a way that preserves their orientations). Show that the resulting surface is homeomorphic to the above handle-attachment. *Hint. Near one of the circles you can pick a closed neighbourhood that looks like $S^1 \times [-1, 0]$, now construct the required map.*

Lemma 4.5. *Attaching a handle decreases χ by 2.*

Proof. To obtain a cellular decomposition of a cylinder $S^1 \times [0,1]$, we declare that $e_1 = [0,1] \cong \{1\} \times [0,1] \subset S^1 \times [0,1]$ is a 1-cell. View each of the circles $S^1 \times \{0\}$ and $S^1 \times \{1\}$ as 1-cells e_2, e_3 which have been attached by identifying both endpoints to the same point, namely the endpoints $v_1 = (1,0)$, $v_2 = (1,1)$ of e_1. The cylinder itself defines a 2-cell bounded by e_1, e_2, e_3. Thus $\chi(\text{cylinder}) = 2 - 3 + 1 = 0$. When we attach the cylinder, we run a construction similar to the picture in the proof of Lemma 4.4. Namely, we remove two discs from the original surface (so a 2-cell), which makes χ drop by 2, whilst the identification of

[1] "Disc" will mean a continuous map $\mathbb{D}^2 \to S$ which is a homeomorphism onto its image.

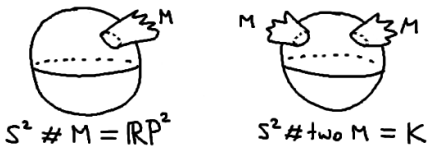
[2] thumb pointing in the normal outward direction, index finger pointing in the oriented circular direction, and middle finger pointing towards the centre of the circle.

the boundary circles does not change χ (viewing the boundary circle as a 1-cell with both endpoints attached to the same 0-cell, we lose a 1-cell and a 0-cell, leaving χ unaffected). □

4.6 Attaching Möbius bands to a sphere

In Exercise sheet 1 you study Möbius bands. The Möbius band M is the quotient of the square $[0,1]\times[0,1]$ by identifying the vertical edges in opposite directions, $(0,y)\sim(1,1-y)$. The boundaries $[0,1]\times\{0\}$ and $[0,1]\times\{1\}$ glue to give a circle.

Attaching a **Möbius band** to S^2 means you remove a "disc" from S^2, and you glue the boundary circle of M onto the boundary of the disc you removed. One cannot draw this in $\mathbb{R}^3$ without self-intersections, so schematically we will draw M as a wiggly cap:



The above are the first two of a sequence of surfaces one obtains by attaching Möbius bands to S^2 (see Exercise Sheet 1, Ex.2).

Lemma 4.6. *Attaching a Möbius band decreases χ by 1.*

Proof. This is similar to Lemma 4.5. M has a 2-cell (the square), three 1-cells (two of the four edges of the square are identified), two 0-cells (vertices are identified in pairs). So $\chi(M)=0$. When we attach M, χ drops by 1 as we remove a disc (a 2-cell) from the original surface. □

5. Classification of surfaces

5.1 Classification of compact topological surfaces

At the end of **Part A Topology** you played with gluing edges of polygons, and "proved":

Theorem 5.1. *Any compact connected topological surface is homeomorphic to:*
(1) *a sphere S^2 with $g\geq 0$ handles attached, or*
(2) *a sphere S^2 with $h\geq 1$ Möbius bands attached.*

Actually you proved the above under the additional assumption that the surface can be **triangulated** (see the Remarks in 4.2 for the definition of a triangulation).

A hard theorem[1] of topology is that every topological surface admits a triangulation.

Once a triangulation has been chosen for the compact connected surface S, by compactness there are only finitely many triangles. You then inductively build a "polygon" (up to homeomorphism, so it may have curved edges) in the plane $\mathbb{R}^2$. Start with a triangle from S, identify it with a "triangle" T_1 (a homeomorphic copy, so it can be curved) in the plane. Then consider an adjacent triangle in S, identify it with a "triangle" T_2 adjacent to T_1, and so on. Since S is connected, once you have exhausted all triangles you end up with a (typically non-convex) "polygon" in $\mathbb{R}^2$: the union of the triangles $T_1, T_2, \ldots$. However, the outer boundary edges will be identified in pairs since in S each edge belongs to two triangles. Thus, up to homeomorphism, the problem reduces to classifying regular polygons with pairwise edge identifications. You solved this combinatorial exercise in Part A Topology.

By Section 4.3, Theorem 5.1 can also be stated as follows:

[1]See http://mathoverflow.net/questions/17578/triangulating-surfaces. It is quite easy to find cellular decompositions, but much harder to triangulate. In fact, in higher dimensions, it is not true that topological manifolds are always "triangulable" (using the higher dimensional analogues of tetrahedra). See http://en.wikipedia.org/wiki/Triangulation_(topology).

Corollary 5.2. *Any compact connected topological surface is homeomorphic to one of:*

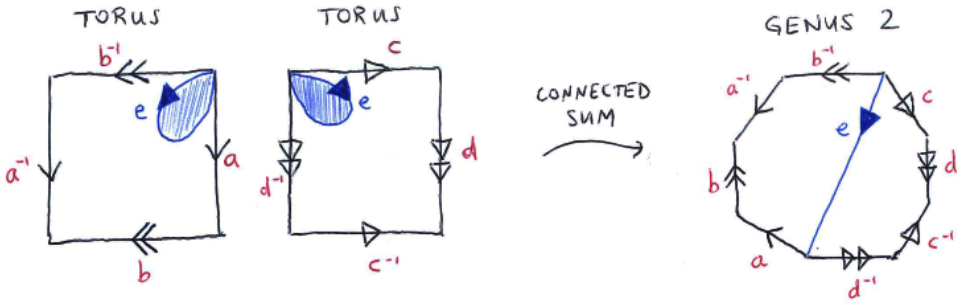
(1) S^2 *or* $T^2\#T^2\#\cdots\#T^2$ *for some number* $g \geq 0$ *of copies of* T^2,
(2) $\mathbb{R}P^2\#\mathbb{R}P^2\#\cdots\#\mathbb{R}P^2$ *for some number* $h \geq 1$ *of copies of* $\mathbb{R}P^2$.

Corollary 5.3. *For the surfaces in Theorem 5.1,*

(1) $\chi(S^2$ *with* $g \geq 0$ *handles attached*$) = 2 - 2g$,
(2) $\chi(S^2$ *with* $h \geq 1$ *Möbius bands attached*$) = 2 - h$.

Proof. Follows by Corollary 5.2 and Lemma 4.4, using that $\chi(T^2) = 0$ and $\chi(\mathbb{R}P^2) = 1$. □

Viewing the torus as a square with the usual side-identifications $aba^{-1}b^{-1}$ (reading clockwise), by performing the connected sum of two such squares we obtain the standard model of the genus 2 surface as a regular octagon with a complete set of side identifications, $aba^{-1}b^{-1}cdc^{-1}d^{-1}$. Repeating this procedure inductively, the regular polygon with $4g$ sides, identified by $a_1b_1a_1^{-1}b_1^{-1}\cdots a_gb_ga_g^{-1}b_g^{-1}$, is the standard model of the genus g surface.



Exercise. Show similarly that the regular polygon with $2h$ sides identified by $a_1a_1a_2a_2\cdots a_ha_h$ is a model for the non-orientable surface with $\chi = 2 - h$ (recall that $\mathbb{R}P^2$ is obtained from a digon with identifications aa, which is the case $h = 1$).

The Euler characteristic does not distinguish the torus T^2 from the Klein bottle K: both have $\chi = 0$. One reason why T^2 and K are not homeomorphic is that K contains a Möbius band but T^2 does not (if they were homeomorphic then T^2 would also contain one). Indeed, one can define orientability by saying that a topological surface is **orientable** if and only if it does not contain a Möbius band. In Section 6 we will discuss orientability, and show:

Corollary 5.4. *The Euler characteristic and orientability uniquely determine the topological surfaces in Theorem 5.1 up to homeomorphism ((1) are orientable, (2) are non-orientable).*

5.2 Classification of compact smooth surfaces

The analogue of Theorem 5.1 is the following hard theorem:

Theorem 5.5. *Any compact connected smooth surface is diffeomorphic to:*

(1) *a sphere* S^2 *with* $g \geq 0$ *handles attached, or*
(2) *a sphere* S^2 *with* $h \geq 1$ *Möbius bands attached.*

Exercise. Convince yourself that you can make attachments smoothly.

5.3 Classification of Riemann surfaces

This Section is non-examinable.

For cultural reasons, I mention the following theorem (we come back to this in the Appendix).

Theorem 5.6 (Uniformization theorem)**.**
Every simply-connected[1] *Riemann surface is biholomorphic to one of:*

(1) $\mathbb{C}P^1$ (*sometimes called the Riemann sphere*),
(2) $\mathbb{C}$ (*the complex plane*)
(3) $D = \{z \in \mathbb{C} : |z| < 1\}$ (*the open disc*).

Remark. Recall the upper half-plane $\mathbb{H} = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$ is biholomorphic to D (Sec. 3.1).
Every connected Riemann surface is biholomorphic to a quotient of one of those three simply-connected models by a discrete[2] *group which acts holomorphically, freely*[3] *and properly.*[4]

Later we'll see how this is related to geometry: the three models have respective curvatures $+1$ (Spherical geometry), 0 (Euclidean geometry), -1 (Hyperbolic geometry).

Example: genus 1 Riemann surfaces are, up to biholomorphism, elliptic curves by Section 3.2, so quotients of $\mathbb{C}$ by a discrete group of translations described by a lattice.

Corollary 5.7. *By studying the possible group actions in the above three cases, it turns out every connected Riemann surface is biholomorphic to one of:*

(1) $\mathbb{C}P^1$,
(2) $\mathbb{C}$, $\mathbb{C}^* = \mathbb{C} \setminus \{0\}$, *or* $\mathbb{C}/\Lambda$ *for a lattice* $\Lambda \subset \mathbb{C}$ *(these three options come from groups isomorphic to* $\{1\}, \mathbb{Z}, \mathbb{Z}^2$ *respectively),*
(3) $\mathbb{H}/G$ *for a discrete subgroup* $G \subset PSL(2,\mathbb{R}) = \{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) : a,b,c,d \in \mathbb{R}, ad-bc = 1\}/\pm I$ *acting freely by Möbius maps on* $\mathbb{H}$.

6. Orientability

6.1 Orientable versus non-orientable surfaces

There are two ways to orient a triangle, i.e. choosing arrows along the edges indicating how we wish to travel once around the triangle. One could equivalently say a surface is **orientable** if in a triangulation of the surface it is possible to pick an orientation for each triangle, so that for any two triangles which share an edge we have picked opposite orientations for that edge for the two triangles. However, the natural[5] definition is Definition 6.1.
We will loosely use the expression **disc in** S to mean a continuous injective[6] map $\mathbb{D} \to S$, and **continuous family of discs in** S to mean a continuous map $F : \mathbb{D} \times [0,1] \to S$, such that each $F_t : \mathbb{D} \to S$, $F_t(z) = F(z,t)$ is a disc in S.

---

[1]Simply-connected means: connected, and every continuous loop can be continuously shrunk to a point (every continuous map $S^1 \to S$ can be extended to a continuous map $\mathbb{D} \to S$ on the closed unit disc).

[2]Discrete means that "points are open": i.e. the one-element subsets $\{g\}$ are open sets, for $g \in G$. Example: $\mathbb{Z}^2 \subset \mathbb{R}^2$ with addition.

[3]Freely means all stabilizers are trivial: $\mathrm{Stab}_G(p) = \{g \in G : g \bullet p = p\} = \{1\}$.

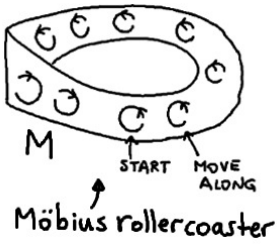
[4]Properly, for a discrete group G acting on a space S, means: any $p, q \in S$ have neighbourhoods U_p, U_q with $U_p \cap (g \bullet U_q) \neq \emptyset$ for only finitely many $g \in G$. This ensures, in particular, that S/G is Hausdorff.

[5]**Non-examinable Remark.** It is the only definition which generalizes to higher dimensions. A topological n-manifolds M is **orientable** if, after moving a disc $\mathbb{D}^n \to M$ continuously in M along any path until its image coincides with the original image, the starting and ending positions yield two embeddings $S^{n-1} = \partial\mathbb{D}^n \to M$ that differ by a homeomorphism $S^{n-1} \to S^{n-1}$ which can be continuously deformed to the identity map (see the analysis handout for the definition of deformation). For a non-orientable n-manifold, there will exist some path yielding a homeomorphism $S^{n-1} \to S^{n-1}$ which can be continuously deformed to the reflection $(x_1, \ldots, x_n) \mapsto (-x_1, x_2, \ldots, x_n)$ (viewing $S^{n-1} \subset \mathbb{R}^n$).

[6]We could strengthen this to "embedding", meaning a homeomorphism onto the image.

Definition 6.1 (Orientable surface)**.** *A topological surface S is **orientable** if for any continuous family of discs $F_t(\mathbb{D})$ in S, starting and ending at the same disc $F_0(\mathbb{D}) = F_1(\mathbb{D})$, the circular boundaries $F_0(S^1), F_1(S^1)$ are oriented in the same direction.*[1]

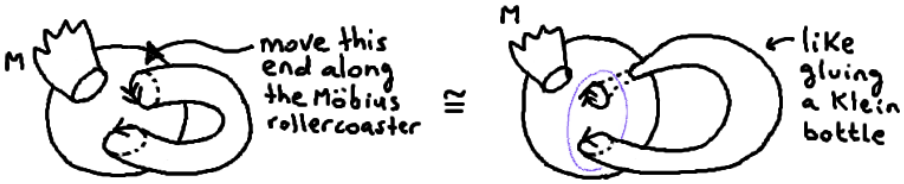
This fails for the Möbius band M, so any surface containing a copy of M is non-orientable:



Thus any surface in family (2) in Theorem 5.1 is non-orientable.

The surfaces in family (1) in Theorem 5.1 are all orientable, since they can be embedded in $\mathbb{R}^3$ with a well-defined outward normal direction, and then a family of discs will have circular boundaries oriented either always agreeing or always disagreeing with the right-hand-rule orientation.[2]

Remark. You may wonder why attaching a handle to a surface S in family (2) in Theorem 5.1 does not give anything new. Up to homeomorphism, you can move one boundary circle of the cylinder once around a Möbius band in S, which will switch the orientation of the boundary circle. If you think of S inside $\mathbb{R}^3$ with (fictitious) "self-intersections", then the cylinder is no longer attached on the outside of the sphere: one of the ends is attached from the inside. The cylinder attached in this way corresponds to connected sum with a Klein bottle $K = \mathbb{R}P^2 \# \mathbb{R}P^2$ (in the picture we drew the circle along which we take the connected sum with K). So attaching a cylinder to a surface which contains a copy of M is the same, up to homeomorphism, as the connected sum $\#\mathbb{R}P^2 \# \mathbb{R}P^2$ with two copies of $\mathbb{R}P^2$. This is what we expected from the classification theorem, since χ drops by 2 when attaching a cylinder.



6.2 Non-orientable compact surfaces cannot be embedded in $\mathbb{R}^3$

Definition 6.2 (Embedding)**.** *For a topological surface S, a map $f : S \to \mathbb{R}^3$ is an **embedding** if $S \to f(S) \subset \mathbb{R}^3$ is a homeomorphism (in particular f is injective and continuous).*
Remark. *For smooth/Riemann surfaces S, one requires in addition that the derivative map*[3] *Df is injective at every point.*

Think of embeddings in $\mathbb{R}^3$ as giving you an identical copy of the surface in $\mathbb{R}^3$.

Theorem 6.3. *A non-orientable compact surface S cannot be embedded in $\mathbb{R}^3$.*

[1] Meaning, $F_1^{-1} \circ F_0 : S^1 \to S^1$ sends the anticlockwise path e^{it} to a path $e^{is(t)}$ for a strictly increasing function $s(t)$ (so $e^{is(t)}$ is also an anticlockwise path).

[2] thumb pointing in the normal direction, index finger pointing in the oriented circular direction, and middle finger pointing towards the centre of the circle.

[3] in local coordinates, the matrix of partial derivatives.

Sketch proof. Choose a point $p \in \mathbb{R}^3$ near infinity, far away from S. For any point $q \in \mathbb{R}^3 \setminus S$, call q **even** if there is a continuous curve $c : [0,1] \to \mathbb{R}^3$ starting at $c(0) = q$, ending at $c(1) = p$, intersecting S in a finite even number of points. Define q **odd** if the number is odd. For example, the straight line segment c from q to p often works. This definition is not quite correct,[1] but some algebraic topology machinery beyond this course ensures that this definition can be made rigorous and that even/oddness is independent of the choice of c. But now, if you consider an ant walking along the equator of a Möbius band, then the positions $q_{\text{start}}, q_{\text{end}}$ of the head of the ant before and after going around the equator has changed parity (to visualise, consider the straight line segments to p). But joining the curve from q_{start} to q_{end} traced out by the head of the ant (which has not intersected S) with a curve from q_{end} to p shows that q_{start} and q_{end} have the same parity. Contradiction. □

Remark. The closed Möbius band embeds into $\mathbb{R}^3$, but is not a surface (we do not allow boundaries in this course). The open Möbius band embeds into $\mathbb{R}^3$ but is non-compact.

Remark. One can improve this proof to show that the complement of any compact surface S embedded in $\mathbb{R}^3$ has two connected components, called the "inside" and the "outside" (the 3-dimensional analogue of the **Jordan curve theorem**, which says that a continuous non-self-intersecting closed curve divides the plane into two regions).

6.3 Orientability of smooth surfaces in terms of the transition maps

If we want to show that a smooth surface S is orientable then we would have to prove that it contains no copy of the Möbius disc – this is hard in practice. It turns out there is an equivalent definition of orientability, which is more practical.

Definition 6.4 (Orientable smooth surface)**.** *A smooth (abstract) surface S is orientable, if the derivatives of all transition maps $F_j^{-1} \circ F_i$ have positive determinant on the overlaps:*

$$\det(DF_j^{-1} \circ DF_i) > 0.$$

(Warning: the failure of this condition does not imply non-orientability)[2]

We now explain this. In $\mathbb{R}^2$ the possible *ordered* bases v_1, v_2 come in two types:

(1) *right-handed bases*: these differ from the standard basis $e_1 = (1,0)$, $e_2 = (0,1)$ by a linear map with positive determinant,[3]
(2) *left-handed bases*: those which differ by a linear map with negative determinant.

The first type, are called *positively oriented bases*, and correspond to the right-hand-rule: the thumb points in the direction v_1, the index finger points in the direction v_2. Notice the unique angle less than 180° from v_1 to v_2 determines an orientation of a circle centred at 0.

Thus, the intuition for $\det(DF_j^{-1} \circ DF_i) > 0$ is that it ensures that two observers (so F_i, F_j) agree on which bases are right-handed and which are not: $e_1 = (1,0), e_2 = (0,1)$ is right-handed for the first observer in their local coordinates, and these correspond to the basis Te_1, Te_2 for the second observer where $T = DF_j^{-1} \circ DF_i$ (the derivative of the transition map $F_j^{-1} \circ F_i$). So $\det T > 0$ ensures Te_1, Te_2 is right-handed also for the second observer.

Example. The reflection $\mathbb{R}^2 \to \mathbb{R}^2$, $(x, y) \to (x, -y)$ (corresponding to complex conjugation) has $\det = -1 < 0$. The right-handed basis e_1, e_2 maps to the left-handed basis $e_1, -e_2$. Notice the reflection flips the orientation of the boundary of the unit disc from anticlockwise to clockwise,

[1] Even if S is smooth some care is needed, e.g. if c touches S tangentially then the intersection should be counted multiple times. Compare with the phenomenon that the polynomial x^2 really has two roots, not one.

[2] You can always compose F_i with the reflection $\mathbb{R}^2 \to \mathbb{R}^2$, $(x, y) \mapsto (x, -y)$, to get a new local parametrization, and the transition maps will still all be smooth. So negativity does not imply non-orientability.

[3] $v_1 = Ae_1$, $v_2 = Ae_2$ and $\det A > 0$. Explicitly A has columns v_1, v_2. So the condition is $\det(v_1|v_2) > 0$.

since $e^{it} \mapsto e^{-it}$.

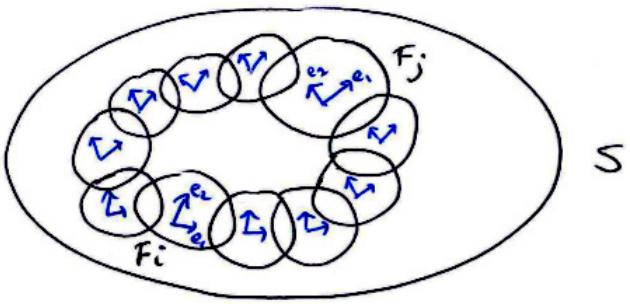
Exercise. *A smooth family of embedded discs $G_t : \mathbb{D} \to S$ starting and ending at the same disc $G_0(\mathbb{D}) = G_1(\mathbb{D})$ will flip the orientation of the boundary* $\Leftrightarrow \det(DG_1^{-1} \circ DG_0) < 0$.

Proof that Definition 6.4 really implies orientability. Let $G : \mathbb{D} \to S$ be a continuously embedded disc, and $F_i : V_i \to S$ a local parametrisation with $F_i(p) = G(0)$. Let γ be a small anti-clockwise Euclidean circle in $V_i \subset \mathbb{R}^2$ centred at p. Then $c(t) = G^{-1}(F_i(\gamma(t)))$ is a curve in $\mathbb{D}$ avoiding 0. The integral of the angle variable of $c(t)$ is either[1] $+2\pi$ or -2π, or equivalently: $c(t)$ is either oriented anti-clockwise or clockwise (if G is smooth, the two cases correspond respectively to whether $\det(F_i^{-1} \circ G)$ is positive or negative). We could have also used a point q different than $0 \in \mathbb{D}$ (by translating $z \mapsto z - q$ and then calculating the angle variable). Changing G, γ or q continuously will change the value of that integral continuously, but as it can only take values $\pm 2\pi$ it must stay constant. By the same reasoning, we can also allow more general continuous loops $\gamma \subset V_i$ avoiding p, not just circles, as long as the angle with respect to the centre p integrates to $+2\pi$ (i.e. "anti-clockwise" loops). Since $\det(DF_j^{-1} \circ DF_i) > 0$ for two overlapping parametrisations, the F_i, F_j agree on what orientation such γ curves have. So on the whole surface we have at our disposal such embedded oriented arbitrarily small "test curves" γ which determine whether an embedded disc G is "positively-oriented" or "negatively-oriented". So any continuous family $G_t : \mathbb{D} \to S$ of embedded discs are either all positively or all negatively oriented, in particular $G_0(\mathbb{D})$, $G_1(\mathbb{D})$ are oriented the same way. □

Lemma 6.5. *If a smooth surface is an orientable topological surface, then we can ensure (by composing with reflections) that the parametrizations F_i satisfy* $\det(DF_j^{-1} \circ DF_i) > 0$.

Sketch proof. Once you pick a local parametrization $F_i : V_i \to S$, this will determine (on the connected component of S where F_i lands) whether or not you need to compose each other F_j with the reflection $r : \mathbb{R}^2 \to \mathbb{R}^2$, $(x, y) \to (x, -y)$. You do this by hopping from F_i to other local parametrizations, each time comparing signs on the overlaps (if you get a negative sign, then replace F_j by $F_j \circ r$ to ensure the derivative of the transition has positive determinant).

The only problem is if there are two paths obtained by hopping from F_i to F_j, and one path requires F_j to be composed with r to obtain positivity and the other path does not require it (see the picture). But in that case, composing the first path with the reverse of the second yields a closed path along which we can move a small smooth embedded disc. Because of the disagreement in signs, the boundary of this disc at the start and end of the closed path will have flipped boundary orientation. But this contradicts that S is orientable.



□

[1]That winding around 0 in the plane more than ± 1 times is impossible without the curve self-intersecting itself is a relatively easy consequence of the Jordan Curve Theorem (a non-self-intersecting continuous curve in the plane divides the plane into two connected components).

6.4 Riemann surfaces are always orientable

Theorem 6.6. *Any Riemann surface S is an orientable surface.*

Proof. The transition maps $F_j^{-1} \circ F_i$ are holomorphic so, viewed as a map $\mathbb{R}^2 \to \mathbb{R}^2$, the derivative matrix is a composition of scaling and rotation, so the determinant is positive. □

7. Local analysis: the inverse and implicit function theorems

7.1 The inverse function theorem

Theorem 7.1 (Inverse function theorem)**.** *For any smooth map $f : \mathbb{R}^n \to \mathbb{R}^n$, if the matrix of partial derivatives at $p \in \mathbb{R}^n$ is invertible, then f is a* ***local diffeomorphism***[1] *near p. Explicitly: the theorem hands us a unique smooth map $g : \mathbb{R}^n \to \mathbb{R}^n$ defined near $f(p)$ such that $f(g(y)) = y$ and $g(f(x)) = x$ for all x, y close enough to $p, f(p)$ respectively.*

Example. Let f be the change of variables from polar coordinates (r, θ) to (x, y). So $f(r, \theta) = (r\cos\theta, r\sin\theta)$, so $Df = \left(\begin{smallmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{smallmatrix}\right)$, so $\det Df = r \neq 0$ for $r \neq 0$. So near any $(r, \theta) \in \mathbb{R}^2$ with $r \neq 0$, there is a unique local inverse of f. Aside from the $r = 0$ issue, there is no global inverse as f is 2π-periodic in θ.

Remarks.

- ◇ Arguably the most important theorem in analysis. It says simple linear algebra (the non-vanishing of a determinant) ensures the smooth invertibility of the map, locally.
- ◇ Invertibility of Df is a necessary condition:[2] if $g(f(x)) = x$ for all x close to p in $\mathbb{R}^n$, then by the chain rule $Dg \circ Df = D(\mathrm{Id}) = \mathrm{Id}$ (the identity map), so $Dg = (Df)^{-1}$.
- ◇ Since surfaces are locally parametrized by $\mathbb{R}^2$, the above theorem also holds for smooth maps between surfaces.
- ◇ It holds also for smooth maps between manifolds, since these are locally $\mathbb{R}^n$.
- ◇ Invertibility of Df can be equivalently phrased as the linear independence of the vectors $\partial_{x_1} f, \ldots, \partial_{x_n} f$ (which form the columns of the matrix Df of partial derivatives).

Corollary 7.2 (Inverse function theorem in complex analysis)**.**
For any holomorphic map $f : \mathbb{C} \to \mathbb{C}$, if $f'(z_0) \neq 0$ then f is a local biholomorphism near z_0.

Non-examinable proof. The previous theorem implies there is a smooth inverse $f^{-1} : \mathbb{R}^2 \to \mathbb{R}^2$ defined near $f(z_0)$. We need to check f^{-1} is holomorphic. So we need to check f^{-1} satisfies the Cauchy-Riemann equations. But this is equivalent to showing the matrix of partial derivatives is a scaling times a rotation (think $f'(z_0) = re^{i\theta}$). By the chain rule, $Df \cdot Df^{-1} = D(\mathrm{Id}) = \mathrm{Id}$, so Df^{-1} is the inverse of Df, so it is a scaling times a rotation since Df is. □

Remarks.

- ◇ As before, the invertibility of Df (that is, non-vanishing of $f'(z_0)$) is a necessary condition. Explicitly: if $g(f(z)) = z$ for all z close to z_0 in $\mathbb{C}$, then by the chain rule $g'(f(z_0)) \cdot f'(z_0) = 1$ so $g'(f(z_0)) = 1/f'(z_0)$.
- ◇ Since Riemann surfaces are locally parametrized by $\mathbb{C}$, the above theorem also holds for holomorphic maps between Riemann surfaces.

---

[1] Meaning: there are open neighbourhoods $U \subset \mathbb{R}^n$ of p and $V \subset \mathbb{R}^n$ of $f(p)$ such that the restriction $f|_U : U \to V$ is a diffeomorphism, so there is a unique smooth inverse $f^{-1} : V \to U$.

[2] As an example, for functions $f : \mathbb{R} \to \mathbb{R}$, for invertibility you need that the line $y =$ constant intersects the graph of f in exactly one point. By the intermediate value theorem, you deduce that f has to either always increase or always decrease. So $f' \geq 0$ or $f' \leq 0$. There are bijective smooth functions $\mathbb{R} \to \mathbb{R}$ with f' sometimes zero, such as $x \mapsto x^3$, but they are not diffeomorphisms: $x \mapsto x^{1/3}$ is not smooth at 0, because the derivative blows up there (the horizontal tangents to f become vertical tangents to f^{-1}).

- ◇ Consider a holomorphic map $f : \mathbb{C}^n \to \mathbb{C}^n$ (i.e. each entry $f_1, \dots, f_n$ is holomorphic in each of the coordinates $z_1, \dots, z_n$). Then linear independence of the complex derivatives $\partial_{z_1} f, \dots, \partial_{z_n} f$ at $p \in \mathbb{C}^n$ implies that f is a local biholomorphism near p.
- ◇ The result holds also for holomorphic maps between complex manifolds, since these are locally $\mathbb{C}^n$.

7.2 The implicit function theorem

Motivation. *When the dimensions n, m are different, there is of course no chance of finding a (local) inverse of a smooth map $f : \mathbb{R}^n \to \mathbb{R}^m$. When $n = m$ and f is invertible, $f(x) = c$ has a unique solution $f^{-1}(c) = (g_1(c), \dots, g_n(c))$, giving rise for each $c \in \mathbb{R}^m$ some unique numbers $g_1, \dots, g_n$ for which $f(g_1, \dots, g_n) = c$. Now assume $n > m$, then the next best thing to finding an inverse (which cannot exist) is finding some functions g_i which can be used to get rid of some of the variables in $\mathbb{R}^n$ and which only depend on the other variables. For example: $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = x^2 - 2y$, then $f(x, g(x)) = c$ if we take $g(x) = \frac{1}{2}(x^2 - c)$. So for the purpose of solving $f = c$ the variable y is redundant since we can replace it with a function $g(x)$ of x, but the variable x is essential. Notice the redundant variable y is the one for which $\partial_y f = -2$ is never zero, whereas $\partial_x f = 2x$ can vanish, at $x = 0$.*

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be smooth, and $n \geq m$. We want to describe the solutions of

$$f(x) = c$$

near a given solution $f(p) = c$, where $x, p \in \mathbb{R}^n$ and $c \in \mathbb{R}^m$.

Theorem 7.3 (Implicit function theorem)**.** *If m columns of $D_p f$ are linearly independent, then the variables $x_{i_1}, \dots, x_{i_m}$ corresponding to those columns are redundant. Namely, they can be replaced by unique smooth functions*

$$g_{i_1}, \dots, g_{i_m} : \mathbb{R}^{n-m} \to \mathbb{R},$$

depending only on the remaining variables x_j (so $j \neq i_1, \dots, i_m$), defined near $x = p$ and satisfying $g_{i_1}(p) = p_{i_1}, \dots, g_{i_m}(p) = p_{i_m}$, so that[1]

$$f(x)|_{(x_{i_1} = g_{i_1}, \dots, x_{i_m} = g_{i_m})} = c$$

describes all solutions x near p.

Examples. Below, we seek solutions of $f = 0$ near $p = (0, \dots, 0)$.

(1) $f(x, y) = y$: $\partial_y f = 1 \neq 0$, so $f(x, g(x)) = 0$ (indeed $g(x) = 0$).
(2) $f(x, y) = x^2 - y$: $\partial_y f = -1 \neq 0$, so $f(x, g(x)) = 0$ (indeed $g(x) = x^2$).
(3) $f(x, y) = (x + 1)^2 - 1 + y^2$: $\partial_x f|_{x=0, y=0} = 2 \neq 0$, so $f(g(y), y) = 0$ (indeed $g(y) = -1 + \sqrt{1 - y^2}$, which is defined near $y = 0$, and notice $g(0) = 0$).
(4) The unit circle S^1 is the solution set $f = 0$ for $f(x, y) = x^2 + y^2 - 1$. For points $(a, b) \in S^1$ with $b \neq 0$, $\partial_y f = 2y \neq 0$ for y close enough to b. So x is a local coordinate: S^1 is described by $(x, g(x))$ near (a, b) (secretly we know $g(x) = \sqrt{1 - x^2}$, which is smooth away from $x = \pm 1$, $y = 0$). For $a \neq 0$, $\partial_x f = 2x \neq 0$ for x close to a, so y is a local coordinate: S^1 is $(g(y), y)$ near (a, b) (secretly we know $g(y) = \sqrt{1 - y^2}$). So we have local coordinates everywhere (a, b cannot both be zero: $f(0, 0) = -1 \neq 0$).
(5) In the previous example, notice that we are locally parametrizing S^1 as the graph of a function, so we get either $(x, g(x))$ or $(g(y), y)$.

[1] More pedantically: $f^{-1}(c) = \{(x_1, x_2, \dots, x_{i_1 - 1}, g_{i_1}(x), x_{i_1 + 1}, \dots) : x = (x_1, x_2, \dots, x_{i_1 - 1}, x_{i_1 + 1}, \dots) \in \mathbb{R}^{n-m}\}$, where we omit the variables $x_{i_1}, x_{i_2}, \dots, x_{i_m}$ and we replace them by the values of $g_{i_1}, g_{i_2}, \dots$.

(6) For the unit sphere in $\mathbb{R}^3$, defined by $f(x,y,z) = x^2+y^2+z^2-1 = 0$, near any point either (x,y) or (x,z) or (y,z) are local smooth coordinates. Indeed, for the implicit function theorem to fail for f in those three cases, it would mean respectively that $\partial_z f = 0$, $\partial_y f = 0$, and $\partial_x f = 0$. But then $x = y = z = 0$, which is not a point of the sphere. Notice that we are locally paramerizing S^2 as the graph of a function, e.g. for the (x,y) case (when $\partial_z f \neq 0$) we deduce that S^2 is locally $(x, y, g(x,y))$ there.

Non-examinable proof of Theorem 7.3. By relabeling coordinates, we may assume the last m columns of $D_p f$ are linearly independent. Abbreviate $k = n - m$. Consider

$$F : \mathbb{R}^n \to \mathbb{R}^n,\ F(x_1, \ldots, x_n) = (x_1, \ldots, x_k, f(x_1, \ldots, x_n)).$$

Then $D_p F$ is invertible (try writing the matrix). By the inverse function theorem,

$$F^{-1}(x_1, \ldots, x_k, c_1, \ldots, c_m) = (x_1, \ldots, x_k, g_{k+1}, \ldots, g_n)$$

for unique functions $g_{k+1}, \ldots, g_n$ of $x_1, \ldots, x_k, c$. □

Corollary 7.4 (Smooth dependence on c in Theorem 7.3)**.** *Notice above $g_{i_1}, \ldots, g_{i_m}$ depend smoothly on c. So there are unique smooth functions*

$$G_{i_1}, \ldots, G_{i_m} : \mathbb{R}^{n-m} \times \mathbb{R}^m \to \mathbb{R}$$

defined near $x = p, y = c$ and depending only on the non-redundant x_j variables (so $j \neq i_1, \ldots, i_m$) and on $y \in \mathbb{R}^m$, so that

$$f(x)|_{(x_{i_1} = G_{i_1}, \ldots, x_{i_m} = G_{i_m})} = y$$

describes all solutions of $f(x) = y$ for x near p, and y near c.

Remark 7.5. *The set of solutions of $f(x) = y$ is therefore locally cut out by the vanishing of m functions: $x_{i_1} - G_{i_1}, \ldots, x_{i_m} - G_{i_m}$.*

Consider the change of coordinates on $\mathbb{R}^n$ (so a local diffeomorphism) defined by first permuting the coordinates of $\mathbb{R}^n$ so that we may assume the $i_1, \ldots, i_m$ above are $n - m + 1, \ldots, m$, and then changing the coordinates by $x_j \mapsto \widetilde{x}_j(x)$ with

$$\widetilde{x}_j(x) = x_j \text{ for } j \leq m \quad \text{and} \quad \widetilde{x}_j(x) = x_j - g_j \text{ for } j = n - m + 1, \ldots, m.$$

In these coordinates the solution set of $f(x) = y$ is parametrized by the first $n-m$ coordinates $\widetilde{x}_j$ and is cut out by the m equations given by the vanishing of the last m coordinates:

$$\widetilde{x}_{n-m+1} = 0,\ \ldots,\ \widetilde{x}_n = 0.$$

So locally, you can think of the inclusion of the solution set $(f(x) = y) \subset \mathbb{R}^n$ as being smoothly "the same" (diffeomorphic) to the standard inclusion $\mathbb{R}^{n-m} \subset \mathbb{R}^n$.

8. Local analysis: Embedded surfaces are locally graphs

8.1 Criterion for a local parametrization of a smooth surface in $\mathbb{R}^3$

Our next goal, is to prove Theorem 2.10, and to show that a smooth surface $S \subset \mathbb{R}^3$ is locally defined by the vanishing of a smooth function (just like S^2 is locally, in fact globally, the zero set of the function $x^2 + y^2 + z^2 - 1 : \mathbb{R}^3 \to \mathbb{R}$). Let $S \subset \mathbb{R}^3$ be a smooth surface, $F : V \to S$ a smooth map, $V \subset \mathbb{R}^2$ an open set, and $F(v_0) = p$.

Theorem 8.1. *F is a smooth local parametrization near p when restricted to a possibly smaller open neighbourhood $V' \subset V$ of $v_0 \iff \partial_x F, \partial_y F$ are linearly independent at v_0.*

Proof. The easy direction is $\Rightarrow$: there is a smooth inverse F^{-1}, so $F^{-1} \circ F = \mathrm{Id} : V' \to V'$, so by the chain rule $DF^{-1} \circ DF = \mathrm{Id}$, so DF is injective, so the two columns $\partial_x F, \partial_y F$ of the matrix DF must be linearly independent at each $v \in V'$, in particular at $v = v_0$.

Now the hard direction $\Leftarrow$. The matrix DF is a 3×2 matrix, and since its columns are linearly independent at v_0, there must be a 2×2 submatrix with non-zero determinant (basic linear algebra). WLOG assume it's the first two rows, so

$$\begin{pmatrix} \partial_x F_1 & \partial_y F_1 \\ \partial_x F_2 & \partial_y F_2 \end{pmatrix}$$

is invertible at $v = v_0$, where we clarify: we use x, y coordinates on $V \subset \mathbb{R}^2$, and we use X, Y, Z coordinates on $\mathbb{R}^3$, so explicitly $F(x, y) = (F_1(x, y), F_2(x, y), F_3(x, y))$.

Compose F with projection to the first two coordinates (X, Y) of $\mathbb{R}^3$,

$$\widetilde{F} = (X, Y) \circ F : \mathbb{R}^2 \supset V \to \mathbb{R}^3 \to \mathbb{R}^2, \ \widetilde{F}(x, y) = (F_1(x, y), F_2(x, y)).$$

Then $D\widetilde{F}$ is the above 2×2 matrix. By the inverse function theorem, $\widetilde{F}$ has a unique inverse $\widetilde{F}^{-1} : \mathbb{R}^2 \to V \subset \mathbb{R}^2$ defined near $\widetilde{F}(v_0)$. Thus $\widetilde{F}(x, y) = (X, Y) \Leftrightarrow \widetilde{F}^{-1}(X, Y) = (x, y)$, so

$$F(\widetilde{F}^{-1}(X, Y)) = F(x, y) = (X, Y, g(X, Y))$$

only depends smoothly on (X, Y) and thus it determines a unique smooth function $g(X, Y)$ (that is: the Z coordinate of points in S is determined by X, Y, near $F(v_0)$).

We now define the map that we hope is the chart inverse to the parametrization F:

$$f : \mathbb{R}^3 \to \mathbb{R}^2, (X, Y, Z) \mapsto \widetilde{F}^{-1}(X, Y).$$

Notice that f is a smooth map $\mathbb{R}^3 \to \mathbb{R}^2$, defined near $F(v_0)$, and f restricted to S near $F(v_0)$ becomes $f(X, Y, g(X, Y)) = \widetilde{F}^{-1}(X, Y) = (x, y)$ so it is the inverse of F. This concludes the proof that F is a local diffeomorphism $V \to S$ near v_0, and hence a local parametrization. □

(Non-examinable) Exercise. *Can you use the idea in the above proof to state and prove a general implicit function theorem for smooth maps $f : \mathbb{R}^n \to \mathbb{R}^m$ when $n < m$?*

Corollary 8.2 (Theorem 2.10)**.** *If F, above, is also injective then F is a smooth local parametrization on all of $V \iff \partial_x F, \partial_y F$ are linearly independent at each point of V.*

Proof. Since $F : V \to F(V)$ is injective, and by construction surjective, it is bijective, so it remains to check that $F^{-1} : F(V) \to V$ is smooth. But $F : V \to F(V) \subset S$ is a local diffeomorphism near each $v_0 \in V$ by the previous Theorem, so F^{-1} is smooth. □

8.2 Smooth surfaces in $\mathbb{R}^3$ are locally graphs

Theorem 8.3. *For any smooth surface S in $\mathbb{R}^3$, near each point $p \in S$,*

1. *S is either[1] a smooth graph $(X, Y, g(X, Y))$ over the coordinates (X, Y), or a graph over the (X, Z) coordinates, or a graph over the (Y, Z) coordinates,*
2. *either (X, Y), or (X, Z), or (Y, Z) are smooth local coordinates for S,*
3. *S is locally cut out as the zero set of a function $\mathbb{R}^3 \to \mathbb{R}$.*

Proof. The proof of Theorem 8.1 showed that S is locally $(X, Y, g(X, Y))$, for smooth g, if the first two rows of DF are linearly independent. The cases $(X, g(X, Z), Z)$, $(g(Y, Z), Y, Z)$ occur if rows $1, 3$, respectively rows $2, 3$ of DF are linearly independent. This proves (1). Also (2) follows since, say in the first case, $(X, Y) \mapsto (X, Y, g(X, Y))$ is a local parametrization.[2]

[1] We always mean *non-exclusive* "either ... or ...", so several options may occur.

[2] The composition of the diffeomorphisms $F \circ \widetilde{F}^{-1}$ in the proof of Theorem 8.1, hence a diffeomorphism.

Finally (3) follows, say in the first case, by considering the function $\mathbb{R}^3 \to \mathbb{R}$, $(X, Y, Z) \mapsto Z - g(X, Y)$ since $Z - g(X, Y) = 0$ cuts out the set $(X, Y, g(X, Y))$ as required. □

8.3 Riemann surfaces in $\mathbb{C}^2$ are locally graphs

Consider a subset of $\mathbb{C}^2$ cut out by a holomorphic equation

$$S = \{(z, w) \in \mathbb{C}^2 : f(z, w) = 0\}$$

where $f : \mathbb{C}^2 \to \mathbb{C}$ is holomorphic (e.g. a complex polynomial in z, w). Analogously to Theorems 8.1 and 8.3 (using Section 7.1 to get holomorphicity), we deduce:

Theorem 8.4. *S is a Riemann surface near $(z_0, w_0) \in S$ if and only if*

(1) *either $\frac{\partial f}{\partial z} \neq 0$ at (z_0, w_0), then S is locally $(g(w), w)$,*

(2) *or $\frac{\partial f}{\partial w} \neq 0$ at (z_0, w_0), then S is locally $(z, g(z))$,*

so S is locally the graph of a holomorphic function $g : \mathbb{C} \to \mathbb{C}$, and S is locally cut out by a holomorphic function (respectively $z - g(w) = 0$, or $w - g(z) = 0$).

Example. Consider $S = \{(z, w) \in \mathbb{C}^2 : f(z, w) = w^2 - (z - 1)(z - 2) = 0\}$ (recall Exercise sheet 1). Then $\partial_w f = 2w$ is non-zero except at $w = 0$. When $w = 0$, either $z = 1$ or 2. But then $\partial_z f = -(2z - 3) \neq 0$. So S is a Riemann surface. Recall we compactify S at $\pm\infty$ by rewriting the equation in the new variables

$$X = 1/z, \quad Y = w/z,$$

so the defining equation for S becomes $\widetilde{f}(X, Y) = Y^2 - (1 - X)(1 - 2X) = 0$, and $\pm\infty$ correspond to the two new points $(X, Y) = (0, \pm 1)$. Finally we check $S \cup \{\pm\infty\}$ is a Riemann surface at those new points: $\partial_Y \widetilde{f} = 2Y \neq 0$ at $Y = \pm 1$.

The following definition and remark are non-examinable, but I hope it will inspire your interest in **B3.3 Algebraic Curves**.

Definition 8.5 (Complex algebraic curve)**.** *A **complex algebraic curve** S is the zero set $f(z, w) = 0$ of a complex polynomial f in two variables z, w. The **singular points** are the $(z_0, w_0) \in S$ where both conditions above fail: $\partial_z f = 0, \partial_w f = 0$. So non-singular complex algebraic curves are Riemann surfaces.*

Remark 8.6 (Projective algebraic curve)**.** *Because of the maximum modulus principle, $S \subset \mathbb{C}^2$ can never be compact (otherwise $|z|, |w|$ would attain maxima, so z, w would be constant functions on S). As in the example above, S is missing some points at infinity and one systematic way of compactifying is to **projectivize the equation**. That means, you view S as a subset of[1] $\mathbb{C}P^2$. So $S \subset \mathbb{C}^2$ are the points of the form $[1 : z : w]$ of the compactification $\overline{S} \subset \mathbb{C}P^2$. To projectivize a polynomial, you make it homogeneous by simply replacing $z = z_1/z_0$, $w = z_2/z_0$ and then rescaling by the least power of z_0 to get rid of denominators.*

Example. *$w^2 - (z - 1)(z - 2) = 0$ becomes $z_2^2 - (z_1 - z_0)(z_1 - 2z_0) = 0$. We already know solutions when $z_0 \neq 0$ (we are then allowed to rescale $z = z_1/z_0$, $w = z_2/z_0$). Suppose instead $z_1 \neq 0$, then we may use local coordinates $X = z_0/z_1$, $Y = z_2/z_1$ so $[z_0 : z_1 : z_2] = [X : 1 : Y]$ (can you see why these are the same as the X, Y in the above example?). The equation becomes: $Y^2 - (1 - X)(1 - 2X) = 0$. So we get two new points $(X, Y) = (0, \pm 1)$, which correspond to $[0 : 1 : \pm 1] \in \mathbb{C}P^2$. Finally, suppose $z_2 \neq 0$ (by definition of $\mathbb{C}P^2$, the z_0, z_1, z_2 cannot all vanish). We could use $X = z_0/z_2$, $Y = z_1/z_2$, so $[z_0 : z_1 : z_2] = [X : Y : 1]$, and rewrite the*

[1] By analogy with $\mathbb{C}P^1$, we define

$$\mathbb{C}P^2 = \{[z_0 : z_1 : z_2] : (z_0, z_1, z_2) \in \mathbb{C}^3 \setminus \{0\}, [z_0 : z_1 : z_2] = [\lambda z_0 : \lambda z_1 : \lambda z_2] \text{ for any } \lambda \in \mathbb{C} \setminus \{0\}\}$$

(geometrically, think of the point $[z_0 : z_1 : z_2]$ as the complex line $\mathbb{C} \cdot (z_0, z_1, z_2) \subset \mathbb{C}^3$).

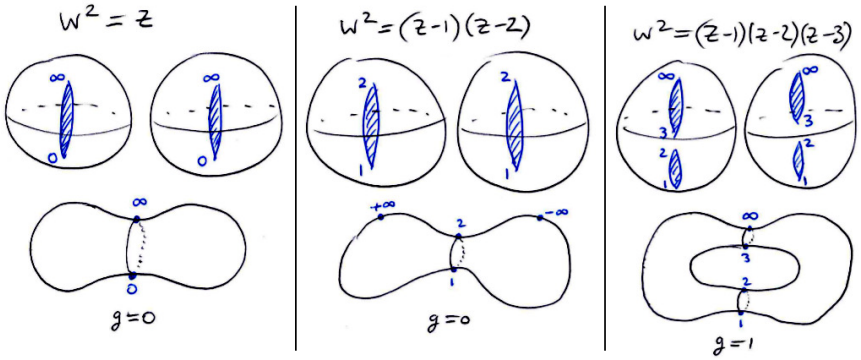
equation, but because we are not in the previous two cases, we may assume that $z_0 = z_1 = 0$, so it only remains to check whether $[0:0:1]$ is a solution, and it isn't.

(Non-examinable) Exercise. *Show that if you projectivize $w^2 = (z-1)(z-2)(z-3)$, you need to add the point $\infty = [0:0:1]$ (and you get a torus). However, show that if you projectivize $w^2 = (z-1)(z-2)\cdots(z-5)$ you get a projective curve which is singular at infinity. A better compactification than this, is described in the next example.*

Example. Consider the polynomial

$$f(w,z) = w^2 - (z-a_1)(z-a_2)\cdots(z-a_n) = 0.$$

Notice $\partial_w f = 2w \neq 0$ unless $w = 0$, and for $w = 0$ we get $z = a_j$ so the condition $\partial_z f \neq 0$ is equivalent to requiring that none of the roots $a_j \in \mathbb{C}$ are repeated. So we get a Riemann surface when the a_j are pairwise distinct.



By the methods of Exercise sheet 1, question 3, one can define a compactification at infinity which yields a Riemann surface. At infinity, we use the coordinates

$$X = \frac{1}{z} \qquad Y = \frac{w}{z^m}$$

where $m = n/2$ if n is even, and $m = (n+1)/2$ if n is odd. The equation becomes $Y^2 = (1-a_1X)\cdots(1-a_nX)$ for n even, and $Y^2 = X(1-a_1X)\cdots(1-a_nX)$ for n odd. So we compactify by adding new points $(X,Y) = (0,\pm 1)$ called $\pm\infty$ for n even, and $(X,Y) = (0,0)$ called ∞ for n odd. In both cases, we declare that Y is a local holomorphic coordinate at infinity.[1] Recall from Exercise sheet 1, question 3, that you can visualize the Riemann surface by gluing two copies of $\mathbb{C}$ with cuts and then compactifying at infinity. In the above picture, we first compactified each $\mathbb{C}$ to a $\mathbb{C}P^1$, then drew the cuts,[2] and in the lower pictures we glued the cuts to obtain the Riemann surface and determined its genus g.

Imagine increasing n by 2: this means we require two extra cuts in the planes $\mathbb{C}P^1$ that we

[1] We need to check that in a small neighbourhood of a point (z,w) for large $z \neq \infty$, the transition map from the local coordinate z to the local coordinate Y is biholomorphic. But near (X,Y) with $X \neq 0$ we can use either X or Y as local coordinate since we can express Y in terms of a holomorphic branch of the square root of a polynomial in X (this is the same argument that proves that for the Riemann surface $w^2 = z$ you may declare that w is a holomorphic coordinate near $(w,z) = (0,0)$, and z is a holomorphic coordinate elsewhere). Thus it suffices to find a holomorphic transition from z to X, away from $X = 0$. But that we know: $z \mapsto X = 1/z$ is the required local biholomorphism in z.

[2] To understand the cut: for $w^2 = (z-1)(z-2)$, why do we cut the segment $(1,2)$? The local model near $z = 1$ and near $z = 2$ is that of the square root, and for the square root we typically choose the cut along the negative real axis. In our case, we make cuts $(-\infty, 1)$ and $(-\infty, 2)$ and we pick branches of $\sqrt{z-1}$ and of $\sqrt{z-2}$. For one copy of $\mathbb{C}$, for $z = 1+ae^{i\theta} = 2+be^{i\psi}$ let's declare $\sqrt{z-1} = a^{1/2}e^{i\theta/2}$ and $\sqrt{z-2} = b^{1/2}e^{i\psi/2}$ for $\theta, \psi \in (-\pi, \pi)$. We would think that this is only acceptable if there is a cut along $(-\infty, 2) \subset \mathbb{R} \subset \mathbb{C}$, but in fact for real $z < 1$ the two discontinuities cancel out: $e^{i\pi/2}e^{i\pi/2} = e^{i\pi} = e^{-i\pi} = e^{i(-\pi/2)}e^{i(-\pi/2)}$.

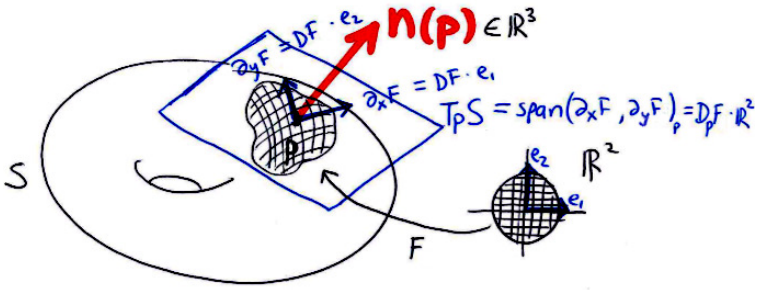
glue, giving rise to an extra handle, so the genus of the compactified Riemann surface increases by 1. For $n = 2$ we get a sphere (genus 0) and for $n = 3$ we get a torus (genus 1), so inductively:

$$\text{genus} = \tfrac{1}{2}(n-2) \text{ for } n \text{ even, and } \tfrac{1}{2}(n-1) \text{ for } n \text{ odd}$$

(corresponding respectively to $\chi = 4 - n$ and $\chi = 3 - n$). These Riemann surfaces are called **hyperelliptic curves**.[1]

9. The tangent space

9.1 Tangent space for smooth surfaces in $\mathbb{R}^3$



Let F be a local parametrization near $p \in S$, for a smooth surface S in $\mathbb{R}^3$. Recall by Theorem 2.10) that $\partial_x F$, $\partial_y F$ are linearly independent at p (we will suppress from the notation that we are evaluating at p). Thus we obtain a 2-dimensional vector subspace of $\mathbb{R}^3$, called the **tangent space**, as follows

$$\boxed{T_pS = \text{span}(\partial_x F, \partial_y F) = \text{span}(DF \cdot e_1, DF \cdot e_2) = DF \cdot \mathbb{R}^2 \subset \mathbb{R}^3}$$

where e_1, e_2 is the standard basis on the domain $\mathbb{R}^2$ of F. Think of the plane T_pS as the plane[2] in $\mathbb{R}^3$ which best approximates S near p.

The picture also shows the unit vector $n(p)$ normal to T_pS, obtained from the cross product of $\partial_x F, \partial_y F$ in $\mathbb{R}^3$ and then normalizing. We will discuss this in Section 9.5.

The above vector subspace T_pS of $\mathbb{R}^3$ is independent of the choice of parametrization F, since for another parametrization $\widetilde{F}$ (so another "observer"), we have

$$D\widetilde{F} \cdot \mathbb{R}^2 = D\widetilde{F} \cdot (D\widetilde{F}^{-1} \circ DF) \cdot \mathbb{R}^2 = DF \cdot \mathbb{R}^2,$$

using that the derivative of the transition, $D\tau = D\widetilde{F}^{-1} \circ DF$, is a linear isomorphism $\mathbb{R}^2 \to \mathbb{R}^2$.

9.2 Tangent space for abstract smooth surfaces

For a surface in $\mathbb{R}^3$, notice that DF identifies the **local tangent space**

$$T_{p_0}\mathbb{R}^2 \equiv \mathbb{R}^2 = \text{span}(e_1, e_2)$$

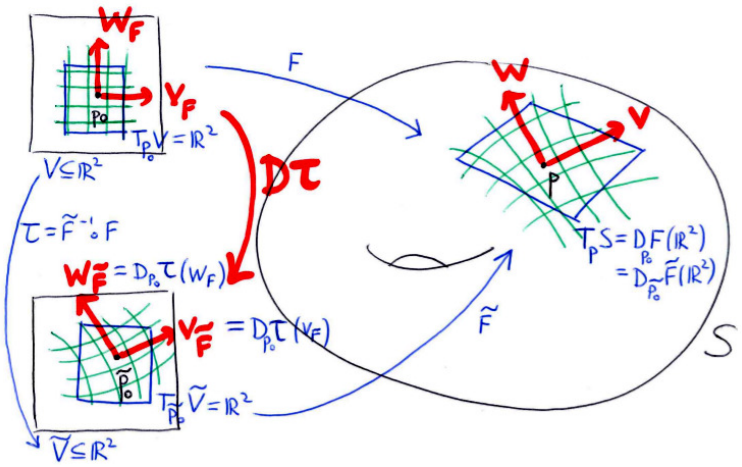
with the tangent space $T_pS = \text{span}(\partial_x F, \partial_y F) \subset \mathbb{R}^3$ (where $F(p_0) = p$). We saw above that two observers agree which plane $T_pS \subset \mathbb{R}^3$ is, because $DF \cdot T_{p_0}\mathbb{R}^2 = D\widetilde{F} \cdot T_{\widetilde{p}_0}\mathbb{R}^2$. We can rewrite this as

$$D\tau \cdot T_{p_0}\mathbb{R}^2 = T_{\widetilde{p}_0}\mathbb{R}^2$$

[1] If you are curious why these clever coordinates work at infinity, unlike the projectivization which typically gives rise to a singularity at infinity, see http://en.wikipedia.org/wiki/Hyperelliptic_curve

[2] To be precise, the plane which best approximates S near p is the translate $p + T_pS \subset \mathbb{R}^3$, but it is more convenient to work with a vector subspace so one uses T_pS.

where $\tau = \widetilde{F}^{-1} \circ F$ is the transition map between the two observers. For abstract smooth surfaces, we simply turn this equation into the definition. For abstract surfaces, there is no common ambient space (such as $\mathbb{R}^3$ above) where local observers can compare tangent spaces, so one simply works with the local tangent spaces and one remembers that $D\tau$ is the map which transforms vectors from one observer's coordinate system to the other.



An equivalent description of the tangent space, is as **equivalence classes of curves**. Namely, in local coordinates, given a local tangent vector $v \in \mathbb{R}^2 \equiv T_pU$, there is a smooth curve $c : (-\varepsilon, \varepsilon) \to U$ passing through $c(0) = p$ with initial velocity $c'(0) = v$. For example, the straight line $c(t) = p + tv$. We only care that the curve is defined for small times, so $\varepsilon > 0$ can be small. There are many choices of such curves $c(t)$, since we only prescribe the first two terms $p + tv$ of a Taylor series, e.g. $c(t) = p + tv + t^2 w$ is another acceptable choice of curve for any w. Thus v corresponds to an equivalence class of curves: two curves c_1, c_2 are equivalent if $c_1(0) = c_2(0)$, $c_1'(0) = c_2'(0)$. We can define the tangent space T_pS as the collection of equivalence classes of smooth curves c in S defined for small times with $c(0) = p$ (the equivalence relation gets checked in any local parametrisation by comparing velocities).

Assuming for simplicity that $F(0,0) = p$, there are two obvious curves in a local parametrisation $t \mapsto (t, 0)$ and $t \mapsto (0, t)$ corresponding to the standard basis vectors $e_1, e_2 \in \mathbb{R}^2 = T_0U$. In T_pS these correspond to the curves $t \mapsto F(t, 0)$ and $t \mapsto F(0, t)$, whose tangent velocity vectors at p are $\partial_x F = DF \cdot e_1$ and $\partial_y F = DF \cdot e_2$. In general, for $F(x_0, y_0) = p$, we can represent the general tangent vector $v = a\partial_x F + b\partial_y F$ by the curve $F(x_0 + at, y_0 + bt)$.

Exercise. If $F, \widetilde{F}$ are two local parametrisations defined near p, let $\tau = \widetilde{F}^{-1} \circ F$ denote the transition map, show that the local curves corresponding to a given curve in S get naturally identified by $\tau : U \to \widetilde{U}$, and that the velocities transform by $D\tau : \mathbb{R}^2 = T_pU \to T_p\widetilde{U} = \mathbb{R}^2$, i.e. by left multiplication by the matrix of partial derivatives of τ.

For smooth maps $\varphi : S_1 \to S_2$ of surfaces, one can also define the derivative map purely in terms of equivalence classes of curves:

$$D_p\varphi \cdot [\text{curve } c(t)] = [\text{curve } \varphi \circ c(t)].$$

Exercise. Check that in local coordinates this corresponds to the matrix of partial derivatives of φ at p, acting by left-multiplication on $c'(0) = v \in \mathbb{R}^2$, thus $D_p\varphi : T_pS_1 \to T_{\varphi(p)}S_2$ is a linear map between the tangent spaces (in local coordinates, $e_1, e_2 \in \mathbb{R}^2$ map to $\partial_x\varphi, \partial_y\varphi \in \mathbb{R}^2$).

Tangent vectors can also be defined as **differential operators** acting on smooth functions. For example, if $f : S \to \mathbb{R}$ is smooth, then locally $e_1 \cdot f$ means the partial derivative $\partial_x f \in \mathbb{R}$.

A tangent vector v acts on a smooth function $f : S \to \mathbb{R}$ by taking the directional derivative

$$v \cdot f = \partial_t|_{t=0}(f \circ c(t)),$$

using any representative curve $c(t)$ for v (exercise: show that the choice of representative does not matter). It tells you how much f varies in the v-direction. Explicitly if $v = a\partial_x F + b\partial_y F$ then $v \cdot f = a\partial_x f_{loc} + b\partial_y f_{loc}$ where $f_{loc}(x, y) = (f \circ F)(x, y)$. For this reason, one often abbreviates the notation by simply writing $v = a\partial_x + b\partial_y$, in particular $\partial_x F = \partial_x$, $\partial_y F = \partial_y$.

9.3 Using the tangent plane to construct local parametrizations

Theorem 9.1. *Let S be a smooth surface in $\mathbb{R}^3$. Near any point $p \in S$, we can locally parametrize S by using the orthogonal projection to the tangent plane T_pS. In this case, S is locally the graph of a function $h : T_pS \to \mathbb{R}$ over the tangent plane (for fixed p).*

Proof. First apply a rotation to $\mathbb{R}^3$ to make the tangent plane T_pS horizontal: so vectors in $T_pS = \mathbb{R}^2 \subset \mathbb{R}^3$ have zero in the third entry. Then the Z coordinate cannot be used together with X or Y to give two local coordinates (e.g. if S were a graph $F(X, Z) = (X, g(X, Z), Z)$ then $\partial_Z F$ would be a tangent vector with a non-zero third entry). Therefore X, Y must be local coordinates and S must be a graph $(X, Y, h(X, Y))$ for a smooth function h (using results from Section 8.2). □

9.4 Vector fields

Definition 9.2 (Vector field)**.** *A **tangent vector field** v on S is a smooth map*

$$v : S \to \mathbb{R}^3 \text{ such that } v(p) \in T_pS \subset \mathbb{R}^3,$$

that is a choice of tangent vector $v(p)$ at each point p of S varying smoothly with $p \in S$.

Locally, we can write:

$$v(x, y) = a(x, y)X_1 + b(x, y)X_2 \qquad (v(x, y) \in T_{F(x,y)}S \subset \mathbb{R}^3),$$

for smooth functions a, b of the local variables x, y, where X_1, X_2 is the basis of $T_{F(x,y)}S$ given by the local vector fields:

$$X_1(x, y) = \partial_x F \qquad X_2(x, y) = \partial_y F.$$

Remark. Vector fields can also be defined for abstract surfaces: $v(x, y) \in \mathbb{R}^2 = T_{(x,y)}V$ is a smooth function $v : V \to \mathbb{R}^2$, and this must transform correctly if we change observer:

$$\widetilde{v}(\tau(x, y)) = (D_{(x,y)}\tau)v(x, y).$$

Example. For the cylinder $X^2 + Y^2 = r^2$, consider the vector fields

$$E_1 = (-\sin\theta, \cos\theta, 0) \qquad E_2 = (0, 0, 1)$$

where $p = F(\theta, Z) = (r\cos\theta, r\sin\theta, Z)$. Notice E_1 points equatorially in the circle direction of the cylinder, and E_2 points in the axis direction. Since $X_1 = (-r\sin\theta, r\cos\theta, 0)$, $X_2 = (0, 0, 1)$,

$$E_1 = \frac{1}{r}X_1 \qquad E_2 = X_2.$$

So a general vector field on the cylinder has the form

$$a(\theta, Z)\, E_1 + b(\theta, Z)\, E_2$$

for any smooth functions a, b which are 2π-periodic in θ.

9.5 Smooth surfaces in $\mathbb{R}^3$: normals and the Gauss map

By Theorem 6.3, a compact smooth surface $S \subset \mathbb{R}^3$ must be orientable. By Lemma 6.5, we can pick a cover of $S = \cup F_i(V_i)$ by local parametrizations $F_i : V_i \to S$ so that on overlaps:

$$\det(DF_j^{-1} \circ DF_i) > 0.$$

Given any point $p \in F_i(V_i)$ we can define a unit normal vector $n(p) \in \mathbb{R}^3$ to S by requiring that the three vectors

$$DF_i \cdot e_1 = \partial_x F_i, \quad DF_i \cdot e_2 = \partial_y F_i, \quad n(p)$$

obey the **right-hand rule**: if $\partial_x F_i, \partial_y F_i$ are the index and middle finger respectively, then $n(p)$ points in the thumb direction.

Explicitly, we take the cross product and normalize:

$$\boxed{n(p) = \frac{\partial_x F_i \times \partial_y F_i}{\|\partial_x F_i \times \partial_y F_i\|} \in \mathbb{R}^3} \qquad \text{(where } \partial_x F_i, \partial_y F_i \text{ are evaluated at } p\text{)}$$

Recall in Section 9.2 we defined the tangent space. Namely (again, evaluating at p):

$$T_pS = \operatorname{span}(\partial_x F_i, \partial_y F_i) = \operatorname{span}(DF_i \cdot e_1, DF_i \cdot e_2) = DF_i \cdot \mathbb{R}^2 \subset \mathbb{R}^3$$

We saw that this vector subspace is independent of the choice of parametrization, since $DF_j \cdot \mathbb{R}^2 = DF_j \cdot (DF_j^{-1} \circ DF_i) \cdot \mathbb{R}^2 = DF_i \cdot \mathbb{R}^2$, using that $DF_j^{-1} \circ DF_i$ is a linear isomorphism $\mathbb{R}^2 \to \mathbb{R}^2$. By construction, the vector $n(p)$ is a unit normal to the 2-dimensional subspace T_pS. Since a plane in $\mathbb{R}^3$ has exactly two unit normals, the question is whether $n(p)$ ever flips if we change parametrizations. But this will never happen because S is orientable: the transitions have $\det(DF_j^{-1} \circ DF_i) > 0$ on overlaps, so the two vectors $\partial_x F_i, \partial_y F_i$ have the same orientation inside T_pS as the two vectors $\partial_x F_j, \partial_y F_j$ (indeed, the two pairs differ by multiplication by the transition map $DF_j^{-1} \circ DF_i$), therefore the cross-product of each pair is oriented in the same direction.

More intuitively said: two observers will agree which plane $T_pS \subset \mathbb{R}^3$ is, they will agree about which ordered basis of T_pS is right-handed, so they will agree about which of the two normals to T_pS gives rise to a right-handed basis for $\mathbb{R}^3$, so they compute the same $n(p)$.

The **Gauss map** is the normal vector field

$$n : S \to \mathbb{R}^3,\ p \mapsto n(p).$$

Remark. The fact that you have a well-defined normal vector field is related to the fact that the field either always points outwards, or always points inwards to the surface.

Example. Let S be the sphere $X^2 + Y^2 + Z^2 = r^2$. For a curve $\gamma(t) = (X(t), Y(t), Z(t)) \in S$, differentiate the defining equation to get:

$$0 = 2XX' + 2YY' + 2ZZ' = (X, Y, Z) \cdot 2(X', Y', Z').$$

For any tangent vector $v \in TS$, there is[1] a curve γ_v in S with that tangent vector $\gamma'(0) = v$. So the above shows that (X, Y, Z) is normal to T_pS (indeed the radial vector is outward and normal to the sphere). Normalizing, we get a Gauss map:

$$n(p) = (X, Y, Z)/r.$$

[1]For F a parametrization, $D_pF : \mathbb{R}^2 \to T_pS = \operatorname{span}(\partial_x F, \partial_y F)$ is surjective. So if $D_{p_0}F(v_F) = v$, then $c_v(t) = p_0 + tv_F \in \mathbb{R}^2$ has $c_v'(0) = v_F$. Then $\gamma_v = F \circ c_v$ works by the chain rule: $\gamma_v'(0) = D_{p_0}F(c_v'(0)) = v$.

Lemma 9.3. *A choice of orientation on a smooth surface $S \subset \mathbb{R}^3$ is the same as a choice of a smooth map*

$$\boxed{\begin{array}{ccc} n : S & \mapsto & S^2 = \{(X,Y,Z) \in \mathbb{R}^3 : X^2+Y^2+Z^2 = 1\} \subset \mathbb{R}^3, \\ p & \mapsto & n(p) \end{array}}$$

such that $n(p)$ is orthogonal to $T_pS \subset \mathbb{R}^3$ for all $p \in S$. More explicitly, a parametrization $F : \mathbb{R}^2 \supset V \to F(V) = U \subset \mathbb{R}^3$ respects the orientation precisely if:

$$n(p) \cdot (\partial_x F \times \partial_y F) \equiv \det \left(\begin{array}{c|c|c} \partial_x F & \partial_y F & n(p) \end{array} \right) > 0.$$

Proof. Declare that a basis $v, w \in T_pS \subset \mathbb{R}^3$ is **right-handed** $\Leftrightarrow v \times w = \lambda n(p)$ for positive $\lambda > 0$ (so in fact, $\lambda = \|v \times w\|$), so $\Leftrightarrow \lambda \equiv n(p) \cdot (v \times w) > 0$. □

Our goal will be to use the Gauss map to define various notions of **curvature** (Gaussian curvature, principal curvatures, mean curvature).

10. Surfaces in $\mathbb{R}^3$: the first fundamental form

10.1 The first fundamental form

Let S be a smooth surface in $\mathbb{R}^3$. Using the dot product[1] $\cdot$ on $\mathbb{R}^3$ we can define an inner product on T_pS called the **first fundamental form**:

$$\boxed{I : T_pS \times T_pS \to \mathbb{R},\ I(v,w) = v \cdot w = v^T w}$$

The properties of an inner product (bilinearity, symmetry, positive-definiteness) all follow from the analogous properties of the dot product.

In a local parametrization $F : \mathbb{R}^2 \supset V \to F(V) = U \subset \mathbb{R}^3$ with $F(p_0) = p$, the vectors $v, w \in T_pS = D_pF(\mathbb{R}^2)$ can be written locally as v_F, w_F, where

$$v = D_{p_0}F(v_F) \qquad w = D_{p_0}F(w_F).$$

So locally the fundamental form is, evaluating at p_0 but omitting that from the notation,

$$\boxed{I_F(v,w) = DF(v_F) \cdot DF(w_F) = v_F^T (DF^T DF)\, w_F}$$

Example. For the standard basis e_1, e_2 of $\mathbb{R}^2$, $I(e_1,e_2) = DF(e_1) \cdot DF(e_2) = \partial_x F \cdot \partial_y F$.

Locally, identifying $T_{p_0}V = \mathbb{R}^2$, the inner product becomes, evaluating at p_0,

$$\boxed{\mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R},\ (v,w) \mapsto v^T A\, w,\ \text{where } A = \begin{pmatrix} e & f \\ f & g \end{pmatrix} = \begin{pmatrix} \partial_x F \cdot \partial_x F & \partial_x F \cdot \partial_y F \\ \partial_y F \cdot \partial_x F & \partial_y F \cdot \partial_y F \end{pmatrix}}$$

indeed, the symmetric matrix $A = DF^T DF$ has entries $A_{ij} = e_i^T A e_j = I_F(e_i, e_j)$, and $DF(e_i)$ is respectively $\partial_x F$ and $\partial_y F$ for $i = 1, 2$.

Example. For the plane $S = \mathbb{R}^2 \subset \mathbb{R}^3$ given by $Z = 0$, and the parametrization $F(r,\theta) = \begin{pmatrix} r\cos\theta \\ r\sin\theta \\ 0 \end{pmatrix}$ we get $\partial_r F = \begin{pmatrix} \cos\theta \\ \sin\theta \\ 0 \end{pmatrix}$ and $\partial_\theta F = \begin{pmatrix} -r\sin\theta \\ r\cos\theta \\ 0 \end{pmatrix}$, thus

$$A = \begin{pmatrix} \cos^2\theta + \sin^2\theta & -r\cos\theta\sin\theta + r\sin\theta\cos\theta \\ -r\cos\theta\sin\theta + r\sin\theta\cos\theta & r^2\sin^2\theta + r^2\cos^2\theta \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix}$$

so $I_F(v,w) = v_1 w_1 + r^2 v_2 w_2$ (written in components, so $v = (v_1, v_2)$, etc.)

[1] Recall, in matrix notation using T for transpose, that $v \cdot w = v^T w$.

10.2 Change in local first fundamental form under coordinate changes

How does the local expression A of the first fundamental form depend on the observer? Consider the picture in Section 9.2. Let $F, \widetilde{F}$ be two local parametrizations, giving fundamental forms $A, \widetilde{A}$. Let

$$\tau = \widetilde{F}^{-1} \circ F : V \to \widetilde{V}$$

be the transition map (defined on the overlap $F^{-1}(\widetilde{U}) \subset V$).

The local tangent spaces get identified via the linear isomorphism

$$D\tau = D\widetilde{F}^{-1} \circ DF : T_{p_0}V \to T_{\widetilde{p}_0}\widetilde{V}.$$

Explicitly, in local coordinates: $\tau(x,y) = \begin{pmatrix} \widetilde{x}(x,y) \\ \widetilde{y}(x,y) \end{pmatrix}$ and $D\tau = \begin{pmatrix} \partial_x\widetilde{x} & \partial_y\widetilde{x} \\ \partial_x\widetilde{y} & \partial_y\widetilde{y} \end{pmatrix}$.

So we expect that $v^T A w = I(v,w) = (D\tau(v))^T \widetilde{A}(D\tau(w)) = v^T (D\tau)^T \widetilde{A}(D\tau)w$. Indeed:

$$\boxed{\begin{array}{rcl} A & = & DF^T DF \\ & = & DF^T (D\widetilde{F}^T)^{-1}(D\widetilde{F}^T D\widetilde{F}) D\widetilde{F}^{-1} DF \\ & = & (D\tau)^T \widetilde{A}\,(D\tau) \end{array}}$$

Often, in practice, you will want to rewrite this as: $\boxed{\widetilde{A} = (D\tau^{-1})^T A\,(D\tau^{-1})}$.

Example. For the plane $S = \mathbb{R}^2 \subset \mathbb{R}^3$, we could use the obvious $F(x,y) = (x,y,0)$ or polar coordinates $\widetilde{F}(r,\theta) = (r\cos\theta, r\sin\theta, 0)$. Since $\tau^{-1}(r,\theta) = (x,y) = (r\cos\theta, r\sin\theta)$,

$$(D\tau)^{-1} = D(\tau^{-1}) = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}.$$

Now $I_F(v,w) = v \cdot w$ and $A = I$, so we confirm the result of the previous example:

$$\widetilde{A} = (D\tau^{-1})^T A\,(D\tau^{-1}) = \begin{pmatrix} \cos\theta & \sin\theta \\ -r\sin\theta & r\cos\theta \end{pmatrix} \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix}.$$

Thus the two local expressions of the first fundamental form are:

$$I_{\widetilde{F}}(\widetilde{v},\widetilde{w}) = \widetilde{v}_1\widetilde{w}_1 + r^2\widetilde{v}_2\widetilde{w}_2 = v_1w_1 + v_2w_2 = I_F(v,w).$$

10.3 Application 1: the length of a curve and isometric surfaces

A **smooth curve** in S is a smooth map $\gamma : [a,b] \to S$. This is the same as saying that $\gamma : [a,b] \to \mathbb{R}^3$ is smooth and $\gamma(t) \in S$ for all $t \in [a,b]$. The **length of the curve** (induced by the norm on $\mathbb{R}^3$) is

$$\boxed{L(\gamma) = \int_a^b \|\gamma'(t)\|\,dt}$$

Lemma 10.1. *The length of a curve is independent of the choice of time-parametrization.*

Proof. Let $\mu(t) = \gamma(s(t))$ be a time-reparametrization of γ, so $s : [A,B] \to [a,b]$ is a smooth strictly increasing function. Integrating by substitution (change of variables) using $s' > 0$:

$$L(\mu) = \int_A^B \|\mu'(t)\|\,dt = \int_A^B \|\gamma'(s(t))\|\,s'(t)\,dt = \int_a^b \|\gamma'(s)\|\,ds = L(\gamma). \quad \square$$

Taking $a = 0$, we say a curve $\gamma : [0,b] \to S$ is **parametrized by arc-length** if $L(\gamma|_{[0,t]}) = t$ for all t. By differentiating in t, this is equivalent to having unit speed: $\|\gamma'(t)\| = 1$ for all t.

Lemma 10.2. *Every smooth curve with $\gamma'(t) \neq 0$ for all t can be parametrized by arc-length.*

Proof. Let $s(t) = \int_0^t \|\gamma'(t)\|\,dt$. Since $s' > 0$, s is invertible and $(s^{-1})'(t) = 1/s'(s^{-1}(t)) = 1/\|\gamma'(s^{-1}(t))\|$. Hence $\mu(t) = \gamma(s^{-1}(t))$ works: $\|\mu'(t)\| = \|\gamma'(s^{-1}(t))\|(s^{-1})'(t) = 1$. $\square$

Notice that for any smooth curve γ, the velocity vector always lies in the tangent space

$$\boxed{\gamma'(t) \in T_{\gamma(t)}S}$$

Indeed: write γ locally using the parametrization F, say $\gamma_{loc}(t) \in V \subset \mathbb{R}^2$, then $\gamma(t) = F \circ \gamma_{loc}(t)$, thus $\gamma'(t) = D_{\gamma_{loc}(t)}F \cdot \gamma'_{loc}(t) \in D_{\gamma_{loc}(t)}F(\mathbb{R}^2) = T_{\gamma(t)}S$.

Theorem 10.3. *Lengths of curves in S are determined by the first fundamental form.*

Proof. $L(\gamma) = \int_a^b \|\gamma'(t)\|\,dt = \int_a^b \sqrt{\gamma'(t)\cdot\gamma'(t)}\,dt = \int_a^b \sqrt{I(\gamma'(t),\gamma'(t))}\,dt$. □

Assuming that γ lies entirely in the parametrization patch U (we can add local contributions by covering $\gamma[a,b]$ by several parametrization patches), we can be even more explicit using the matrix $A = \left(\begin{smallmatrix} e & f \\ f & g \end{smallmatrix}\right)$ defined previously and writing $\gamma_{loc}(t) = (x(t), y(t)) \in V \subset \mathbb{R}^2$:

$$L(\gamma) = \int_a^b \sqrt{e\left(\tfrac{dx}{dt}\right)^2 + 2f(\tfrac{dx}{dt})(\tfrac{dy}{dt}) + g(\tfrac{dy}{dt})^2}\,dt$$

where $e = e(x(t), y(t))$, etc. depend on the local coordinates, so depend on t.

Example. Letting $\gamma_{loc}(t) = p_0 + (t, 0)$ for $t \in [0, \varepsilon]$, with $\varepsilon \geq 0$ a variable, by the fundamental theorem of calculus:

$$\frac{d}{d\varepsilon}L(\gamma_{loc}) = \frac{d}{d\varepsilon}\int_0^\varepsilon \sqrt{e(p_0 + (t,0))}\,dt = \sqrt{e(p_0 + (\varepsilon, 0))},$$

from which we recover $e(p_0)$ by taking $\varepsilon = 0$. So if we know the values $L(\gamma)$ of lengths of all curves in V, we recover the values of e on V.

Exercise. For any continuous real-valued function f prove that $\lim \frac{1}{\varepsilon}\int_0^\varepsilon f(t)\,dt = f(0)$, as $\varepsilon \to 0$. Deduce that $\lim \frac{1}{\varepsilon}L(\gamma_{loc}) = \sqrt{e(p_0)}$ for the curve in the example.

Theorem 10.4. *Lengths of curves determine I locally.*

Proof. The example above recovered e. Similarly, we recover g by considering $\gamma_{loc}(t) = p_0 + (0, t)$ and we recover f by using $\gamma_{loc}(t) = p_0 + (t, t)$ (in the latter case $\frac{d}{d\varepsilon}\big|_{\varepsilon=0} L(\gamma_{loc}) = \sqrt{e(p_0) + 2f(p_0) + g(p_0)}$, but we know e, g so we recover f). □

Definition 10.5 (Isometric surfaces)**.** *Two surfaces S_1, S_2 in $\mathbb{R}^3$ are **isometric** if there is a diffeomorphism $\varphi : S_1 \to S_2$ preserving lengths of curves: $L(\varphi \circ \gamma) = L(\gamma)$ for $\gamma \subset S_1$.*

Theorem 10.6. *Two surfaces S_1, S_2 in $\mathbb{R}^3$ are locally isometric near p_1, p_2 if and only if there are local parametrizations $F_1 : V \to U_1$, $F_2 : V \to U_2$ near p_1, p_2 yielding the same local first fundamental form $I_{F_1} = I_{F_2}$ on V.*

Proof. ($\Leftarrow$): is immediate from the local expression of I, and the local calculation of $L(\gamma)$. ($\Rightarrow$): take $F_2 = \varphi \circ F_1$ where φ is the local iso, and apply Theorem 10.4. □

Example. Suppose we have a cone made out of paper and we cut out a straight ray from the vertex. When we unfold this piece of paper we get a pie-sliced piece of paper. So these two surfaces are obviously isometric. Let's prove it. Consider the cone $S = \{(X, Y, Z) \in \mathbb{R}^3 : X^2 + Y^2 = a^2Z^2,\ Z > 0\}$ with angle $\tan^{-1}(a)$ to the axis. Calculate:

$$F(x,y) = \begin{pmatrix} ax\cos y \\ ax \sin y \\ x \end{pmatrix} \quad \partial_x F = \begin{pmatrix} a\cos y \\ a \sin y \\ 1 \end{pmatrix} \quad \partial_y F = \begin{pmatrix} -ax\sin y \\ ax\cos y \\ 0 \end{pmatrix} \quad A = \begin{pmatrix} 1+a^2 & 0 \\ 0 & a^2x^2 \end{pmatrix}$$

Remove the line $(-aX, 0, X)_{X\in\mathbb{R}}$ from S: F parametrizes $S \setminus (\text{line})$ for $(x,y) \in V = (0,\infty)\times(-\pi,\pi)$. We claim $S\setminus(\text{line})$ is isometric to a pie-shape in $\mathbb{R}^2$ bounded by two rays. Parametrize a pie-shape by $(x,y)\mapsto (xb\cos(cy), xb\sin(cy), 0)\in\mathbb{R}^3$ for $(x,y)\in V$. To get the same A let $b=(1+a^2)^{1/2}$, $c=a/b$ (note: $c\le\frac{1}{2}$).

10.4 Quadratic form for I, differentials, a fast change of coordinates

It is convenient to abbreviate the quadratic form corresponding to I, written locally, by:

$$\boxed{I = e\,dx^2 + 2f\,dx\,dy + g\,dy^2 = (\partial_x F\cdot\partial_x F)\,dx^2 + 2(\partial_x F\cdot\partial_y F)\,dx\,dy + (\partial_y F\cdot\partial_y F)\,dy^2}$$

where e, f, g are functions of the coordinates $(x,y)\in V\subset\mathbb{R}^2$. The symbols dx, dy are called **differentials** (or **differential 1-forms**, or **covectors**). They are elements of the dual vector space

$$(T_pS)^* = \{\text{linear functions } T_pS\to\mathbb{R}\}$$

called **cotangent space**. Locally, dx, dy are the dual basis of the standard basis $e_1, e_2\in\mathbb{R}^2 = T_{p_0}V$. Explicitly, if $v=(v_1,v_2)\in\mathbb{R}^2=T_{p_0}V$:

$$dx : T_{p_0}V\to\mathbb{R},\ dx(v)=v_1, \qquad dy : T_{p_0}V\to\mathbb{R},\ dy(v)=v_2.$$

Example. As before, writing $\gamma_{loc}(t) = (x(t), y(t))$,

$$dx(\gamma'_{loc}(t)) = \frac{dx}{dt}, \qquad dy(\gamma'_{loc}(t)) = \frac{dy}{dt}.$$

One often denotes the standard basis $e_1, e_2\in\mathbb{R}^2 = T_{p_0}V$ by the symbols $e_1 = \frac{\partial}{\partial x}, e_2 = \frac{\partial}{\partial y}$, and so the condition of being a dual basis are the memorable looking formulas

$$dx(\tfrac{\partial}{\partial x}) = 1,\ dx(\tfrac{\partial}{\partial y}) = 0,\ dy(\tfrac{\partial}{\partial x}) = 0,\ dy(\tfrac{\partial}{\partial y}) = 1.$$

Other formulas also become more memorable: $DF(\frac{\partial}{\partial x}) = \partial_x F$, $DF(\frac{\partial}{\partial y}) = \partial_y F$.

The change of I under changes of coordinates also becomes easier in this notation. Writing $\widetilde{\gamma}_{loc}(t) = (\widetilde{x}(t), \widetilde{y}(t))$ in the parametrization $\widetilde{F}$, by the chain rule

$$\frac{d\widetilde{x}}{dt} = \partial_x\widetilde{x}\,\frac{dx}{dt} + \partial_y\widetilde{x}\,\frac{dy}{dt}, \qquad \frac{d\widetilde{y}}{dt} = \partial_x\widetilde{y}\,\frac{dx}{dt} + \partial_y\widetilde{y}\,\frac{dy}{dt}$$

therefore

$$d\widetilde{x} = \partial_x\widetilde{x}\,dx + \partial_y\widetilde{x}\,dy, \qquad d\widetilde{y} = \partial_x\widetilde{y}\,dx + \partial_y\widetilde{y}\,dy.$$

Example. For the plane $S=\mathbb{R}^2\subset\mathbb{R}^3$, $F(r,\theta) = (r\cos\theta, r\sin\theta, 0) = (x,y,0) = \widetilde{F}(x,y)$,

$$dx = \cos\theta\,dr - r\sin\theta\,d\theta, \qquad dy = \sin\theta\,dr + r\cos\theta\,d\theta.$$

Recall that for τ the transition, $D\tau = \begin{pmatrix}\partial_x\widetilde{x} & \partial_y\widetilde{x}\\ \partial_x\widetilde{y} & \partial_y\widetilde{y}\end{pmatrix}$, so

$$\begin{pmatrix} d\widetilde{x}\\ d\widetilde{y}\end{pmatrix} = D\tau\cdot\begin{pmatrix} dx\\ dy\end{pmatrix}$$

Let's check this is consistent with how the local quadratic form I changes, in matrix notation:

$$(\,d\widetilde{x}\ d\widetilde{y}\,)\begin{pmatrix}\widetilde{e} & \widetilde{f}\\ \widetilde{f} & \widetilde{g}\end{pmatrix}\begin{pmatrix} d\widetilde{x}\\ d\widetilde{y}\end{pmatrix} = \begin{pmatrix} dx\\ dy\end{pmatrix}^T (D\tau)^T\begin{pmatrix}\widetilde{e} & \widetilde{f}\\ \widetilde{f} & \widetilde{g}\end{pmatrix} D\tau\begin{pmatrix} dx\\ dy\end{pmatrix} = (\,dx\ dy\,)\begin{pmatrix} e & f\\ f & g\end{pmatrix}\begin{pmatrix} dx\\ dy\end{pmatrix}.$$

So if you happen to know $I = e\,dx^2 + 2f\,dx\,dy + g\,dy^2$ and you want to compute I in the other coordinates, $I = \widetilde{e}\,d\widetilde{x}^2 + 2\widetilde{f}\,d\widetilde{x}\,d\widetilde{y} + \widetilde{g}\,d\widetilde{y}^2$, then simply write $dx = \cdots$, $dy = \cdots$ in terms of $d\widetilde{x}$, $d\widetilde{y}$, then simply substitute and formally square/multiply.

Example. For the plane $S=\mathbb{R}^2\subset\mathbb{R}^3$, $F(x,y) = (x,y,0)$ and $\widetilde{F}(r,\theta) = (r\cos\theta, r\sin\theta, 0)$:

$$I = dx^2 + dy^2 = (\cos\theta\,dr - r\sin\theta\,d\theta)^2 + (\sin\theta\,dr + r\cos\theta\,d\theta)^2 = dr^2 + r^2\,d\theta^2.$$

10.5 Examples of calculations of I

Sphere of radius a: $\boxed{I = a^2 \sin^2 y\, dx^2 + a^2 dy^2}$ using:

$$F(x,y) = \begin{pmatrix} a\cos x \sin y \\ a \sin x \sin y \\ a \cos y \end{pmatrix} \quad \partial_x F = \begin{pmatrix} -a \sin x \sin y \\ a \cos x \sin y \\ 0 \end{pmatrix} \quad \partial_y F = \begin{pmatrix} a \cos x \cos y \\ a \sin x \cos y \\ -a \sin y \end{pmatrix}$$

Cylinder of radius a: $\boxed{I = dx^2 + a^2 dy^2}$ using:

$$F(x,y) = \begin{pmatrix} a\cos y \\ a \sin y \\ x \end{pmatrix} \quad \partial_x F = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad \partial_y F = \begin{pmatrix} -a \sin y \\ a \cos y \\ 0 \end{pmatrix}$$

Cone with angle $\tan^{-1}(a)$ to the axis: $\boxed{I = (1+a^2)dx^2 + a^2 x^2 dy^2}$ using:

$$F(x,y) = \begin{pmatrix} ax\cos y \\ ax \sin y \\ x \end{pmatrix} \quad \partial_x F = \begin{pmatrix} a \cos y \\ a \sin y \\ 1 \end{pmatrix} \quad \partial_y F = \begin{pmatrix} -ax \sin y \\ ax \cos y \\ 0 \end{pmatrix}$$

Surface of revolution, rotate $Y = f(Z)$ about Z-axis: $\boxed{I = (1 + f'(x)^2)\, dx^2 + f(x)^2 dy^2}$ using:

$$F(x,y) = \begin{pmatrix} f(x)\cos y \\ f(x) \sin y \\ x \end{pmatrix} \quad \partial_x F = \begin{pmatrix} f'(x) \cos y \\ f'(x) \sin y \\ 1 \end{pmatrix} \quad \partial_y F = \begin{pmatrix} -f(x) \sin y \\ f(x) \cos y \\ 0 \end{pmatrix}$$

10.6 A substantial example: ruled surfaces in $\mathbb{R}^3$

Remark. You do not need to memorize for exams the terminology or formulas that appear in this example, it is only supposed to be an interesting example to see.

Recall a ruled surface S is swept out by lines $p(t) + \mathbb{R}n(t)$ along a curve $t \mapsto p(t)$, where $n(t)$ is the unit direction of the line at time t. So we can parametrize S by:

$$\boxed{F(x,y) = p(x) + yn(x)}$$

As written, S may self-intersect, and even worse S may not be smooth locally.
Example. Let S be the double cone $\{(X,Y,Z) \in \mathbb{R}^2 : X^2 + Y^2 = a^2 Z^2\}$. This arises as a ruled surface $p(x) + yn(x)$ taking:

$$p(x) = \begin{pmatrix} a\cos x \\ a \sin x \\ 1 \end{pmatrix} \qquad n(x) = \frac{p(x)}{\|p(x)\|} = \frac{p(x)}{\sqrt{1+a^2}} \qquad p(x)+yn(x) = (1+\tfrac{y}{\sqrt{1+a^2}}) \begin{pmatrix} a\cos x \\ a \sin x \\ 1 \end{pmatrix}$$

S does not self-intersect, but it fails to be locally smooth at the vertex $(0,0,0)$.
In general S is a smooth surface near $F(x,y) \Leftrightarrow \partial_x F, \partial_y F$ are linearly independent.

$$\partial_x F = p'(x) + yn'(x) \qquad \text{and} \qquad \partial_y F = n(x).$$

So smoothness at $F(x,y)$ is equivalent to linear independence of $p'(x)+yn'(x)$ and $n(x)$, which is equivalent to the non-vanishing of the cross-product:[1]

$$(p'(x)+yn'(x))\times n(x)\neq 0.$$

We often use, without mentioning, the basic trick:

Trick: If $n(t)\in\mathbb{R}^3$ has unit norm $\|n(t)\|=1$, then the velocity $n'(t)$ is perpendicular to the curve $n(t)$. *Proof:* $0=\frac{d}{dt}(1)=\frac{d}{dt}(n(t)\cdot n(t))=2n'(t)\cdot n(t)$. $\square$

Since $n(x)\cdot n(x)=1$, we get $n'(x)\cdot n(x)=0$, thus:

$$\boxed{I=(\|p'\|^2+y^2\|n'\|^2+2yp'\cdot n')\,\mathbf{dx^2}+2(p'\cdot n)\,\mathbf{dx\,dy}+\mathbf{dy^2}}$$

Example. For the double cone S from the previous example,

$$p'(x)=\begin{pmatrix}-a\sin x\\ a\cos x\\ 0\end{pmatrix}\qquad n'(x)=\frac{p'(x)}{\sqrt{1+a^2}}$$

so: $\|p'\|^2=a^2$, $\|n'\|^2=\frac{\|p'\|^2}{1+a^2}=\frac{a^2}{1+a^2}$, $p'\cdot n'=\frac{\|p'\|^2}{\sqrt{1+a^2}}=\frac{a^2}{\sqrt{1+a^2}}$, $p'\cdot n=\frac{p'\cdot p}{\sqrt{1+a^2}}=0$. Substituting in the formula above we obtain:

$$I=(a+\tfrac{ay}{\sqrt{1+a^2}})^2\,dx^2+dy^2.$$

The general formula for I becomes $I=(\|q'\|^2+y^2\|n'\|^2)\,\mathbf{dx^2}+2(q'\cdot n)\,\mathbf{dx\,dy}+\mathbf{dy^2}$ if one replaces $p(x)$ by a clever choice of curve $q(x)$ called **line of striction** which satisfies

$$q'(x)\cdot n'(x)=0.$$

Let's find $q(x)$. We want a curve

$$q(x)=p(x)+y(x)n(x)\in S$$

such that $q'(x)\cdot n'(x)=0$. Notice we can then replace p by q because the surface is made up of the same straight lines $p(x)+\mathbb{R}n(x)=q(x)+\mathbb{R}n(x)$. Compute:

$$q'\cdot n'=(p'+y'n+yn')\cdot n'=p'\cdot n'+y\|n'\|^2$$

so taking $y(x)=-\frac{p'(x)\cdot n'(x)}{\|n'(x)\|^2}$ works. Thus:

$$q(x)=p(x)-\frac{p'(x)\cdot n'(x)}{\|n'(x)\|^2}\,n(x)\qquad S=\{q(x)+yn(x)\in\mathbb{R}^3:x,y\in\mathbb{R}\}$$

This is well-defined provided we assume that $n'(x)\neq 0$ for all x. The ruled surface is called **non-cylindrical** if it satisfies $n'(x)\neq 0$ for all x. One can break up a ruled surface into pieces where this condition holds, and then separately study the cylindrical pieces where $n'=0$ for an interval of values of x (these pieces are very simple: $n(x)$ is constant in x, so we just draw parallel lines through the points $p(x)$ in the constant direction $n(x)$).

Example. For the double cone S from the previous example,

$$q(x)=\begin{pmatrix}a\cos x\\ a\sin x\\ 1\end{pmatrix}-\frac{\frac{a^2}{\sqrt{1+a^2}}}{\frac{a^2}{1+a^2}}\frac{1}{\sqrt{1+a^2}}\begin{pmatrix}a\cos x\\ a\sin x\\ 1\end{pmatrix}=0.$$

[1] Recall the cross-product in the standard basis $\mathbf{i},\mathbf{j},\mathbf{k}$ of $\mathbb{R}^3$ is:

$$a\times b=\det\begin{pmatrix}a_1 & b_1 & \mathbf{i}\\ a_2 & b_2 & \mathbf{j}\\ a_3 & b_3 & \mathbf{k}\end{pmatrix}$$

which points in the right-hand thumb direction if a is the index and b is the middle finger, and which has length $\|a\times b\|=\|a\|\,\|b\|\,|\sin\theta|$ if θ is the angle between a,b.

So the line of striction is the constant curve at the vertex $(0,0,0)$.

The condition $q' \cdot n' = 0$, makes it is also easy to check where S is a locally smooth surface. We require the non-vanishing of the cross-product $(q'(x) + yn'(x)) \times n(x) \neq 0$. Since n' is perpendicular to both q', n, we deduce[1] that $q' \times n = \lambda n'$ for some $\lambda = \lambda(x) \in \mathbb{R}$. Thus:

$$\|(q' + yn') \times n\|^2 = \|\lambda n' + yn' \times n\|^2 = \lambda^2 \|n'\|^2 + y^2 \|n' \times n\|^2 = (\lambda^2 + y^2)\, \|n'\|^2,$$

where we used that $n', n' \times n$ are perpendicular, and some cross-product tricks.[2]

Theorem 10.7. *For non-cylindrical ruled surfaces $p(x) + yn(x)$ (meaning $n'(x) \neq 0$ for all x), one can parametrize S by $q(x) + yn(x)$ with $q'(x) \cdot n'(x) = 0$, in which case $q(x)$ is called the* ***line of striction****. Moreover S is locally smooth everywhere except at those points on the line of striction where q', n become linearly dependent. The first fundamental form is*

$$I = \begin{pmatrix} \|q'\|^2 + y^2 \|n'\|^2 & (q' \cdot n) \\ (q' \cdot n) & 1 \end{pmatrix}.$$

Proof. This follows by the above calculation, since $\lambda^2 + y^2 = 0$ if and only if $y = 0$ and $\lambda = 0$, and the latter implies $q' \times n = 0$, equivalently: q', n are linearly dependent. □

Example. For the double cone S, the line of striction is $q(x) = (0,0,0)$, and $q'(x) \times n(x) = 0$ since $q' = 0$. So $(0,0,0)$ is the only singular point (as expected).

10.7 Application 2: the angle between curves in a surface

The angle θ between two vectors $v, w \in TS \subset \mathbb{R}^3$ satisfies $v \cdot w = \|v\|\, \|w\| \cos\theta$, therefore

$$\boxed{\cos\theta = \frac{v \cdot w}{\|v\|\, \|w\|} = \frac{I(v,w)}{\sqrt{I(v,v)}\, \sqrt{I(w,w)}}}$$

which only depends on I (and the vectors v, w). Notice this can be used to measure angles between intersecting curves: if γ_1, γ_2 are smooth curves in S which intersect at $p = \gamma_1(t) = \gamma_2(s)$, then we can measure the angle between their tangent vectors $v = \gamma_1'(t)$ and $w = \gamma_2'(s)$.

10.8 Application 3: the area of a region in a surface

Motivation. Consider the region near p where the surface S is parametrized by F. The region is approximated by infinitesimal parallelograms with edges the vectors $(\partial_x F)\, dx, (\partial_y F)\, dy$ where we think of dx, dy as infinitesimal increments of x, y. The area of this parallelogram is

$$(\|\partial_x F\|\, dx)\, (\|\partial_y F\|\, dy)\, |\sin\theta| = \|\partial_x F \times \partial_y F\|\, dx\, dy,$$

where θ is the angle between the two edges. Using the following rule about cross-products:[3]

$$(a \times b) \cdot (a \times b) = (a \cdot a)(b \cdot b) - (a \cdot b)^2,$$

we obtain, in terms of the first fundamental form $I = e\, dx^2 + 2f\, dx\, dy + g\, dy^2 = \left(\begin{smallmatrix} e & f \\ f & g \end{smallmatrix}\right)$,

$$\|\partial_x F \times \partial_y F\|\, dx\, dy = \sqrt{eg - f^2}\, dx\, dy = \sqrt{\det(I)}\, dx\, dy.$$

---

[1] This follows for $\lambda \neq 0$ when q', n are linearly independent, and when they are dependent just take $\lambda = 0$.

[2] Using the cyclic symmetry

$$c \cdot (a \times b) = \det \begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix} = a \cdot (b \times c)$$

we get $(n' \times n) \cdot (n' \times n) = n' \cdot (n \times (n' \times n)) = n' \cdot n'$ (the last equality follows because the directions agree, and the lengths agree using $\|a \times b\| = \|a\|\, \|b\|\, |\sin\theta|$, or alternatively use: $a \times (b \times c) = (a \cdot c)b - (a \cdot b)c$).

[3] coming from the more general rule: $(a \times b) \cdot (c \times d) = (a \cdot c)(b \cdot d) - (a \cdot d)(b \cdot c)$.

As mathematicians, we just turn this into a definition:

Definition 10.8 (Area of a region in a surface)**.** *The area of $R = F(V) \subset S$ is defined as*

$$\text{Area}(R) = \int_V \|\partial_x F \times \partial_y F\|\, dx\, dy = \int_V \sqrt{eg - f^2}\, dx\, dy = \int_V \sqrt{\det(I)}\, dx\, dy.$$

More generally, for any open region $R \subset S$, we cover R by (closures of) such sets and add the areas.

Theorem 10.9. *The area is well-defined independently of choices of parametrization.*

Proof. Suppose (restricting to an overlap) $F : V \to S$, $\widetilde{F} : V \to S$ are two parametrizations, yielding local first fundamental forms $A, \widetilde{A}$. Using the transition τ,

$$\sqrt{\det(A)} = \sqrt{\det((D\tau)^T \widetilde{A}(D\tau))} = |\det\, D\tau| \sqrt{\det(\widetilde{A})}.$$

Recall (see the Analysis handout) the change of variables formula for integrals will replace $dx\, dy$ by $|\det\, D\tau^{-1}|\, d\widetilde{x}\, d\widetilde{y}$, thus $|\det\, D\tau|\,|\det\, D\tau^{-1}| = 1$ disappears when we integrate. □

Example. For the surface of revolution $F(x, y) = (f(x)\cos y, f(x)\sin y, x)$ (where $f(x) > 0$) we got $I = (1 + f'(x)^2)\, dx^2 + f(x)^2 dy^2$, so the area for $(x, y) \in [a, b] \times [0, 2\pi]$ is the familiar

$$\int_0^{2\pi} \int_a^b f(x)\sqrt{1 + f'(x)^2}\, dx\, dy = \int_a^b 2\pi f(x)\sqrt{1 + f'(x)^2}\, dx.$$

10.9 Riemannian metric: first fundamental forms for abstract surfaces

For any surface (or submanifold) in $\mathbb{R}^n$ one can define a first fundamental form by using the dot product on $\mathbb{R}^n$. However, for an abstract smooth surface S, there is no preferred way to embed it in $\mathbb{R}^n$, so there is no preferred inner product on T_pS.

Definition 10.10 (Riemannian metric)**.** *A Riemannian metric for a surface (or manifold) S is a well-defined inner product on each tangent space T_pS, which in local coordinates depends smoothly on $p \in S$.*

Let's unpack this definition. By tangent space T_pS we mean locally $T_{p_0}V = \mathbb{R}^2$ for a parametrization $F : V \to S$ subject to the rule that if we change parametrization to $\widetilde{F}$ then we identify the local tangent spaces using the derivative of the transition $\tau = \widetilde{F}^{-1} \circ F$:

$$T_{p_0}V = \mathbb{R}^2 \xrightarrow{D\tau} \mathbb{R}^2 = T_{\widetilde{p}_0}\widetilde{V}.$$

The Riemannian metric is locally given by an inner product

$$T_{p_0}V \times T_{p_0}V = \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R},\ (v, w) \mapsto v^T A w \quad \text{where} \quad A = \begin{pmatrix} e & f \\ f & g \end{pmatrix},$$

and we require that $A = A(p_0)$ depends smoothly on p_0, that is: the functions e, f, g are smooth functions of the local coordinates $p_0 = (x_0, y_0)$.

In order for this to be an inner product, we need: bilinearity (automatic), symmetry (automatic), and positive definiteness which by linear algebra is ensured by the conditions

$$\boxed{e > 0 \quad \text{and} \quad eg - f^2 > 0} \qquad \text{(it follows that also } g > 0\text{)}.$$

Asking that the inner product is well-defined means that we want observers to agree on what inner product is being used having identified their local tangent spaces by $D\tau$ as mentioned above. So by Section 10.2 we require:

$$\boxed{A = (D\tau)^T \widetilde{A}\, (D\tau)}$$

It follows that, once a Riemannian metric is chosen on S (for example, the first fundamental form obtained from a particular embedding $S \to \mathbb{R}^n$), we can define as before: lengths of smooth curves, angles between intersecting curves, and areas of open sets in S.

Examples.

(1) On the torus $\mathbb{R}^2/(\mathbb{Z}\omega_1 + \mathbb{Z}\omega_2)$ we can use the Riemannian metric [1]
$$I = dx^2 + dy^2 = dz\, d\overline{z} = |dz|^2$$
but there are lots of other choices: e.g. rescale the above by any (smooth) strictly positive doubly periodic function, meaning $f(x+\omega, y+\omega) = f(x,y)$ for $\omega \in \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$.

(2) On the upper half-plane $\mathbb{H} = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$, the **hyperbolic metric** is the Riemannian metric
$$I = \frac{dx^2 + dy^2}{y^2} = \frac{dz\, d\overline{z}}{\mathrm{Im}(z)^2} = \frac{|dz|^2}{\mathrm{Im}(z)^2}$$

(3) On the unit disc $D = \{z \in \mathbb{C} : |z| < 1\}$, the **hyperbolic metric** is the Riemannian metric
$$I = \frac{4(dx^2 + dy^2)}{(1 - x^2 - y^2)^2} = \frac{4\, dz\, d\overline{z}}{(1-|z|^2)^2} = \frac{4\,|dz|^2}{(1-|z|^2)^2}$$

Remark. (Non-examinable) A connected[2] surface with a Riemannian metric is a metric space, by defining the **distance function** $d : S \times S \to \mathbb{R}$ by letting $d(p,q)$ be the infimum of the lengths of all smooth curves from p to q (you can also allow piecewise smooth curves without affecting d). In particular, the open balls for this metric are a basis for the topology of S.

Given two smooth surfaces S_1, S_2 with choices of Riemannian metric, we say S_1, S_2 are **isometric** if there is a diffeomorphism preserving lengths of curves.

Example. Recall from Section 3.1 that there are biholomorphisms
$$D \to \mathbb{H},\ z \mapsto \tau(z) = \frac{iz+i}{-z+1} \qquad \mathbb{H} \to D,\ z \mapsto \tau^{-1}(z) = \frac{z-i}{z+i}.$$
Let's check this is an isometry if we use the hyperbolic metrics from the previous example. The above specifies a change of coordinates $z \mapsto \widetilde{z} = \tau(z)$. Rather than switching to real coordinates, recall that differentials change by multiplication by $D\tau$, and since we may identify $D\tau$ with $\tau'(z)$ when identifying $\mathbb{R}^2 \equiv \mathbb{C}$, we deduce:[3]
$$d\widetilde{z} = \tau'(z)\, dz.$$
Now calculate:
$$\tau'(z) = \frac{i(-z+1) - (iz+i)(-1)}{(-z+1)^2} = \frac{2i}{(z-1)^2}$$
$$\mathrm{Im}(\widetilde{z}) = \mathrm{Im}(\tau(z)) = \mathrm{Im}\frac{(iz+i)(-\overline{z}+1)}{|z-1|^2} = \mathrm{Im}\frac{i(-|z|^2 + 2i\mathrm{Im}\, z + 1)}{|z-1|^2} = \frac{1-|z|^2}{|z-1|^2}$$
$$I_{\mathbb{H}} = \frac{|d\widetilde{z}|^2}{\mathrm{Im}(\widetilde{z})^2} = \frac{4}{|z-1|^4}\frac{|z-1|^4}{(1-|z|^2)^2}|dz|^2 = \frac{4|dz|^2}{(1-|z|^2)^2} = I_D.$$

---

[1] where $z = x + iy$, $dz = dx + i\, dy$, $\overline{z} = x - iy$, $d\overline{z} = dx - idy$. Here $dz : T_pS \to \mathbb{C}$ is again viewed as a linear functional on the tangent space of the given surface S.

[2] For surfaces (and manifolds) connectedness implies path-connectedness, so there is always a continuous path $[0,1] \to S$ between two given points p, q. Using local parametrisations one can approximate any continuous path by a piecewise smooth path, and then one can round off corners to get a smooth path.

[3] I'm saying we can identify $\begin{pmatrix} dx \\ dy \end{pmatrix} \equiv dz$ and hence $d\widetilde{z} = \begin{pmatrix} d\widetilde{x} \\ d\widetilde{y} \end{pmatrix} = D\tau \cdot \begin{pmatrix} dx \\ dy \end{pmatrix} \equiv \tau'(z)\, dz$. Although we won't need it, you may be curious how $d\overline{z}$ changes: the rule is as expected: $d\widetilde{\overline{z}} = \overline{\tau'(z)}\, d\overline{z}$. Indeed as a linear function, $d\overline{z} = dx - i\, dy$ maps $\partial_x \mapsto 1, \partial_y \mapsto -i$, whereas $dz = dx + i\, dy$ maps $\partial_x \mapsto 1, \partial_y \mapsto i$. So $d\overline{z}$ is the conjugate of the linear function dz. So $d\widetilde{\overline{z}} = \overline{d\widetilde{z}} = \overline{\tau'(z)\, dz} = \overline{\tau'(z)}\, d\overline{z}$

11. Surfaces in $\mathbb{R}^3$: the second fundamental form

11.1 A basic toy model: what is the curvature of a curve?

Let γ be a smooth curve in $\mathbb{R}^3$, $\gamma : [0, b] \to \mathbb{R}^3$ parametrized by arc-length (i.e. unit speed: $\|\gamma'(t)\| = 1$, see Sec.10.3). Then the **unit tangent vector** is the velocity $\gamma'(t)$, and the **curvature** is the norm of the acceleration $\gamma''(t)$:

$$\boxed{\kappa(t) = \|\gamma''(t)\|}$$

Remark. As $\|\gamma'\| = 1$, the velocity γ' is perpendicular to the acceleration γ'' by the Trick:
Trick. $\|\gamma'\| = 1 \Rightarrow \gamma' \cdot \gamma' = 1 \Rightarrow \frac{\partial}{\partial t}(\gamma' \cdot \gamma') = 0 \Rightarrow 2\gamma'' \cdot \gamma' = 0 \Rightarrow \gamma'' \perp \gamma'$.

Example. Consider the circle $X^2 + (Y - r)^2 = r^2$ inside the plane $Z = 0$ of $\mathbb{R}^3$, with centre $(0, r, 0)$ and radius r. Near $0 \in \mathbb{R}^3$, it equals the curve

$$\gamma_r(t) = (r \sin \tfrac{t}{r}, r - r \cos \tfrac{t}{r}, 0)$$

for t close to 0. We use $\frac{t}{r}$ instead of t so that γ_r has unit speed:

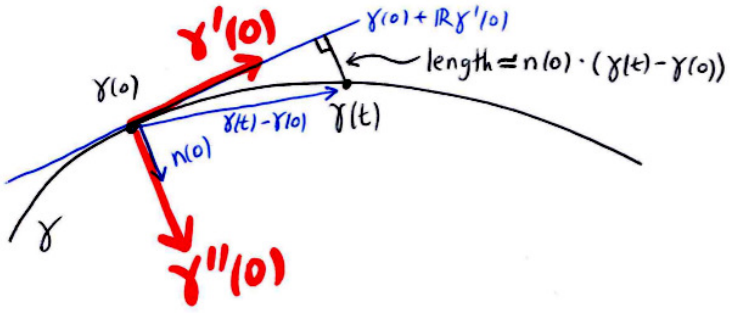
$$\|\gamma_r'(t)\| = \|(\cos \tfrac{t}{r}, \sin \tfrac{t}{r}, 0)\| = 1.$$

As $\gamma_r''(t) = (-\frac{1}{r} \sin \frac{t}{r}, \frac{1}{r} \cos \frac{t}{r}, 0)$, the curvature is

$$\kappa_r(t) = \sqrt{\frac{1}{r^2}} = \frac{1}{r}.$$

In general, a rotation and translation in $\mathbb{R}^3$ will not change the curvature of a curve γ nor the property that γ is parametrized by arc-length (rotations and translations preserve lengths). By rotating and translating, we may assume $\gamma(0) = 0 = \gamma_r(0)$, $\gamma'(0) = (1, 0, 0) = \gamma_r'(0)$, and $\gamma''(0) = (0, \kappa(0), 0)$. Pick $r = 1/\kappa(0)$. Then also $\gamma_r''(0) = \gamma''(0)$, so the Taylor series for γ, γ_r agree up to the second order. So that circle of radius $r = 1/\kappa(0)$ is the best quadratic curve which approximates γ at 0 (when $\kappa(0) = 0$, we can think of the circle γ_∞ of infinite radius as the straight line equal to the x-axis, indeed γ is "flat" up to second-order). The circle γ_r is called the **osculating circle** for γ at $\gamma(0) = 0$, and r is called the **radius of curvature**.

A more geometric way of interpreting the curvature is as follows.



Recall γ'' is orthogonal to γ'. For this reason,

$$n = \frac{\gamma''}{\|\gamma''\|}$$

is called the normal vector to γ (equivalently: $\gamma''(t) = \kappa(t)n(t)$). We now ask: by how much does $\gamma(t)$ swerve away from the line $\gamma(0) + \mathbb{R}\gamma'(0)$ tangent to γ at $t = 0$?

The distance of $\gamma(t)$ from the straight line $\gamma(0)+\mathbb{R}\gamma'(t)$ is, up to order t^3 errors,[1]

$$\begin{aligned} n(0)\cdot(\gamma(t)-\gamma(0)) &= \tfrac{\gamma''(0)}{\|\gamma''(0)\|}\cdot(\gamma'(0)t+\tfrac{1}{2}\gamma''(0)t^2+\cdots) \\ &= \tfrac{1}{2}\|\gamma''(0)\|t^2+\cdots \\ &= \tfrac{1}{2}\kappa(0)t^2+\cdots, \end{aligned}$$

where we used that γ'', γ' are orthogonal. So the curvature $\kappa(0)$ measures how much the curve $\gamma(t)$ deviates from the tangent line.

11.2 The local second fundamental form

Let $S\subset\mathbb{R}^3$ be a smooth surface with a Gauss map $n: S\to\mathbb{R}^3$, so $n(p)\cdot T_pS=0$ and $n(p)\cdot n(p)=1$. How much does S swerve away from the plane $p+T_pS$ tangent to S at p?

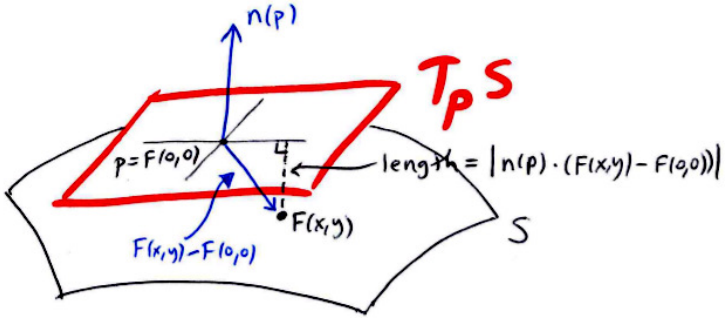
Let F be a parametrization near p. By translating $V\subset\mathbb{R}^2$, we can assume for simplicity that $F(0,0)=p$. Abbreviate $0=(0,0)$. Recall the Taylor series of a smooth function $G(x,y)\in\mathbb{R}$ in two variables is

$$\begin{array}{rcccccc} G(x,y) &=& G(0) &+& D_0G\cdot\left(\begin{smallmatrix}x\\y\end{smallmatrix}\right) &+& \frac{1}{2}\left(\begin{smallmatrix}x\\y\end{smallmatrix}\right)^T \mathrm{Hess}_0\, G\left(\begin{smallmatrix}x\\y\end{smallmatrix}\right) + \cdots \\ &=& G(0) &+& x\partial_xG+y\partial_yG &+& \frac{1}{2}\left(x^2\,\partial_{xx}G+2xy\,\partial_{xy}G+y^2\,\partial_{yy}G\right)+\cdots \end{array}$$

where the partial derivatives of G are all evaluated at $(0,0)$, the **Hessian** $\mathrm{Hess}_0\,G$ is just the matrix of second-order partial derivatives, and we abbreviate $\partial_{xy}G=\partial_x(\partial_yG)$, etc. We can compute this Taylor series taking $G=$ one of the three components of $F=(F_1,F_2,F_3)\in\mathbb{R}^3$.

Example. Consider $F(x,y)=(x,y,ax^2+by^2)$ near $(x,y)=0$. The above becomes:

$$F(x,y)=\left(\begin{smallmatrix}0\\0\\0\end{smallmatrix}\right)+x\left(\begin{smallmatrix}1\\0\\0\end{smallmatrix}\right)+y\left(\begin{smallmatrix}0\\1\\0\end{smallmatrix}\right)+\frac{1}{2}\left(x^2\left(\begin{smallmatrix}0\\0\\2a\end{smallmatrix}\right)+2xy\left(\begin{smallmatrix}0\\0\\0\end{smallmatrix}\right)+y^2\left(\begin{smallmatrix}0\\0\\2b\end{smallmatrix}\right)\right)$$



Therefore, using that $\partial_xF,\partial_yF\in T_pS$ are orthogonal to $n(p)$, the signed distance of a nearby point $F(x,y)$ from the plane $p+T_pS$ is (again evaluating partial derivatives at p):

$$n(p)\cdot(F(x,y)-F(0,0))=\frac{1}{2}\left(x^2\,\mathbf{n}\cdot\partial_{\mathbf{xx}}\mathbf{F}+2xy\,\mathbf{n}\cdot\partial_{\mathbf{xy}}\mathbf{F}+y^2\,\mathbf{n}\cdot\partial_{\mathbf{yy}}\mathbf{F}\right)+\cdots$$

Thus, the analogue for surfaces of the curvature of a curve is the form

$$\boxed{I\!I_F:\mathbb{R}^2\times\mathbb{R}^2\to\mathbb{R},\ (v,w)\mapsto v^T\begin{pmatrix} n\cdot\partial_{xx}F & n\cdot\partial_{xy}F\\ n\cdot\partial_{yx}F & n\cdot\partial_{yy}F\end{pmatrix}w}$$

which is clearly bilinear and symmetric ($\partial_{xy}F=\partial_{yx}F$ by smoothness of F).

Example. $I\!I_F$ need not be positive definite, indeed it can vanish: for the plane $\mathbb{R}^2\subset\mathbb{R}^3$ taking

[1] Here $n(t)\cdot(\gamma(t)-\gamma(0))$ would be the correct distance if the curve lies inside the plane $\gamma(0)+\mathrm{span}(\gamma'(0),\gamma''(0))$. If we truncate the Taylor series of γ after t^3 terms, then the curve does lie in this plane.

$F(x,y) = (x,y,0)$, the second partial derivatives of F are zero.

Example. Continuing with the example $F(x,y) = (x,y,ax^2+by^2)$, we pick:

$$n(0) = \left.\frac{\partial_x F \times \partial_y F}{\|\partial_x F \times \partial_y F\|}\right|_{(x,y)=(0,0)} = \begin{pmatrix}1\\0\\0\end{pmatrix} \times \begin{pmatrix}0\\1\\0\end{pmatrix} / \text{norm} = \begin{pmatrix}0\\0\\1\end{pmatrix}$$

So $2n(0)\cdot(F(x,y)-F(0,0)) = 2ax^2+2by^2$, therefore at $p = F(0,0) = (0,0,0)$:

$$II_F(v,w) = v^T \begin{pmatrix} 2a & 0 \\ 0 & 2b \end{pmatrix} w = 2av_1w_1 + 2bv_2w_2.$$

Example. Let S be the sphere of radius r. The normal is $n(p) = p$. At the North pole $p = (0,0,r)$, S is locally the graph $F(X,Y) = (X,Y,\sqrt{r^2-X^2-Y^2})$, and $n(p) = (0,0,1)$. Dotting with $n(p)$ means we take the third entry, so we need the Hessian of $h = \sqrt{r^2-X^2-Y^2}$ at $(X,Y) = (0,0)$. Compute: $\partial_X h = \frac{-X}{\sqrt{r^2-X^2-Y^2}}$, so $\partial_{XX}h|_{(0,0)} = -\frac{1}{r}$ and $\partial_{YX}h|_{(0,0)} = 0$, and by symmetry also $\partial_{YY}h|_{(0,0)} = -\frac{1}{r}$. Thus

$$II_F = v^T \begin{pmatrix} -\frac{1}{r} & 0 \\ 0 & -\frac{1}{r} \end{pmatrix} w = -\frac{1}{r}v_1w_1 - \frac{1}{r}v_2w_2.$$

We get this same form at each point of the sphere by rotational symmetry.[1]

To avoid confusion later, we make a clear distinction between $n : S \to \mathbb{R}^3$ and the Gauss map in local coordinates $(nF) : \mathbb{R}^2 \supset V \to \mathbb{R}^3$:

$$nF(x,y) = n(F(x,y)).$$

Lemma 11.1. *Locally the matrix for the second fundamental form is, evaluating at p,*

$$\boxed{B = \begin{pmatrix} L & M \\ M & N \end{pmatrix} = \begin{pmatrix} (nF)\cdot\partial_{xx}F & (nF)\cdot\partial_{xy}F \\ (nF)\cdot\partial_{yx}F & (nF)\cdot\partial_{yy}F \end{pmatrix} = -\begin{pmatrix} \partial_x(nF)\cdot\partial_x F & \partial_x(nF)\cdot\partial_y F \\ \partial_y(nF)\cdot\partial_x F & \partial_y(nF)\cdot\partial_y F \end{pmatrix}}$$

Note: $\partial_x(nF)\cdot\partial_y F = \partial_y(nF)\cdot\partial_x F$ as the right-hand side is symmetric using $\partial_{xy}F = \partial_{yx}F$.

Proof. Since n is orthogonal to $TS = \text{span}(\partial_x F, \partial_y F)$,

$$(nF)\cdot\partial_x F = 0 \qquad (nF)\cdot\partial_y F = 0.$$

Differentiating in x or in y gives the equality between the matrices in the claim. □

Using matrix notation, with columns the first partial derivatives:

$$DF = \begin{pmatrix} \partial_x F \mid \partial_y F \end{pmatrix} \qquad \text{and} \qquad D(nF) = \begin{pmatrix} \partial_x(nF) \mid \partial_y(nF) \end{pmatrix},$$

then by the Lemma, $II_F(v,w) = v^T B w$ where:

$$\boxed{B = -D(nF)^T DF}$$

[1] Apply a rotation R about the origin. Then for the parametrization $R\circ F$ near $R(p)$, using $n(R(p)) = R(p)$, we get the same II_F: $n(R(p))^T\partial_{ij}(R\circ F) = p^T R^T R\partial_{ij}F = p^T\partial_{ij}F = n(p)^T\partial_{ij}F$, using that rotations are orthogonal maps ($R^TR = \text{id}$).

11.3 The local second fundamental form under a change of coordinates

Suppose we change coordinates using a transition $\tau = \widetilde{F}^{-1} \circ F$. Since $F(x,y) = \widetilde{F}(\tau(x,y))$, we have $nF(x,y) = n\widetilde{F}(\tau(x,y))$. Differentiating using the chain rule:

$$DF = D\widetilde{F}\, D\tau \qquad \text{and} \qquad D(nF) = D(n\widetilde{F})\, D\tau.$$

Thus, the change in the second fundamental form is:

$$B = -D(nF)^T DF = -(D(n\widetilde{F})\, D\tau)^T (D\widetilde{F}\, D\tau) = -D\tau^T\, D(n\widetilde{F})^T\, D\widetilde{F}\, D\tau = D\tau^T\, \widetilde{B}\, D\tau,$$

thus as expected: $\boxed{B = D\tau^T\, \widetilde{B}\, D\tau}$

11.4 The second fundamental form

Since the local forms change correctly under the transition ($D\tau : TV \to T\widetilde{V}$ is the correct identification between the local tangent spaces, so $B = D\tau^T\, \widetilde{B}\, D\tau$ is the correct change of coordinates for bilinear forms), the local forms

$$I\!I_F : \mathbb{R}^2 \times \mathbb{R}^2 = T_pV \times T_pV \to \mathbb{R}$$

determine a well-defined global bilinear form

$$I\!I : T_pS \times T_pS \to \mathbb{R}$$

independent of the observer. More explicitly, if $v = DF(v_F), w = DF(w_F)$ are the actual vectors in $TS \subset \mathbb{R}^3$, rather than the local versions $v_F, w_F \in TV = \mathbb{R}^2$, then by the chain rule:

$$I\!I_F(v_F, w_F) = -v_F^T D(nF)^T\, DF w_F = -(Dn\, DF\, v_F) \cdot (DF\, w_F) = -Dn(v) \cdot w = I\!I(v,w)$$

is independent of the choice of F. Therefore:

Theorem 11.2. *The local forms $I\!I_F$ determine a well-defined symmetric bilinear form, called the* ***second fundamental form****,*

$$\boxed{I\!I : T_pS \times T_pS \to \mathbb{R},\ I\!I(v,w) = -(D_pn)(v) \cdot\, w}$$

We should now clarify better what exactly $-Dn$ means. We want to avoid using extensions[1] of n as in definition 2.6, which is messy and requires us to make choices (although the choice of extension does not matter in the end). The derivative map

$$Dn : T_pS \to T_{n(p)}\mathbb{R}^3 = \mathbb{R}^3$$

of $n : S \to \mathbb{R}^3$ is a well-defined map independent of parametrizations: in Section 9.2 this map was defined in terms of curves in S without using parametrizations. A vector $v \in TS$ is an equivalence class of curves $[c_v]$ in S, satisfying $p = c_v(0)$, $v = c_v'(0)$, and

$$Dn[c_v] = [n \circ c_v].$$

Notice that this makes sense by the chain rule: we identify $[c_v] \equiv \partial_t|_{t=0} c_v = c_v'(0) = v$, so

$$[n \circ c_v] \equiv \partial_t|_{t=0}(n \circ c_v) = (D_{c_v(0)}n)\,(c_v'(0)) = D_pn(v).$$

Notice the above description gives a very useful formula, where $\gamma'(t) = (X'(t), Y'(t), Z'(t))$ is a general vector in TS in terms of a curve $\gamma(t) = (X(t), Y(t), Z(t)) \in S$:

$$\boxed{Dn(\gamma'(t)) = Dn \begin{pmatrix} X'(t) \\ Y'(t) \\ Z'(t) \end{pmatrix} = \left.\frac{\partial}{\partial t}\right|_{t=0} n(\gamma(t))}$$

[1] n is a map $S \to \mathbb{R}^3$, so first choose an extension of n (at least locally) to a neighbourhood of S, then n becomes a map $n : \mathbb{R}^3 \to \mathbb{R}^3$ defined near S, hence we know what Dn means (matrix of partial derivatives).

Example. For the plane $S = \mathbb{R}^2 \subset \mathbb{R}^3$, $n = (0,0,1)$ is constant, so $Dn = 0$, so

$$II = 0.$$

Example. The cylinder S of radius r, $X^2 + Y^2 = r^2$, has Gauss map $n(X,Y,Z) = (X,Y,0)/r$ (this is the outward normal to the cylinder). For $\gamma(t) = (X(t), Y(t), Z(t))$ in S, we have $n(\gamma(t)) = (X(t), Y(t), 0)/r$ so

$$Dn \begin{pmatrix} X'(t) \\ Y'(t) \\ Z'(t) \end{pmatrix} = \partial_t|_{t=0} n(\gamma(t)) = \begin{pmatrix} X'(t)/r \\ Y'(t)/r \\ 0 \end{pmatrix}.$$

A basis for T_pS consists of a vector E_1 tangent to the equatorial circle of the cylinder (so taking $Z(t) =$ constant), and a vector E_2 parallel to the axis of the cylinder (so taking $X(t), Y(t)$ constant). It follows by the above calculation that $Dn(E_1) = E_1/r$, $Dn(E_2) = 0$. Thus $Dn : TS \times TS \to TS$, $Dn = \left(\begin{smallmatrix} 1/r & 0 \\ 0 & 0 \end{smallmatrix}\right)$ in the basis E_1, E_2. Therefore

$$II(v,w) = -v^T \begin{pmatrix} \frac{1}{r} & 0 \\ 0 & 0 \end{pmatrix} w = -\frac{1}{r} v_1 w_1.$$

For example, an orthonormal choice of E_1, E_2 at $p = (r\cos\theta, r\sin\theta, Z)$ is

$$E_1 = (-\sin\theta, \cos\theta, 0) \quad \text{and} \quad E_2 = (0,0,1).$$

Remark. Even if two surfaces are isometric, the second fundamental form can be substantially different. E.g. the cylinder is locally isometric to the flat plane, but it has a non-trivial II. So the second fundamental form depends on **extrinsic** *information: the choice of embedding into $\mathbb{R}^3$. Whereas we think of the first fundamental form as* **intrinsic** *(due to Theorem 10.6).*

11.5 Summary of first and second fundamental forms

Fundamental form $T_pS \times T_pS \to \mathbb{R}$ $v, w \in T_pS \subset \mathbb{R}^3$	Local fundamental form $T_pV \times T_pV \to \mathbb{R}$ $v_F, w_F \in T_pV = \mathbb{R}^2$ (So $DF(v_F) = v$, etc.)	Local matrix	Coordinate change $\tau = \widetilde{F}^{-1} \circ F$
$I(v,w) = v \cdot w$	$I_F(v_F, w_F) = v_F^T A w_F$	$A = DF^T DF$	$A = D\tau^T \, \widetilde{A} \, D\tau$
$II(v,w) = -Dn(v) \cdot w$	$II_F(v_F, w_F) = v_F^T B w_F$	$B = -D(nF)^T DF$	$B = D\tau^T \, \widetilde{B} \, D\tau$

11.6 The second fundamental form is the variation of the first

For the purposes of this course, the following is not a central result, it is just a curiosity.

Theorem 11.3. *The second fundamental form is the variation of the first fundamental form,*

$$\tfrac{\partial}{\partial t}\big|_{t=0} I_t = -2\, II,$$

when we deform the surface in the normal direction, meaning we vary the local parametrization by $(x,y) \mapsto F(x,y) + t\, n(F(x,y))$ in terms of time $t \in [0, \text{small}]$.

Proof. Writing $F_t = F + tnF$, we get $\partial_x F_t = \partial_x F + t\partial_x(nF)$, $\partial_y F_t = \partial_y F + t\partial_y(nF)$, so

$$I_t = DF_t^T DF_t = I_0 + t \begin{pmatrix} 2\partial_x(nF)\cdot\partial_x F & \partial_x(nF)\cdot\partial_y F + \partial_y(nF)\cdot\partial_x F \\ \partial_x(nF)\cdot\partial_y F + \partial_y(nF)\cdot\partial_x F & 2\partial_y(nF)\cdot\partial_y F \end{pmatrix} + \text{order } t^2.$$

Thus the claim follows by using the relations from the proof of Lemma 11.1 (in particular the symmetry $\partial_x(nF)\cdot\partial_y F = \partial_y(nF)\cdot\partial_x F$). □

12. Curvature

12.1 The shape operator $\mathbb{S} = -Dn : T_pS \to T_pS$

Lemma 12.1. *$T_pS = T_{n(p)}S^2 \subset \mathbb{R}^3$ where S^2 is the unit sphere in $\mathbb{R}^3$.*

Proof. T_pS is the vector subspace of $\mathbb{R}^3$ orthogonal to the normal $n(p)$ at $p \in S$. Similarly, $T_{n(p)}S^2$ is the vector subspace of $\mathbb{R}^3$ orthogonal to the normal to the sphere S^2 at $n(p)$. But recall that the normal to S^2 at $n(p)$ is just $n(p)$. So those vector subspaces equal. □

Corollary 12.2. *$Dn : T_pS \to T_{n(p)}S^2$ can be viewed as a linear endomorphism of the 2-dimensional vector space T_pS, so $Dn : T_pS \to T_pS$.*

Proof. We need to show D_pn lands in $T_{n(p)}S^2$. Let $\gamma \subset S$ be any curve through $\gamma(0) = p$. Then $Dn(\gamma') = \partial_t(n \circ \gamma)$ is the velocity of a curve $n \circ \gamma \subset S^2$ so it is tangent to S^2 (at $n(\gamma(0)) = n(p)$). Another proof: differentiate $n(\gamma) \cdot n(\gamma) = 1$ (since n is unit length) in time: $2n(\gamma) \cdot Dn(\gamma') = 0$ so $Dn(\gamma')$ is perpendicular to n, so it lies in $T_{n(p)}S^2$. □

Definition 12.3 (Shape operator)**.** $\boxed{\mathbb{S} = -Dn : T_pS \to T_pS}$ *is the **shape operator**.*

Theorem 12.4. *The fundamental forms are related by:*

$$\boxed{I\!I(v, w) = \mathbb{S}(v) \cdot w = I(\mathbb{S}(v), w)}$$

Proof. $I\!I(v, w) = -Dn(v) \cdot w$ and $Dn(v) \in T_{n(p)}S^2 = T_pS$, so that dot product can be computed using $I : T_pS \times T_pS \to \mathbb{R}$ as both $-Dn(v), w$ lie in T_pS. □

Corollary 12.5. *$\mathbb{S}$ is self-adjoint[1] with respect to the inner product I*

$$I(\mathbb{S}v, w) = I(v, \mathbb{S}w).$$

Proof. I and $I\!I$ are symmetric, so $I(\mathbb{S}v, w) = I\!I(v, w) = I\!I(w, v) = I(\mathbb{S}w, v) = I(v, \mathbb{S}w)$. □

Lemma 12.6. *Let $nF = n \circ F$ be the local expression for n in the parametrization F, then*

$$\partial_x(nF), \partial_y(nF) \in TS. \tag{12.1}$$

Let $[\mathbb{S}_{ij}]$ be the matrix[2] for $\mathbb{S}$ in the basis $\partial_x F$, $\partial_y F$ of TS, then

$$\begin{aligned} &-\partial_x(nF) = \mathbb{S}_{11}\,\partial_x F + \mathbb{S}_{21}\,\partial_y F \quad \text{and} \quad -\partial_y(nF) = \mathbb{S}_{12}\,\partial_x F + \mathbb{S}_{22}\,\partial_y F \\ &-D(nF) = -\begin{pmatrix} \partial_x(nF) \mid \partial_y(nF) \end{pmatrix} = \begin{pmatrix} \partial_x F \mid \partial_y F \end{pmatrix} \begin{pmatrix} \mathbb{S}_{11} & \mathbb{S}_{12} \\ \mathbb{S}_{21} & \mathbb{S}_{22} \end{pmatrix} = DF\ \mathbb{S}. \end{aligned} \tag{12.2}$$

Proof. By differentiating $n \cdot n = 1$ we deduce that $\partial_x(nF) \cdot nF = 0$, $\partial_y(nF) \cdot nF = 0$. Hence (12.1) follows, since TS is the plane orthogonal to n.

By the chain rule, $-Dn(\partial_x F) = -\partial_x(nF)$ and $-Dn(\partial_y F) = -\partial_y(nF)$.

Thus, abbreviating x, y by indices $1, 2$, using the basis $X_i = \partial_i F$ we have:[3] $-Dn(X_i) = \sum_{j=1}^{2} \mathbb{S}_{ji} X_j$. So the matrix $\mathbb{S}_{ij}$ represents $-Dn$ in the basis $X_i = \partial_i F$. □

---

[1] Self-adjointness means that the matrix for $\mathbb{S}$ in a basis which is *orthonormal* with respect to I will be symmetric. However, the matrix $\mathbb{S}_{ij}$ discussed below is not symmetric in general. The symmetry $\partial_i X_k = \partial_i \partial_k F = \partial_k \partial_i F = \partial_k X_i$ implies symmetry in i, k in $-Dn(X_i) \cdot X_k = -\partial_i(nF) \cdot X_k = (nF) \cdot \partial_i X_k$ (differentiating the orthogonality relation $(nF) \cdot X_k = 0$). This however does not imply the symmetry $\mathbb{S}_{ij} = \mathbb{S}_{ji}$, it only implies that $I\!I(X_i, X_k) = -Dn(X_i) \cdot X_k$ is symmetric in i, k.

[2] So the $\mathbb{S}_{ij}$ are smooth functions in the local variables x, y.

[3] The i-th column of the matrix $\mathbb{S}$ is the image of the i-th basis vector X_i, written in the basis X_j.

Using the notation of the previous proof, Theorem 12.4 can be rewritten locally as

$$II(X_i, X_k) = -Dn(X_i) \cdot X_k = \sum_{j=1}^{2} \mathbb{S}_{ji} X_j \cdot X_k = \sum_{j=1}^{2} \mathbb{S}_{ji} I(X_j, X_k).$$

Notice this implies $B = \mathbb{S}^T A$ (we come back to this proof in Theorem 12.8). Intuitively $\mathbb{S}$ is "just the same as" II: we turned the 2-form II into a linear map using the inner product I.[1]

12.2 Principal curvatures, mean curvature, Gaussian curvature

Since the shape operator $\mathbb{S}$ is self-adjoint, it can be diagonalized using an orthonormal[2] basis E_1, E_2 of eigenvectors for TS (where orthonormality is computed using I).

$$\mathbb{S} = \begin{pmatrix} \kappa_1 & 0 \\ 0 & \kappa_2 \end{pmatrix}$$

Principal directions	E_1, E_2	o.n. eigenvectors of $\mathbb{S}$
Principal curvatures	κ_1, κ_2	eigenvalues of $\mathbb{S}$
Mean curvature	$H = \frac{1}{2}(\kappa_1 + \kappa_2)$	$H = \frac{1}{2}\text{trace}(\mathbb{S})$
Gaussian curvature	$K = \kappa_1 \kappa_2$	$K = \det(\mathbb{S})$

We will see later, that the principal directions are the directions that a curve in the surface must travel along to have maximum and minimum curvature (= the two principal curvatures).

Remark. Explicitly, we are diagonalizing the symmetric bilinear form $II : TS \times TS \to \mathbb{R}$, so

$$II(E_i, E_j) = \kappa_i \delta_{ij} \qquad \text{and} \qquad I(E_i, E_j) = \delta_{ij}$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ for $i \neq j$. If you think of E_i, E_j as vectors in $\mathbb{R}^3$, then of course orthonormality just means $E_i \cdot E_j = \delta_{ij}$ for the usual dot product.

Example. For the cylinder $X^2 + Y^2 = r^2$ at the end of Sec.11.4, at $p = (r\cos\theta, r\sin\theta, Z)$ the eigenvectors $E_1 = (-\sin\theta, \cos\theta, 0)$, $E_2 = (0, 0, 1)$ are orthonormal w.r.t. dot product in $\mathbb{R}^3$, and we found $II = -\frac{1}{r} dx^2$, so $\kappa_1 = -\frac{1}{r}$ and $\kappa_2 = 0$ (as expected since in the E_1 direction the surface looks like a circle of radius r, and in the E_2 direction the surface looks like a straight line).

Example. In the example $F(x, y) = (x, y, ax^2 + by^2)$, we computed at $(x, y) = 0$:

$$II_F = \left(\begin{smallmatrix} 2a & 0 \\ 0 & 2b \end{smallmatrix}\right)$$

in the basis $v_F = (1, 0)$, $w_F = (0, 1)$. The first fundamental form is

$$I_F = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right),$$

so the basis v_F, w_F is orthonormal. So $\kappa_1 = 2a$, $\kappa_2 = 2b$.

---

[1] Compare the raising/lowering of indices in Sect.13.3. We are doing $II_{ik} = \mathbb{S}_i^j g_{jk}$, where $g_{jk} = I(X_j, X_k)$.

[2] **Warning.** If you compute locally, you must ensure orthonormality $I(E_i, E_j) = \delta_{ij}$. For example, if you use $F(\lambda x, \lambda y)$ instead of $F(x, y)$, then the local matrix B would become $\lambda^2 B$, so curvatures computed for these matrices would locally change by λ^2 or λ^4: but we want them to be independent of parametrizations! More generally, if we pick a clever non-orthogonal linear transition $\tau = S$, so $D\tau = S$, then we can turn B into a matrix $\widetilde{B} = S^T B S$ of the form $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} \pm 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, or $\left(\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ (a bilinear form is determined up to congruency by its signature). So locally, if you ignore orthonormality, you could arbitrarily rescale by positive numbers the eigenvalues of the local matrix B using changes of coordinates. So only the following signs are preserved:

$$\text{sign}\, K = \text{sign}\, \det B \qquad (\text{signs of } \kappa_1, \kappa_2) = (\text{signs of eigenvalues of } B).$$

For example $\det(\widetilde{B}) = \det(D\tau^T B D\tau) = \det(D\tau)^2 \det(B)$. For the same reasons, you have no control over the sign of the mean curvature.

Lemma 12.7. *If $\partial_x F$, $\partial_y F \in \mathbb{R}^3$ are orthonormal at p, then*

$$\begin{aligned} \kappa_1, \kappa_2 &= \text{(the eigenvalues of the local matrix } B \text{ for } II_F) \\ K &= \det B. \end{aligned}$$

Proof. The change of basis from $\partial_x F, \partial_y F$ to E_1, E_2 will be an orthogonal matrix Q, and so $II_F = Q^T \left(\begin{smallmatrix} \kappa_1 & 0 \\ 0 & \kappa_2 \end{smallmatrix}\right) Q = Q^{-1} \left(\begin{smallmatrix} \kappa_1 & 0 \\ 0 & \kappa_2 \end{smallmatrix}\right) Q$, using orthogonality: $Q^T = Q^{-1}$. But conjugation does not change the characteristic polynomial, so the eigenvalues are still κ_1, κ_2. □

Theorem 12.8. *Writing $\mathbb{S}$ for the matrix of the shape operator in the basis $\partial_x F, \partial_y F$ yields a relation between the local matrices A, B for I_F, II_F:*[1]

Weingarten equations	$\mathbb{S} = A^{-1}B$
Gaussian curvature	$K = \det \mathbb{S} = \frac{\det B}{\det A}$

These can be written out explicitly in terms of the coefficient functions of I_F, II_F:

$$II = \mathbb{S} = I_F^{-1}\, II_F = \begin{pmatrix} e & f \\ f & g \end{pmatrix}^{-1} \begin{pmatrix} L & M \\ M & N \end{pmatrix} = \frac{1}{eg - f^2} \begin{pmatrix} g & -f \\ -f & e \end{pmatrix} \begin{pmatrix} L & M \\ M & N \end{pmatrix}$$

$$K = \det(\mathbb{S}) = \frac{\det(II_F)}{\det(I_F)} = \frac{LN - M^2}{eg - f^2}.$$

Proof. It is enough to show that $\mathbb{S}^T A = B$, since then $\mathbb{S} = (BA^{-1})^T = (A^T)^{-1}B^T = A^{-1}B$, using that A, B are symmetric. Below are two ways to check that $\mathbb{S}^T A = B$.

In local coordinates, for $v, w \in \mathbb{R}^2$, we have $I(v, w) = v^T A w$, $II(v, w) = v^T B w$. So $I(\mathbb{S}v, w) = II(v, w)$ becomes $v^T \mathbb{S}^T A w = v^T B w$. As this holds for all v, w, we deduce $\mathbb{S}^T A = B$.

Alternatively, combine the matrix equation in Lemma 12.6 with Section 11.5,

$$B = -D(nF)^T\, DF = (DF\, \mathbb{S})^T\, DF = \mathbb{S}^T DF^T DF = \mathbb{S}^T A. \quad \square$$

12.3 Normal curvature

The **normal curvature** of a curve γ in S through $p = \gamma(0)$, with γ parametrized by arc-length so speed $\|\gamma'\| = 1$, is the component of the acceleration γ'' in the normal direction[2]

$$\boxed{\gamma''(0) \cdot n(p) = -Dn(\gamma'(0)) \cdot \gamma'(0) = II(\gamma'(0), \gamma'(0))}$$

Intuitively, if you are racing with a car on a surface, the normal curvature tells you how much pull away from the surface you feel when you accelerate the car.

Since $\|\gamma'(0)\| = 1$, the normal curvatures are measured by the quadratic form $II(v, v)$ on unit vectors $v = \cos\theta\, E_1 + \sin\theta\, E_2 \in T_p S$. Explicitly, we get the **Euler formula**:

$$II(v, v) = \kappa_1 \cos^2\theta + \kappa_2 \sin^2\theta$$

so κ_1, κ_2 are the extreme values (min and max)[3] of the possible normal curvatures at p.

---

[1] It is easy to check that K does not change if we change local parametrization. More strikingly, Gauss' Theorem Egregium (Section 13.2) says that K does not change under isometries, so it is **intrinsic** to the surface with its Riemannian metric, it is not extrinsic (= dependent on the choice of embedding).

[2] where we used the usual tricks: $\nu(t) = n(\gamma(t))$ is orthogonal to TS, so $\nu(t) \cdot \gamma'(t) = 0$, differentiating: $\nu' \cdot \gamma' + \nu \cdot \gamma'' = 0$, and finally $\nu'(0) = \partial_t|_{t=0} n(\gamma(t)) = Dn(\gamma'(0))$. The formula then follows.

[3] Proof: if $\kappa_1 \le \kappa_2$, then $\kappa_1 = \kappa_1(\cos^2\theta + \sin^2\theta) \le \kappa_1 \cos^2\theta + \kappa_2 \sin^2\theta \le \kappa_2(\cos^2\theta + \sin^2\theta) = \kappa_2$.

Corollary 12.9. *The principal curvatures κ_1, κ_2 are the min and max of the ratio of the two quadratic forms:*

$$\frac{Lx^2 + 2Mxy + Ny^2}{ex^2 + 2fxy + gy^2}$$

Proof. We can rescale (x, y) without affecting the above ratio so that $ex^2+2fxy+gy^2 = 1$. But this is precisely the condition that $I_F(v_F, v_F) = 1$ for the local vector $v_F = (x, y) \in TV = \mathbb{R}^2$. Then $II_F(v_F, v_F)$ is equal to the numerator. We proved above that for unit vectors $v \in \mathbb{R}^3$ the min and max of the quadratic form $II(v, v)$ are the principal curvatures κ_1, κ_2. □

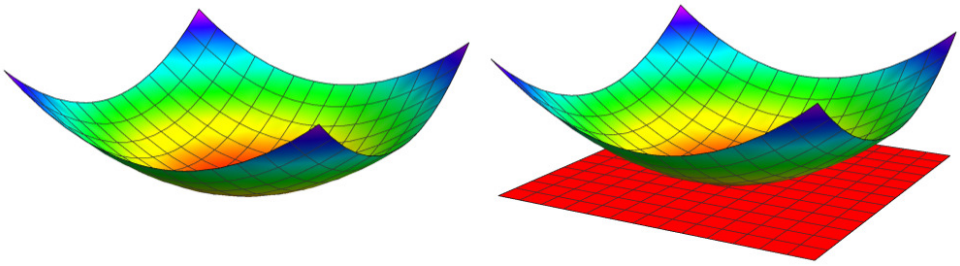
12.4 Qualitative interpretation of the curvatures

Loosely speaking:

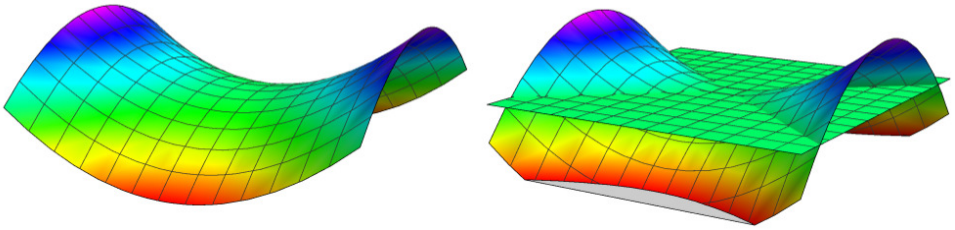
(1) The principal directions E_1, E_2 tell you the directions of steepest increase/decrease,
(2) If $\kappa_1 > 0$, a curve is S with tangent E_1 is accelerating in the normal direction, since $\gamma'' \cdot n = II(\gamma', \gamma') = II(E_1, E_1) = \kappa_1 > 0$ (using Sec.12.3). So near p, the curve lies on the same side of TS as $n(p)$ does.
(3) If $\kappa_1 < 0$, a curve in direction E_1 accelerates away from $n(p)$ so it lies on the other side of TS.
(4) $K > 0$ means κ_1, κ_2 have the same sign, so either all curves accelerate towards $n(p)$ or away from $n(p)$, so S is locally all on the same side of the tangent plane and S can be locally approximated by an ellipsoid.
We call a point $p \in S$ **elliptic** if κ_1, κ_2 have the same sign ($\Leftrightarrow K > 0$)
(5) We call a point $p \in S$ **hyperbolic** if κ_1, κ_2 have opposite sign ($\Leftrightarrow K < 0$). This means S locally looks like a saddle and S lies on both sides of the tangent plane.

Example. In the above example $F(x, y) = (x, y, ax^2 + by^2)$ near $(x, y) = 0$:

For $a, b > 0$, the origin is an elliptic point: the surface is $Z = aX^2 + bY^2$, it lies all above the tangent plane $Z = 0$, and we get an ellipse when we slice S with a plane $Z =$(positive constant) which is parallel to the tangent plane:



For $a < 0$, $b > 0$, the surface lies on both sides of the tangent plane $Z = 0$, and we get a hyperbola when we slice S:



12.5 Curvatures in terms of the Hessian

Theorem 12.10. *Parametrizing $S \subset \mathbb{R}^3$ near p as the graph of a function $h : T_pS \to \mathbb{R}$ over the tangent plane (Theorem 9.1), II_F at p becomes the Hessian of h at p:*

$$II_F = \begin{pmatrix} L & M \\ M & N \end{pmatrix} = \pm \begin{pmatrix} h_{xx} & h_{xy} \\ h_{yx} & h_{yy} \end{pmatrix} = \pm \operatorname{Hess}_p h,$$

where the sign depends on the choice of normal. In particular the Gaussian curvature at p is

$$K(p) = h_{xx}h_{yy} - h_{xy}^2 = \det \operatorname{Hess}_p h.$$

Proof. In the proof of Theorem 9.1, $F(x, y) = (x, y, h(x, y))$, so $\partial_x F = (1, 0, \partial_x h)$ and $\partial_y F = (0, 1, \partial_y h)$ is a basis for $T_pS = (xy\text{-plane})$. This forces $\partial_x h = \partial_y h = 0$ at $(0, 0)$, so $(0, 0)$ is a critical point of h. Thus the normal is $\pm n(p) = (1, 0, 0) \times (0, 1, 0) = (0, 0, 1)$, and dot product with n just means taking $\pm$ the third component of a vector (the sign $\pm$ depends on the choice of Gauss map). The basis $\partial_x F, \partial_y F$ is orthonormal, and in this basis II is the matrix in the claim. Now use Lemma 12.7. □

Notice that, rotating as in the proof with S locally $(X, Y, h(X, Y))$, the signs of the Hessian of h are precisely what you studied in applied courses to discuss minima, maxima and saddles of a function $h(X, Y)$ of two variables. So

$$\begin{array}{rcll} \kappa_1, \kappa_2 > 0 & \Rightarrow & Z = h(X, Y) \text{ has a minimum at } 0 & \Rightarrow \quad S \text{ lies above } Z = 0 \\ \kappa_1, \kappa_2 < 0 & \Rightarrow & Z = h(X, Y) \text{ has a maximum at } 0 & \Rightarrow \quad S \text{ lies below } Z = 0 \\ K = \kappa_1\kappa_2 < 0 & \Rightarrow & Z = h(X, Y) \text{ has a saddle at } 0 & \Rightarrow \quad S \text{ lies on both sides of } Z = 0 \end{array}$$

Example. Consider the sphere S of radius r, $X^2+Y^2+Z^2 = r^2$, near the North pole $p = (0, 0, 1)$. Locally S is the graph $F(X, Y) = (X, Y, \sqrt{r^2 - X^2 - Y^2})$ with normal $n(p) = (0, 0, 1)$. The Hessian of $\sqrt{r^2 - X^2 - Y^2}$ at 0 gives

$$II = \begin{pmatrix} -\frac{1}{r} & 0 \\ 0 & -\frac{1}{r} \end{pmatrix} \qquad K = \det II = \frac{1}{r^2}$$

This was expected: the great arcs are osculating circles of radius r which accelerate away from the outward normal direction $(0, 0, 1)$. Notice the curvature becomes small as $r \to \infty$ since the surface looks more and more "flat" for large r.

12.6 Locally flat surfaces

Theorem 12.11. *If $II = 0$ near p, then S lies in a plane near p.*

Proof. Let's use the simpler notation $n = n(x, y)$ instead of $n(F(x, y))$ for the local expression of n in this proof. We need to show that locally the Gauss map n is constant and that S satisfies the equation for a plane orthogonal to n:

$$n \cdot F = \text{constant.} \qquad (\text{equivalently } n \cdot (F(x, y) - F(0, 0)) = 0)$$

From $II = 0$ we obtain $\partial_x n \cdot \partial_x F = 0$, $\partial_x n \cdot \partial_y F = 0$, etc. and hence by linearity $\partial_x n, \partial_y n$ are orthogonal to all of $TS = \operatorname{span}(\partial_x F, \partial_y F)$. Differentiating $n \cdot n = 1$ shows they are also orthogonal to n: $\partial_x n \cdot n = 0$ and $\partial_y n \cdot n = 0$. So they are orthogonal to all of $\mathbb{R}^3$ and thus must vanish. So n is locally constant in x, y. So $n \cdot F$ is also constant: $\partial_x(n \cdot F) = n \cdot \partial_x F = 0$ (since n is orthogonal to $\partial_x F \in TS$), and similarly $\partial_y(n \cdot F) = 0$. □

12.7 The Gaussian curvature as a ratio of areas

Since n is a unit vector, it maps $\boxed{n : S \to S^2 \subset \mathbb{R}^3}$ into the unit sphere S^2 of $\mathbb{R}^3$. So we can compare the areas of two regions $n(U) \subset S^2$ and $U \subset S$. The following says the Gaussian curvature K precisely measures the infinitesimal ratio of those two areas.

Theorem 12.12. *The Gaussian curvature at $p \in S$ equals*

$$K(p) = \lim_{U \to p} \pm \frac{\text{Area}(n(U) \subset S^2)}{\text{Area}(U \subset S)},$$

where the limit is over shrinking neighbourhoods U of p, and where $\pm = sign(K(p))$.

Proof. Work locally with a parametrization $F = F(x, y)$ such that $F(0, 0) = p$. Recall by Section 10.8, the area of $U = F(V) \subset S$ is defined as

$$\text{Area}(U) = \int_V \|\partial_x F \times \partial_y F\| \, dx \, dy.$$

The area of $n(U)$ is analogously:[1]

$$\text{Area}(n(U)) = \int_V \|\partial_x (nF) \times \partial_y (nF)\| \, dx \, dy.$$

We may assume that F is correctly oriented, so that $n = (\partial_x F \times \partial_y F)/\|\partial_x F \times \partial_y F\|$. Using the shape operator $\mathbb{S}$ written in the basis $\partial_x F, \partial_y F$, we compute:

$$\begin{aligned} \partial_x (nF) \times \partial_y (nF) &= (\partial_x F \, \mathbb{S}_{11} + \partial_y F \, \mathbb{S}_{21}) \times (\partial_x F \, \mathbb{S}_{12} + \partial_y F \, \mathbb{S}_{22}) \\ &= (\mathbb{S}_{11}\mathbb{S}_{22} - \mathbb{S}_{21}\mathbb{S}_{12}) \, \partial_x F \times \partial_y F \\ &= \det(\mathbb{S}) \, \partial_x F \times \partial_y F \\ &= \det(\mathbb{S}) \, \|\partial_x F \times \partial_y F\| \, n. \end{aligned}$$

Since $K(x, y) = \det \mathbb{S}$ is the Gaussian curvature at $F(x, y)$, and $\|n\| = 1$,

$$\text{Area}(n(U)) = \int_V |K(x, y)| \, \|\partial_x F \times \partial_y F\| dx \, dy.$$

Writing $|K(x, y)| = |K(p)| + (|K(x, y)| - |K(p)|)$, we obtain

$$\begin{aligned} \text{Area}(n(U)) &= |K(p)| \int_V \|\partial_x F \times \partial_y F\| dx \, dy + \int_V (|K(x, y)| - |K(p)|) \, \|\partial_x F \times \partial_y F\| dx \, dy \\ &= |K(p)| \, \text{Area}(U) + \int_V (|K(x, y)| - |K(p)|) \, \|\partial_x F \times \partial_y F\| dx \, dy. \end{aligned}$$

Divide that by $\text{Area}(U)$, move the first term on the right to the left, and take absolute values:

$$\begin{aligned} \left| \tfrac{\text{Area}(n(U))}{\text{Area}(U)} - |K(p)| \right| &\leq \tfrac{1}{\text{Area}(U)} \cdot \max_{(x,y) \in \overline{V}} ||K(x, y)| - |K(p)|| \cdot \int_V \|\partial_x F \times \partial_y F\| dx \, dy \\ &= \max_{(x,y) \in \overline{V}} ||K(x, y)| - |K(p)|| \end{aligned}$$

and the final expression converges to 0 as we shrink V to $(0, 0)$ since $K(x, y) \to K(0, 0) = K(p)$ by continuity of K (so $x, y \to 0$). □

[1] *Non-examinable technical remark.* If $\partial_x(nF), \partial_y(nF)$ are linearly independent at $(0, 0)$ then near $n(p) \in S^2$ we have a local parametrization $n \circ F : V \to S^2$, and the formula for $\text{Area}(n(U))$ is valid. If $\partial_x(nF), \partial_y(nF)$ are linearly dependent at $(0, 0)$ then we cannot say that. However, one can still argue that the ratio of the areas in the claim converges to zero (notice $K(p) = 0$ here, as the columns of II are linearly dependent). Indeed, one can check that $\text{Area}(n(U)) \leq \varepsilon \, \text{Area}(U)$ for small enough V where $\varepsilon \to 0$ as we shrink V. We will not carry out these details.

Lemma 12.13. *We record two formulae, one from Sec.10.8 and one from the proof above:*

$$\boxed{\partial_x F \times \partial_y F = \pm\sqrt{\det I_F}\, n \qquad \text{and} \qquad \partial_x n \times \partial_y n = \pm K\sqrt{\det I_F}\, n}$$

where the sign is + precisely if $\partial_x F, \partial_y F$ is a right-handed basis ($(\partial_x F \times \partial_y F)\cdot n > 0$).

13. Tangential derivatives and Gauss' Theorema Egregium

13.1 Tangential derivative (Levi-Civita connection)

Let S be a smooth surface in $\mathbb{R}^3$. Recall in Section 9.4 we defined vector fields on S, and we abbreviated $X_1 = \partial_x F$, $X_2 = \partial_y F$ for a local parametrization F near $p \in S$. Thus, at each point $p \in S$, we have a basis of $\mathbb{R}^3$ given by:

$$X_1(p), X_2(p), n(p)$$

since $\mathbb{R}^3 = T_pS \oplus \mathbb{R}n(p)$, as n is orthogonal to T_pS. Therefore, the derivatives of a smooth local tangent vector field $v(x,y) \in T_{F(x,y)}S$ defined near $p \in S$ can be written in terms of this basis. In particular, the orthogonal projection to T_pS of such derivatives is called the **tangential derivative**, which you study in Exercise Sheet 2:

$$\boxed{\begin{array}{rcl} \nabla_x v & = & \text{orthogonal projection of } \partial_x v \text{ onto } TS \\ & = & \partial_x v - (n\cdot\partial_x v)\, n \\ & = & \partial_x v + (\partial_x n \cdot v)\, n \end{array}}$$

where $n = n(x,y)$ is the local expression $n(F(x,y))$ for the Gauss map (also $n\cdot\partial_x v = -\partial_x n \cdot v$ by differentiating the orthogonality relation $n\cdot v = 0$, compare Exercise 4 of Exercise Sheet 2). Abbreviate x, y by indices $1, 2$. Then

$$\boxed{\partial_j v = \nabla_j v + (Bv)_j n}$$

since, writing $v = \sum v^i X_i$ by abbreviating the coefficient functions $v^1 = a(x,y)$, $v^2 = b(x,y)$, and recalling the definition of the second fundamental form $B_{ij} = n\cdot\partial_i X_j = -\partial_i n \cdot X_j$,

$$\partial_j n \cdot v = \partial_j n \cdot \sum v^i X_i = -\sum B_{ji} v^i = -(Bv)_j.$$

The symbol ∇ is called **nabla**, and the operator ∇ is called a **connection** for the surface S. Notice that a connection is a way to differentiate vector fields without ever leaving the tangent space TS (it turns out connections exist also for abstract surfaces and manifolds).

Notice that in fact you can differentiate any vector field by any other vector field. Given a vector field $X = \sum a^j X_j$, we define ∇_X linearly in the differentiating variable X:

$$\nabla_X v = \sum a^j \nabla_j v.$$

Note $\nabla_X v$ is of course not linear (with respect to smooth functions) in the v variable since

$$\boxed{\nabla_j(fv) = (\partial_j f)v + f\nabla_j v}$$

where $f = f(x,y)$ is a function (here we used that $\partial_j(f)v$ is already in TS, so doesn't change under orthogonal projection). That equation is called, of course, **Leibniz rule**.

Lemma 13.1. *The tangential derivative only depends on the Riemannian metric I (the first fundamental form), so it is an invariant of the surface up to isometries.*[1]

[1] i.e. it does not change even if you pick a different embedding into $\mathbb{R}^3$, provided the two embedded surfaces are isometric.

Proof. This is a calculation:

$$\nabla_i v = \nabla_i \sum v^j X_j = \sum \partial_i(v^j) X_j + v^j \nabla_i X_j$$

so we just need to check that $\nabla_i X_j$ depends only on I_F. You will do this rather explicitly in Exercise Sheet 2 by expressing

$$\boxed{\nabla_i X_j = \sum \Gamma_{ij}^k X_k}$$

in terms of the basis X_k, where the functions Γ_{ij}^k are called **Christoffel symbols** and showing that there is a formula for these in terms of the first fundamental form and its derivatives. Here we will just illustrate an example: since dotting with X_j kills the normal term (as n is orthogonal to $TS = \text{span}(X_1, X_2)$), we get

$$X_1 \cdot \nabla_1 X_2 = X_1 \cdot \partial_1 X_2 = \partial_x F \cdot \partial_x \partial_y F = \partial_x F \cdot \partial_y \partial_x F = \frac{1}{2}\partial_y(\partial_x F \cdot \partial_x F) = \frac{1}{2}\partial_y A_{11}$$

where A_{11} is the first entry of the matrix A for I_F. Similarly, all $X_k \cdot \nabla_i X_j$ are determined by derivatives of A_{ij}, hence this determines $\nabla_i X_j$ (by linear algebra). □

Cultural Remark. *In Exercise Sheet 2 you prove the formula:*

$$\Gamma_{ij}^k = \frac{1}{2}\sum_\ell g^{k\ell}(\partial_i g_{j\ell} + \partial_j g_{i\ell} - \partial_\ell g_{ij})$$

*which gives Γ_{ij}^k explicitly in terms of the Riemannian metric $g_{ij} = I(X_i, X_j) = X_i \cdot X_j$ (the first fundamental form). Conversely, given any abstract surface or manifold, with a Riemannian metric, you simply define Γ_{ij}^k by that formula, thus you obtain a connection ∇ defined by $\nabla_i X_j = \sum \Gamma_{ij}^k X_k$, called **Levi-Civita connection**, which defines tangential derivatives! More of this in* **C3.3: Differentiable Manifolds**

13.2 Gauss' Theorema Egregium

Theorem 13.2. *The Gaussian curvature only depends on the first fundamental form.*

Proof 2. Let $v(x, y) \in TS$ be any non-zero local vector field (for example $v = \partial_x F$). Recall $\nabla_i v = \partial_i v - ((\partial_i v) \cdot n)n$. Consider the following expression:

$$\begin{aligned}(\nabla_x \nabla_y - \nabla_y \nabla_x)v &= \partial_x \partial_y v - (\partial_y v \cdot n)\partial_x n - \partial_y \partial_x v + (\partial_x v \cdot n)\partial_y n \\ &= -(\partial_y v \cdot n)\partial_x n + (\partial_x v \cdot n)\partial_y n\end{aligned}$$

where we dropped all the terms that were multiples of n since we know the result must be in TS, we used that v is smooth so partial derivatives commute, and we used that $\partial_i n = \nabla_i n \in TS$ (recall this is because differentiating $n \cdot n = 1$ shows that $\partial_i n$ is orthogonal to n and hence is in TS). Recall that $v \cdot \partial_i n = -\partial_i v \cdot n$ (by differentiating the orthogonality relation $v \cdot n = 0$), so the above becomes:

$$\begin{aligned}(\nabla_x \nabla_y - \nabla_y \nabla_x)v &= (v \cdot \partial_y n)\partial_x n - (v \cdot \partial_x n)\partial_y n \\ &= -(\partial_x n \times \partial_y n) \times v && \text{Cross-product tricks}[1] \\ &= \mp K\sqrt{\det I_F}\; n \times v && \text{Lemma 12.13}\end{aligned}$$

where the sign is $-$ precisely if $\partial_x F, \partial_y F$ is right-handed.

Now notice that: by Lemma 13.1 the function $(\nabla_x \nabla_y - \nabla_y \nabla_x)v$ only depends on the first fundamental form. Of course $\sqrt{\det I_F}$ only depends on the first fundamental form, but so does $\mp n \times v$ because that is just a rotation by ∓ 90 degrees of the vector v inside TS and we know by 10.7 that the first fundamental form can be used to measure angles. It follows by

[1] $(a \times b) \times c = (a \cdot c)b - (b \cdot c)a$. For example, $(e_1 \times e_2) \times e_1 = e_3 \times e_1 = e_2 = (e_1 \cdot e_1)e_2 - (e_2 \cdot e_1)e_1$.

the above formula that also K only depends on I (using that $\mp\sqrt{\det I_F}\ n \times v \neq 0$ for $v \neq 0$). *Technical Remark. The sign ambiguity above is not an issue: the choice of normal $\pm n$ affects the choice of orientation for the surface, and thus the notion of clockwise/anti-clockwise rotation by 90 degrees, but the vector $\mp n \times v$ is independent of this choice (the signs cancel).* □

Remark. *The Theorema Egregium implies that the Gaussian curvature is an* **intrinsic** *invariant, i.e. it is an isometry invariant, because it depends only on the choice of Riemannian metric on the surface (whereas other curvature invariants, such as principal curvatures, are* **extrinsic***: they depend heavily on the choice of embedding of the surface into $\mathbb{R}^3$).*

Lemma 13.3. *We record for later, the useful formula from the above proof:*

$$\boxed{(\nabla_x \nabla_y - \nabla_y \nabla_x)v = \mp K\sqrt{\det I_F}\ n \times v}$$

with $-$ sign precisely if $\partial_x F$, $\partial_y F$ is right-handed ($\det(\partial_x F | \partial_y F | n) > 0$).

13.3 Riemann curvature tensor

This Section is non-examinable.

Motivation. *A natural way to think of curvature is to ask: how much do the tangential derivatives ∇_x, ∇_y fail to commute? For a small local vector field v, if you think intuitively of $\nabla_x v, \nabla_y v$ as being small arrows in an infinitesimal parallelogram, the failure of this parallelogram to close up is measured by $\nabla_x \nabla_y v - \nabla_y \nabla_x v$. This encodes how curved the space is.*

The **Riemann curvature tensor** R^m_{ijk} is defined by:

$$\boxed{\begin{aligned} R(X_i, X_j)X_k &= \nabla_i \nabla_j X_k - \nabla_j \nabla_i X_k \\ &= \textstyle\sum R^m_{ijk} X_m \end{aligned}}$$

Cultural Remark. *We have only defined R on the basis $X_i = \partial_i F$, but the Riemann curvature tensor $R(X, Y)Z$ can be defined for general vector fields X, Y, Z. The meaning of* ***tensor*** *is that it must be linear with respect to smooth functions (not just linear with respect to constants): $R(fX, gY)hZ = fghR(X, Y)Z$ for any smooth functions f, g, h. For sake of comparison: $\nabla_X Y$ is tensorial only in X, whereas in Y it satisfies the Leibniz rule. You can now check by calculation, using the Leibniz rule for ∇, that the tensorial condition on R implies the general formula for R has an additional term:*

$$R(X, Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z$$

where the ***Lie bracket*** *$[X, Y]$ is defined by: $[\sum v^i X_i, \sum w^j X_j] = \sum v^i \partial_i(w^j) X_j - \sum w^j \partial_j(v^i) X_i$. The Lie bracket measures how much the flow of two vector fields fails to commute: you will encounter this again in* **C3.5 Lie Groups** *and secretly in* **C2.1 Lie Algebras***. In our case: $[X_i, X_j] = \partial_i(1)X_j - \partial_j(1)X_i = 0$ since 1 is a constant coefficient.*

In Exercise Sheet 3, you will show that R^m_{ijk} is completely determined by the Christoffel symbols Γ^k_{ij} and their derivatives, and by Exercise Sheet 2 the Christoffel symbols only depend on the Riemannian metric I (first fundamental form). Hence:

Theorem 13.4. *The Riemann curvature tensor R only depends on the Riemannian metric (first fundamental form), so it is an invariant of the surface up to isometries.*[1]

[1] i.e. it does not change even if you pick a different embedding into $\mathbb{R}^3$, provided the two embedded surfaces are isometric.

It's often useful to dot the above with another basis vector X_ℓ, which defines

$$\boxed{\begin{array}{rcl} R_{ijk\ell} & = & R(X_i, X_j)X_k \cdot X_\ell \\ & = & I(R(X_i, X_j)X_k, X_\ell) \end{array}}$$

The two Riemann curvature tensors are related by the lowering/raising of indices using the Riemannian metric $g_{ij} = I_{ij} = X_i \cdot X_j$, explicitly

$$R_{ijk\ell} = \sum R^m_{ijk} g_{m\ell} \qquad\qquad R^m_{ijk} = \sum R_{ijk\ell} g^{\ell m}$$

where g^{ij} is the inverse matrix of g_{ij}, thus: $\sum g^{ij} g_{jk} = \delta^i_k$.

At first, it seems that $R_{ijk\ell}$ is a lot of information ($2^4 = 16$ choices of values for the four indices), but in fact by definition it is antisymmetric in i, j, so

$$R_{ijk\ell} = -R_{jik\ell}$$

so we might as well choose $i = 1, j = 2$. Then, by Exercise Sheet 2 you know that tangential derivatives are compatible with the Riemannian metric:

$$\boxed{\partial_i I(v, w) = I(\nabla_i v, w) + I(v, \nabla_i w)}$$

so using the symmetry $\partial_j \partial_i I(v, w) = \partial_i \partial_j I(v, w)$ (since $I(v, w)$ is a smooth function), you will deduce in Exercise Sheet 3 that:

$$R_{ijk\ell} = -R_{ij\ell k}$$

so $R_{ijk\ell}$ is anti-symmetric also in k, ℓ so we might as well take $k = 1$, $\ell = 2$. Thus, only one value is interesting up to symmetries,[1] and you will show in Exercise Sheet 3 that:[2]

$$R_{1212} = -K \det I_F.$$

Theorem 13.5 (Theorema Egregium)**.** *The Gaussian curvature only depends on the first fundamental form, not on the second (so it depends on the Riemannian metric but not on the particular choice of embedding into $\mathbb{R}^3$). Indeed:*

$$K = -\frac{I(R(v, w)v, w)}{|v \times w|^2} = -\frac{I(R(v, w)v, w)}{I(v, v)I(w, w) - I(v, w)^2}$$

for any two linearly independent vectors v, w.

Proof. The first part follows from $R_{1212} = -K \det I_F$ since $R_{ijk\ell}$ only depends on I_F and derivatives of I_F. In the second part, the second equality is just the expansion of the cross-product $(v \times w) \cdot (v \times w) = (v \cdot v)(w \cdot w) - (v \cdot w)^2$. We only need to check the second part

[1] **Cultural remark.** In addition to the above symmetries, there is one last symmetry that holds in general for manifolds. From the **Jacobi identity** for Lie brackets,

$$[X, [Y, Z]] + [Y, [Z, X]] + [Z, [X, Y]] = 0$$

one obtains the **first Bianchi identity**:

$$R(X, Y)Z + R(Y, Z)X + R(Z, X)Y = 0.$$

From this one deduces the cyclic symmetry $R_{ijk\ell} + R_{jki\ell} + R_{kij\ell} = 0$. By adding the four equations you get from this, if you reorder $ijk\ell$ cyclically, and using anti-symmetries, you can deduce that $R_{j\ell ik} = R_{ikj\ell}$, or relabelling: $R_{ijk\ell} = R_{k\ell ij}$ so you can interchange the first and last pair of indices.

[2] From this it follows, using $R^m_{ijk} = \sum R_{ijk\ell} g^{\ell m}$, that

$$R^2_{121} = -Ke, \quad R^2_{122} = -Kf, \quad R^1_{121} = Kf, \quad R^1_{122} = Kg$$

where K is the Gaussian curvature, e, f, f, g are the entries of the first fundamental form I_F.

for specific $v = X_1, w = X_2$ (by linear algebra it then holds for any basis v, w: indeed just change F by the change of basis which sends X_1, X_2 to v, w). By the above:

$$\frac{I(R(X_1,X_2)X_1,X_2)}{I(X_1,X_1)I(X_2,X_2)-I(X_1,X_2)^2} = \frac{R_{1212}}{eg-f^2} = -\frac{K\det I_F}{\det I_F} = -K. \quad \square$$

Remark 13.6 (Why is that proof conceptually different?)**.** *The calculation in the above proof is the same as in Section 13.2 which calculated the Riemann curvature*

$$R(\partial_x F, \partial_y F)v = \mp K\sqrt{\det I_F}\, n \times v.$$

However, the proof in Section 13.2 made explicit reference to the normal n and the fact that S is embedded in $\mathbb{R}^3$ (even though at the end we conclude that K only depends on the embedding up to isometry), so it appears to only apply to surfaces S whose Riemannian metric arises from the first fundamental form of an embedding of S into $\mathbb{R}^3$. However, not all Riemannian metrics arise in this way, for example: the flat torus $T^2 = \mathbb{R}^2/\mathbb{Z}^2$, that is with Riemannian metric locally induced by the standard dot product in $\mathbb{R}^2$, is locally isometric to $\mathbb{R}^2$ so $II = 0$ (in particular $K = 0$), so it cannot isometrically embed into $\mathbb{R}^3$ because by Theorem 12.11 it would have to lie in a plane (more intuitively: obviously a torus embedded in $\mathbb{R}^3$ is going to be curved!). The approach of Section 13.3 is more general, since it works for any abstract smooth surface with any choice of Riemannian metric g_{ij}. Indeed, in Exercise Sheet 2 you found a formula for Γ^k_{ij} in terms of g_{ij}, this in turn defines $\nabla_{X_i}X_j$, which in turn defines $R(X_i,X_j)X_k$ and thus K. For example, this applies to the hyperbolic plane $\mathbb{H}$ without ever worrying about whether or not $\mathbb{H}$ embeds isometrically into $\mathbb{R}^3$.

The are two more famous curvatures, called **Ricci curvature tensor** Ric_{ij} and **scalar curvature** R, they are obtained from the Riemann curavature by taking traces.

The **Ricci curvature** is the trace

$$\mathrm{Ric}(X,Z) = \mathrm{trace}(Y \mapsto R(X,Y)Z)$$

So, defining $R_{ik} = \mathrm{Ric}(X_i, X_k)$, we have to put $X = X_i$, $Y = X_j$, $Z = X_k$ above, then take the j-th entry of the result, and sum over j (and also over ℓ on the right):

$$R_{ik} = \sum R^j_{ijk} = \sum R_{ijk\ell}g^{\ell j}$$

The lowering of two indices using the inverse metric g^{ij} is called a **metric trace**. Similarly, the **scalar curvature** is defined as the metric trace of Ric:

$$R = \sum g^{ik}R_{ik}$$

summing over both i, k. For surfaces,

$$\mathrm{R}_{ij} = -Kg_{ij} \qquad R = -2K.$$

The course **C3.3: Differentiable Manifolds** develops these ideas further.

14. Geodesic curvature and the Gauss-Bonnet theorem

14.1 Geodesic curvature and normal curvature

Recall that for a curve $\gamma : [0, b] \to \mathbb{R}^3$ parametrized by arc-length, we defined the curvature of γ as the norm of the acceleration:

$$\kappa(t) = \|\gamma''(t)\|.$$

However, we saw that on a surface the correct notion of differentiation (i.e. the one which only depends on the Riemannian metric, and not on the particular choice of embedding into $\mathbb{R}^3$), is the tangential derivative:

$$\begin{aligned}\nabla_t\gamma' &= \partial_t(\gamma') - (\partial_t\gamma'\cdot n)n \\ &= \gamma''(t) - I\!I(\gamma',\gamma')\,n\end{aligned}$$

where recall in Sec.12.3 we already met the normal curvature $\gamma''\cdot n = -Dn(\gamma')\cdot\gamma' = I\!I(\gamma',\gamma')$. So the acceleration breaks up into two parts, the tangential and the normal part:

$$\gamma'' = \nabla_t\gamma' + I\!I(\gamma',\gamma')\,n.$$

Correspondingly the curvature breaks up into two parts (using that n is normal to $\nabla_t\gamma' \in TS$),

$$\begin{aligned}\kappa^2 &= \|\gamma''\|^2 \\ &= \|\nabla_t\gamma'\|^2 + I\!I(\gamma',\gamma')^2 \\ &= \kappa^2_{geodesic} + \kappa^2_{normal}.\end{aligned}$$

Definition 14.1 (Geodesic curvature)**.** *The geodesic curvature of a curve γ in S parametrized by arc-length is:*

$$\boxed{\kappa_{geodesic} = \|\nabla_t\gamma'\|}$$

*A curve γ in S is a **geodesic** if $\kappa_{geodesic} = 0$.*

Example. For the plane $\mathbb{R}^2$, $I\!I = 0$ so $\kappa_{geodesic} = \kappa$. Therefore $\kappa_{geodesic} = 0$ implies $\gamma'' = 0$ and hence, integrating, $\gamma(t) = p + tv$ is a straight line. So geodesics in the plane are straight lines.

Example. For a sphere S of radius r in $\mathbb{R}^3$, we saw that the normal curvatures are $-1/r$ in all directions (as $I\!I = -\frac{1}{r}\mathrm{id}$). A great circle (a circle in S of maximal radius, r) has curvature $1/r$ and normal curvature $-1/r$, so $\kappa_{geodesic} = 0$. So great circles are geodesics: they look "straight" from the viewpoint of the surface. A circle of smaller radius $s < r$ in S has curvature $1/s$, but normal curvature $-1/r$, so $\kappa_{geodesic} = \sqrt{s^{-2} - r^{-2}} \neq 0$ So smaller circles are not geodesics.

Lemma 14.2. *For γ parametrized by arc-length,*

$$\boxed{\begin{aligned}\pm\kappa_{geodesic} &= \gamma''\cdot(n\times\gamma') \\ &= \det(n|\gamma'|\gamma'') \\ &= \nabla_t\gamma'\cdot(n\times\gamma')\end{aligned}}$$

Proof. γ is parametrized by arc-length, so $\gamma'\cdot\gamma' = 1$, so differentiating:[1]

$$\nabla_t\gamma'\cdot\gamma' = 0$$

(the tangential acceleration is perpendicular to the velocity). So $\nabla_t\gamma'$ is orthogonal to γ' and it is also orthogonal to n as it lies in TS. So $n\times\gamma'$ is parallel to $\nabla_t\gamma'$. Moreover, $n\times\gamma'$ has unit length since n,γ' have unit length and they are perpendicular (since $\gamma' \in TS$). Therefore

$$\nabla_t\gamma' = \pm\kappa_{geodesic} n\times\gamma'.$$

Since n is orthogonal to $n\times\gamma'$, we also know that $\gamma''\cdot(n\times\gamma') = \nabla_t\gamma'\cdot(n\times\gamma')$. □

Corollary 14.3. *For a curve γ in S which is parametrized by arc-length, γ is a geodesic $\Leftrightarrow \gamma''$ is normal to S. For a curve γ not parametrized by arc length, with $\gamma'(t) \neq 0$,*

$$\boxed{\textit{After arc-length reparametrization } \gamma \textit{ becomes a geodesic } \Leftrightarrow \gamma', \gamma'', n \textit{ are linearly dependent}}$$

[1] For smooth vector fields $v(t), w(t) \in T_{\gamma(t)}S$, the analogue of the compatibility equation $\partial_i I(v,w) = I(\nabla_i v, w) + I(v, \nabla_i w)$ from Section 13.3 holds: $\frac{d}{dt}(v\cdot w) = \nabla_t v\cdot w + v\cdot\nabla_t w$. Indeed $\frac{d}{dt}(v\cdot w) = v'\cdot w + v\cdot w'$, and $v' = \nabla_t v + (n\cdot v')n$ has $v'\cdot w = \nabla_t v\cdot w$ since $w \in TS$ is perpendicular to n (and similarly for $v\cdot w'$).

Proof. If γ is parametrized by arc-length this follows by the lemma since $\det(n|\gamma'|\gamma'') = 0$ precisely if n, γ', γ'' are linearly dependent. If $\widetilde{\gamma}(t) = \gamma(s(t))$ is a reparametrization of γ so that $\widetilde{\gamma}$ is parametrized by arc-length, then $\widetilde{\gamma}' = s'\gamma'(s)$ and $\widetilde{\gamma}'' = s''\gamma'(s) + (s')^2\gamma''(s)$. Since $s' > 0$, linear dependence of $n, \widetilde{\gamma}', \widetilde{\gamma}''$ is equivalent to linear dependence of n, γ', γ''. □

Example. Consider the torus $T^2 \subset \mathbb{R}^3$ parametrized by

$$F(\theta, \psi) = ((a + b\cos\psi)\cos\theta,\ (a + b\cos\psi)\sin\theta,\ b\sin\psi).$$

Consider the quotient map:

$$\mathbb{R}^2 \to S^1 \times S^1 \cong T^2,\ (\theta, \psi) \mapsto (e^{i\theta}, e^{i\psi}) \mapsto F(\theta, \psi)$$

Consider the straight line $(t, 0)$ in $\mathbb{R}^2$, which gives rise to the first circle S^1 factor $\gamma(t) = F(t, 0) = ((a + b)\cos t, (a + b)\sin t, 0)$ in T^2. Along γ, putting $\psi = 0$,

$$n = \begin{pmatrix} \cos t \\ \sin t \\ 0 \end{pmatrix} \quad \gamma' = \partial_\theta F = (a+b)\begin{pmatrix} -\sin t \\ \cos t \\ 0 \end{pmatrix} \qquad \gamma'' = \partial_{\theta\theta} F = (a+b)\begin{pmatrix} -\cos t \\ -\sin t \\ 0 \end{pmatrix}$$

so γ'' is normal so γ is a geodesic. Similarly for the other circle factor $\gamma(t) = F(0, t)$ we get γ'' is a multiple of n, so γ is a geodesic. However, it is not true that any straight line in $\mathbb{R}^2$ will give rise to a geodesic in this torus: draw a picture for the circle $\gamma(t) = F(\frac{\pi}{2}, t)$, check that n, γ', γ'' are clearly linearly independent. This is not surprising because the Riemannian metric is

$$I = (a + b\cos\psi)^2 d\theta^2 + b^2 d\psi^2,$$

so rotation in θ is an isometry (since I will be invariant), but rotation in ψ is not: indeed it cannot be because you show in Exercise Sheet 3 that the Gaussian curvature K varies in the ψ direction. This torus is therefore not "flat", in the sense that the quotient map $\mathbb{R}^2 \to T^2$ is not locally an isometry.

Example. An example of a flat torus T^2 is

$$F(\theta, \psi) = (e^{i\theta}, e^{i\psi}) \in S^1 \times S^1 \subset \mathbb{C} \times \mathbb{C} = \mathbb{R}^4$$

using the dot product from $\mathbb{R}^4$ to define the Riemannian metric, then $\mathbb{R}^2 \to T^2$ will be a local isometry, and geodesics in T^2 correspond to quotients of straight lines in $\mathbb{R}^2$.

14.2 The local Gauss-Bonnet theorem

Definition 14.4 (Signed geodesic curvature)**.** *Following Lemma 14.2, we define the* ***signed geodesic curvature*** $\kappa_g = \pm\kappa_{geodesic}$ *of a smooth curve* γ *in* S *parametrized by arc-length by*

$$\boxed{\kappa_g = \nabla_t\gamma' \cdot (n \times \gamma') = \gamma'' \cdot (n \times \gamma') = \det(n|\gamma'|\gamma'')}$$

Remark. If γ is not parametrised by arc-length, we define the geodesic curvature as that obtained for the unit-speed reparametrised curve $\widetilde{\gamma}$. This yields:

$$\kappa_g = \frac{\det(n|\gamma'|\gamma'')}{\|\gamma'\|^3}.$$

Proof: in the notation of the proof of Corollary 14.3, $\det(n|\widetilde{\gamma}'|\widetilde{\gamma}'') = (s')^3 \det(n|\gamma'|\gamma'')$ (using that the determinant is linear in each column, and the fact that $\det(n|\gamma'|\gamma') = 0$ as the columns are linearly dependent). Finally the speed $\|\widetilde{\gamma}'\| = 1$, so $s' = 1/\|\gamma'\|$.

Theorem 14.5 (Local Gauss-Bonnet Theorem)**.** *Let γ be a smooth simple*[1] *closed curve, which bounds a region R that lies entirely inside a parametrisation patch. Assume that γ travels anti-clockwise around R. Then*

$$\boxed{\int_R K\,dA = 2\pi - \int_\gamma \kappa_g\,ds}$$

where ds is the element[2] *of arc-length of γ, and dA is the element*[3] *of area of S.*

If γ travels clockwise around R, then the $-$ sign in the last term is $+$. If γ is piecewise smooth, so $\gamma'(t)$ is discontinuous at finitely many times $t = t_j$, then 2π becomes

$$2\pi - \sum \alpha_j$$

where α_j is the anti-clockwise angle (measured using I) that γ' jumps by at $t = t_j$.

We'll prove this after the examples.

Examples.

(1) For S the plane $\mathbb{R}^2$, notice that $\gamma'(s) \in S^1 \subset \mathbb{R}^2$ since it is a unit length vector. So $\gamma'(s) = (\cos\theta(s), \sin\theta(s))$ for a certain smooth angle $\theta(s)$. Choosing $n = (0,0,1)$, with $\mathbb{R}^2$ as the plane $Z = 0$ in $\mathbb{R}^3$, then $n \times \gamma' = (-\sin\theta(s), \cos\theta(s))$ is γ' rotated by 90 degrees. Finally $\gamma''(s) = \theta'(s)(-\sin\theta(s), \cos\theta(s))$, so $\kappa_g = \gamma'' \cdot (n \times \gamma') = \theta'(s)$. This confirms the theorem:

$$\int_\gamma \kappa_g\,ds = \int \theta'(s)\,ds = \theta(\text{end}) - \theta(\text{start}) = 2\pi = 2\pi - \int 0\,dA.$$

(2) For S the sphere of radius r, recall $K = \frac{1}{r^2}$. Take $\gamma =$ the equator: this is a great circle, so a geodesic, so $\kappa_g = 0$. Thus:

$$\int_{\text{upper hemisphere}} K dA = \frac{1}{r^2}(\text{Area of the hemisphere}) = 2\pi - \int 0\,ds = 2\pi.$$

Hence the area of the hemisphere is $2\pi r^2$, so the area of the sphere is $4\pi r^2$.

(3) Take any curve γ in the unit sphere S^2. This divides the sphere into two regions R_1, R_2 with areas A_1, A_2. The curve travels anti-clockwise around one region, say R_1, and clockwise around R_2. Thus:

$$\text{Area}(S^2) = \int_{R_1} K\,dA + \int_{R_2} K\,dA = (2\pi - \int_\gamma \kappa_g\,ds) + (2\pi + \int_\gamma \kappa_g\,ds) = 4\pi.$$

(4) The deep reason why the 2π appears, is because the vector field γ' does not extend to a non-vanishing vector field $v = v(x,y)$ on all of the region R. Indeed, in local coordinates $\gamma'_{loc} \in \mathbb{R}^2 = TV$ swings around by an angle 2π, and the number of times the vector field swings around (which is an integer) depends continuously on the curve, hence this integer is constant. Shrinking the curve γ_{loc} to a point, then along one of the shrinking curves the vector field v should also swing around once. But once you've shrunk the curve to a point, v is constant along the constant curve, so it swings around zero times. Contradiction.

---

[1] **simple** means it does not intersect itself, i.e. $\gamma : [a,b] \to S$ is injective.

[2] $ds = \sqrt{I(\gamma'(t), \gamma'(t)}\,dt$, recall we used this to define lengths $L(\gamma) = \int_\gamma ds$. If γ is parametrized by arc-length then $I(\gamma', \gamma') = 1$, so we just get $ds = dt$.

[3] $dA = \sqrt{\det I_F}\,dx\,dy$, recall we used this to define areas: $\text{Area}(R) = \int_R dA$.

Non-examinable *Proof of The Local Gauss-Bonnet Theorem.*

Step 0. We may assume γ is parametrized by arc-length, so $ds = dt$, and that for the parametrization $F(x, y)$, the basis $\partial_x F, \partial_y F$ is right-handed (otherwise switch the sign of y).

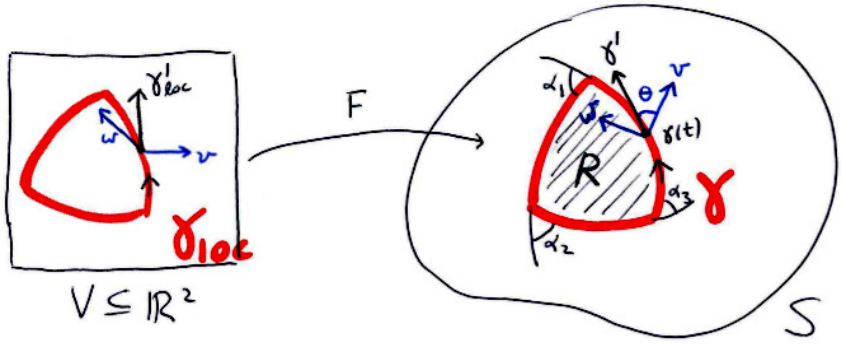
Step 1. We first build an orthonormal basis of vector fields for TS over the region R. Take $v = \partial_x F/\text{norm}$, where $\text{norm} = \sqrt{I(\partial_x F, \partial_x F)}$. Then take

$$w = n \times v = (v \text{ rotated by 90 degrees inside } TS).$$

So v, w is an orthonormal basis for TS over R with $v \times w = n$. Differentiating the relation $w \cdot w = 1$, we get that $\nabla_x w$, $\nabla_y w \in TS$ are orthogonal to w, so they are proportional to v:

$$\nabla_x w = Pv \qquad \nabla_y w = Qv$$

for some smooth functions $P = P(x, y), Q = Q(x, y)$ of the local coordinates (x, y).



Step 2. Compute the Riemann curvature of w in two ways. First by Lemma 13.3,

$$(\nabla_x \nabla_y - \nabla_y \nabla_x) w = -K\sqrt{\det I_F}\; n \times w = K\sqrt{\det I_F}\; v.$$

Then explicitly in terms of P, Q,

$$\begin{aligned} (\nabla_x \nabla_y - \nabla_y \nabla_x) w &= \nabla_x (Qv) - \nabla_y (Pv) \\ &= (\partial_x Q - \partial_y P) v + Q \nabla_x v - P \nabla_y v \\ &= (\partial_x Q - \partial_y P) v \end{aligned}$$

where in the second equality we knew the last two terms had to cancel because the result should be a multiple of v, namely $K\sqrt{\det I_F}\; v$, whereas $\nabla_x v, \nabla_y v$ are orthogonal to v (by differentiating the relation $v \cdot v = 1$).

Step 3. Now we can compute the integral of the Gaussian curvature:

$$\begin{aligned} \textstyle\int_R K\, dA &= \textstyle\int_R K\sqrt{\det I_F}\, dx\, dy \\ &= \textstyle\int_R (\nabla_x \nabla_y - \nabla_y \nabla_x) w \cdot v\, dx dy && \text{Using } v \cdot v = 1 \\ &= \textstyle\int_R (\partial_x Q - \partial_y P)\, dx\, dy \\ &= \textstyle\int_R d(P dx + Q dy) \\ &= \textstyle\int_\gamma P dx + Q dy && \text{Green's theorem in } \mathbb{R}^2 \\ &= \textstyle\int (x' P + y' Q)\, dt && \text{Meaning of } \textstyle\int_\gamma, \text{ locally } \gamma(t) = (x(t), y(t)) \\ &= \textstyle\int (x' \nabla_x w + y' \nabla_y w) \cdot v\, dt && \text{Using } v \cdot v = 1 \\ &= \textstyle\int \nabla_t w \cdot v\, dt && \text{Chain rule } \nabla_t = x' \nabla_x + y' \nabla_y \end{aligned}$$

Step 4. Now we compute κ_g in terms of v, w.

Since $\gamma'(t)$ is a unit vector in TS, we can write it as follows in the basis v, w

$$\gamma'(t) = \cos\theta(t)\, v + \sin\theta(t) w$$

where v, w are evaluated at $\gamma(t)$ of course. Then

$$\nabla_t \gamma' = \theta'(t)(-\sin\theta(t)\, v + \cos\theta(t) w) + \cos\theta(t)\, \nabla_t v + \sin\theta(t) \nabla_t w.$$

The signed geodesic curvature becomes:

$$\begin{aligned} \kappa_g &= \nabla_t \gamma' \cdot (n \times \gamma') \\ &= \nabla_t \gamma' \cdot (\cos\theta\, n \times v + \sin\theta\, n \times w) \\ &= \nabla_t \gamma' \cdot (\cos\theta\, w - \sin\theta\, v) \\ &= \theta' \cos^2\theta + \theta' \sin^2\theta + \cos^2\theta\, \nabla_t v \cdot w - \sin^2\theta \nabla_t w \cdot v \end{aligned}$$

using orthonormality of v, w, and using that $\nabla_t v$ is orthogonal to v (from differentiating $v \cdot v = 1$), similarly $\nabla_t w \cdot w = 0$. Differentiating $v \cdot w = 0$ we get $\nabla_t v \cdot w = -v \cdot \nabla_t w$. Thus:

$$\kappa_g = \theta' - \nabla_t w \cdot v$$

Step 5. Combine Step 3 and Step 4:

$$\boxed{\int_R K\, dA = \int \theta'(t)\, dt - \int \kappa_g(t)\, dt}$$

Now $\theta(t)$ is the angle between v and γ'. Passing to local coordinates, $v = \partial_x F/\text{norm} = DF(e_1)/\text{norm}$ where e_1 is the standard basis vector $(1,0)$ (so the x-direction) and $\gamma' = DF(\gamma'_{loc})$, so locally v, γ' are represented by e_1, γ'_{loc}. So $\theta(t)$ is the angle (measured using I, not the usual Euclidean angle) that the local tangent vector $\gamma'_{loc}(t) \in \mathbb{R}^2$ makes with the positive x-direction. Since $\theta(t)$ makes one full circle in the anti-clockwise direction when γ travels anti-clockwise around the simple curve, we must have $\theta(\text{end}) - \theta(\text{start}) = 2\pi$ (even though θ is not the angle we may expect with Euclidean eyes). If there are discontinuities, then the integral $\int \theta'(t)\, dt$ does not notice the jumps by α_j so it equals $2\pi - \sum \alpha_j$. When γ travels clockwise around R, simply consider the reversed path $\widetilde{\gamma}(t) = \gamma(-t)$ and notice that κ_g switches sign since $\widetilde{\gamma}'(t) = -\gamma'(-t)$ and $\widetilde{\gamma}''(t) = +\gamma''(-t)$. □

Cultural Remark. *Notice the key step in the proof is a way to pass from an integral around the boundary ∂R of R to an integral over the region R. This was Green's theorem:*

$$\int_R d\omega = \int_{\partial R} \omega$$

where $\omega = P dx + Q dy$ is a differential form. Observe that Green's theorem is the 2-dimensional analogue of the Fundamental Theorem of Calculus:

$$\int_a^b f'(x)\, dx = \int_{[a,b]} df = \int_{\partial[a,b]} f = f(b) - f(a).$$

The above Green's formula holds in great generality: R can be any smooth n-dimensional manifold with boundary, and ω can be any differential $n-1$ form (meaning ω is a sum where each term looks like

$$f\, dx \wedge dy \wedge \cdots \wedge dz$$

*where f is a smooth function, and $x, y, \ldots, z$ are any $n-1$ of the local coordinates, the symbol $\wedge$ reminds us that 1-forms anti-commute: $dx \wedge dy = -dy \wedge dx$, just like for cross-products). This generalization of Green's formula is called **Stokes's theorem**, it is arguably the most important result in geometry. More of this in* **C3.3 Differentiable Manifolds**.

14.3 The sum of the angles in a geodesic triangle

A **geodesic triangle** consists of a region R and a piecewise smooth geodesic γ moving anti-clockwise around the boundary of R having exactly three discontinuities at points $p, q, r \in S$, called the **vertices** of the triangle. So the boundary of R consists of three geodesic arcs joining the points p, q, r.

Call $\alpha_p, \alpha_q, \alpha_r$ the internal angles between the two geodesic arcs which meet at the points p, q, r (internal angle means the one swept out inside R).

Corollary 14.6. *The sum of the angles of a geodesic triangle that lies entirely in a parametrization patch is*

$$\boxed{\alpha_p + \alpha_q + \alpha_r = \pi + \int_R K \, dA.}$$

For example if $K > 0$ in R then the sum is strictly larger than π, if $K = 0$ in R then the sum is π just like in Euclidean geometry, and if $K < 0$ in R then the sum is strictly less than π.

Proof. The angles by which γ' jumps are the external angles: $\alpha_1 = \pi - \alpha_p$, $\alpha_2 = \pi - \alpha_q$, $\alpha_3 = \pi - \alpha_r$, so by Theorem 14.5 without using the geodesic assumption:

$$\begin{aligned} \alpha_p + \alpha_q + \alpha_r &= \pi + [2\pi - (\alpha_1 + \alpha_2 + \alpha_3)] \\ &= \pi + \textstyle\int_R K \, dA + \int_\gamma \kappa_g \, ds. \end{aligned}$$

When the triangle is geodesic then the last integral vanishes since $\kappa_g = 0$. □

Examples. Consider a geodesic triangle with angles α, β, γ and vertices A, B, C:

(1) **Spherical geometry.** On the unit sphere, $K = 1$, so the sum of the angles in a geodesic triangle is:

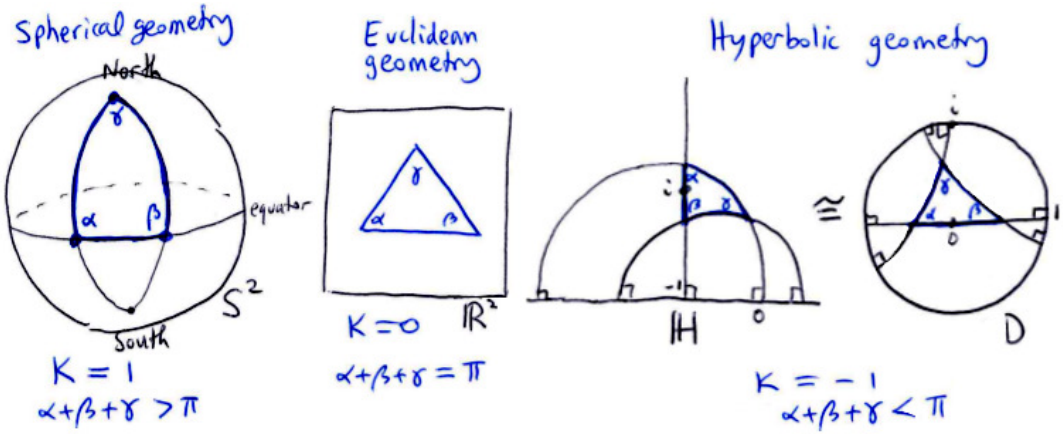
$$\alpha + \beta + \gamma = \pi + \text{Area}(ABC) \geq \pi.$$

(2) **Euclidean geometry.** In the plane, $K = 0$, so the sum of the angles is

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

(3) **Hyperbolic geometry.** In the hyperbolic plane, so $\mathbb{H} = \{z \in \mathbb{C} : \text{Im} z > 0\}$ with $I = \frac{dx^2+dy^2}{y^2} = \frac{1}{y^2} I_{\text{Euclidean}}$, we will see later that $K = -1$, so

$$\alpha + \beta + \gamma = \pi - \text{Area}(ABC) \leq \pi.$$



14.4 The Gauss-Bonnet theorem

Theorem 14.7. *For any smooth compact orientable surface S with a Riemannian metric,*

$$\chi(S) = \frac{1}{2\pi}\int_S K\,dA.$$

Example. For the unit sphere, $K = 1$ so $\int_S K\,dA = \text{Area}(S^2) = 4\pi$, so

$$\frac{1}{2\pi}\int_S K\,dA = 2 = \chi(S^2).$$

Proof. Apply the Local Gauss-Bonnet theorem to each triangle of a triangulation of S (where we subdivide the triangulation if necessary so that each triangle is small enough to lie in a parametrization patch, so that Theorem 14.5 applies). The integrals of κ_g all cancel because we integrate κ_g twice along each edge but in opposite directions (see the comments in Theorem 14.5 about anti-clockwise/clockwise curves). Thus, as in the proof of Corollary 14.6,

$$\int_R K\,dA = \sum_{\text{triangles}} (\sum(\text{internal angles}) - \pi).$$

Let V, E, F denote the total number of vertices, edges, triangles. Then the above equals:

$$2\pi V - \pi F.$$

Since each edge belongs to exactly two faces, and each face has exactly three edges, $3F = 2E$. So $2\pi V - \pi F = 2\pi V - 3\pi F + 2\pi F = 2\pi(V - E + F) = 2\pi\chi(S)$. □

15. Morse functions, Poincaré-Hopf and Hairy Ball Theorem

15.1 Critical points of functions, Morse functions, gradient vector field

A function $f(x, y)$ in two variables has a critical point at p if its first derivatives there are zero, equivalently the differential vanishes:

$$df = (\partial_x f)\,dx + (\partial_y f)\,dy = 0 \qquad \text{(evaluated at } p\text{)}.$$

At critical points, you want to know whether f has a minimum, maximum or a saddle. Recall, from calculus, that when the Hessian (evaluated at p)

$$\text{Hess}\, f = \begin{pmatrix} \partial_{xx} f & \partial_{yx} f \\ \partial_{yx} f & \partial_{yy} f \end{pmatrix}$$

is non-singular (determinant is non-zero) then you know the answer by looking at the signs[1] of the eigenvalues λ_1, λ_2 (see Section 12.5).

A critical point p is **non-degenerate** if the Hessian at p is non-singular.

Definition 15.1. *A function is **Morse** if all critical points are non-degenerate.*

It turns out that *generically* a function is Morse, in the sense that given any function (even the zero function!) if you wiggle it randomly then it will become Morse. The word "generic" has a very precise and rigorous meaning in mathematics[2]

[1] You can bypass finding the eigenvalues by first finding $\det \text{Hess} = \lambda_1\lambda_2$: if it is negative you have a saddle, otherwise it is a max/min. In the max/min case: if $\partial_{xx} f > 0$ it must be a min, if $\partial_{xx} f < 0$ it must be a max.

[2] *Non-examinable:* A **Baire set** is a set which contains a countable intersection of dense open sets. The **Baire category theorem** says that in a complete metric space every Baire set is dense. Intuitively you can think of Baire set as roughly meaning "everything outside of a measure zero set", so you have "probability 1" that a point lies in the Baire set. It turns out that one can put a topology on all smooth functions so that the Morse functions form a Baire subset.

It turns out that after a change of coordinates (namely a diagonalization argument for the Hessian), a Morse function near a critical point p always has the form:

$$f(x,y) = f(p) + \lambda_1 x^2 + \lambda_2 y^2,$$

and in fact by rescaling x, y you can assume the λ_j are ± 1 (but you cannot get rid of the signs by coordinate changes, those are invariants). So there are no other critical points near p (locally p corresponds to $(x,y) = 0$). Therefore the critical points of a Morse function f are isolated, so on a compact surface there are only finitely many critical points.

Definition 15.2 (Gradient vector field)**.** *For a smooth function $f : S \to \mathbb{R}$, the gradient vector field $\nabla f \in TS$ is defined*[1] *by the equation*

$$I(\cdot, \nabla f) = df,$$

where I is the Riemannian metric (first fundamental form).[2]

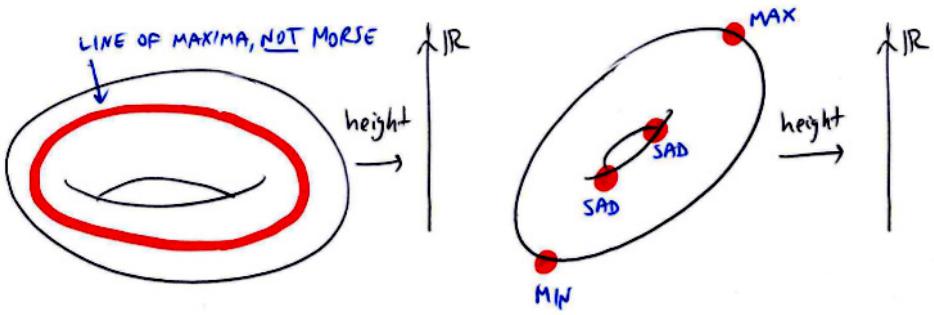
Lemma 15.3.

(1) *∇f is orthogonal to the **level sets** $f =$ constant (the contour lines),*
(2) *the points where $\nabla f = 0$ are precisely the critical points of f,*
(3) *f increases*[3] *in the direction of ∇f.*

Proof. (1): for a vector v tangent to the level set the function f does not vary, so $df(v) = 0$. Then by definition $I(v, \nabla f) = df(v) = 0$, so v is orthogonal to ∇f. (2) holds by definition. For (3): $df(\nabla f) = I(\nabla f, \nabla f) = \|\nabla f\|^2 \geq 0$ so f increases in the direction of ∇f. □

For a surface $S \subset \mathbb{R}^3$, it turns out that a linear functional $f : \mathbb{R}^3 \to \mathbb{R}$, say $f(p) = p \cdot v$ (for some fixed $v \in \mathbb{R}^3$) is a Morse function when restricted to S for almost all choices of v (but not all choices: $v = 0$ is bad). Notice that after rotating $\mathbb{R}^3$ to make $v = (0,0,1)$, you can think of these functions as just measuring the height function Z for the surface. So generically (that is, after a small generic rotation of the surface if necessary), the function $f(X,Y,Z) = Z$ becomes a Morse function.

Example. For our usual torus T^2 in $\mathbb{R}^3$, the function $f(X,Y,Z) = Z$ has a circle of maxima and a circle of minima. Since these critical points are not isolated, $f : T^2 \to \mathbb{R}$ is not Morse. But slightly rotating T^2 makes it Morse:

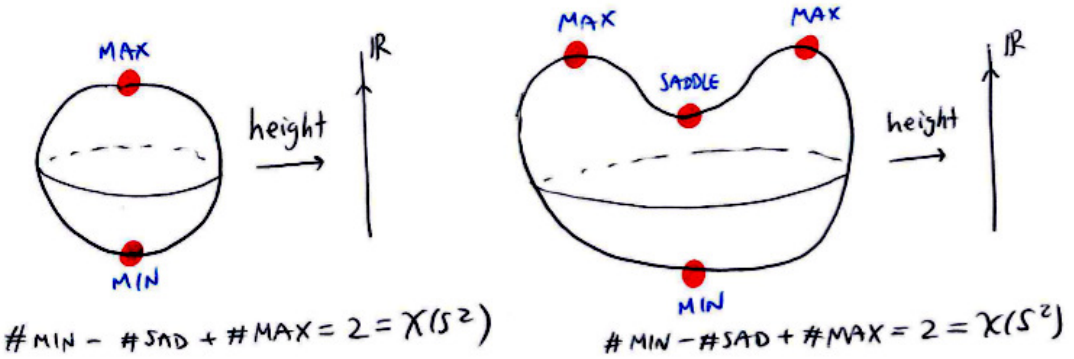


[1]**Remark.** *Explicitly, locally $A\nabla f = \begin{pmatrix} \partial_x f \\ \partial_y f \end{pmatrix} = \nabla_{Eucl} f$ where $f = f(x,y)$ in local coordinates, $A =$ the local matrix for I, and $\nabla_{Eucl} f =$ the Euclidean gradient you are used to (the first partial derivatives). So $(\nabla f)_{local} = A^{-1} \nabla_{Eucl} f \in \mathbb{R}^2 = TV$. Mapping by DF we get the resulting vector field in TS, so $\nabla f = DF(\nabla f)_{local} = DF\, A^{-1} \nabla_{Eucl} f \in TS$.*

[2]For example: evaluating on the basis vector $X_1 = \partial_x F$, we get $I(X_1, \nabla f) = df(X_1) = \partial_x f$ locally.

[3]Indeed, $\nabla f / \|\nabla f\|$ is the direction of maximal increase for f since $|df(v)| = |I(v, \nabla f)| \leq \|\nabla f\|$ for unit vectors v, by Cauchy-Schwarz, and $df(\nabla f / \|\nabla f\|) = \|\nabla f\|$ achieves equality.

Notice above that #(minima) − #(saddles) + #(maxima) = 0 = $\chi(T^2)$. Is this a coincidence? Let's check for the sphere: no matter how much you deform the sphere, a generic height function seems to always have #(minima) − #(saddles) + #(maxima) = 2 = $\chi(S^2)$:

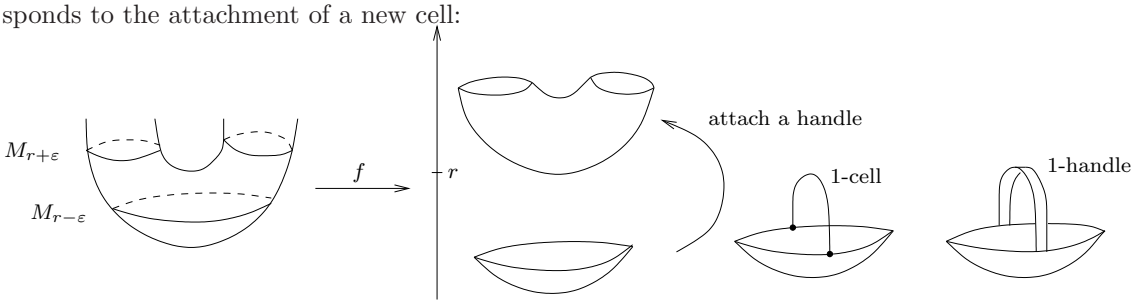


15.2 Critical points of a Morse function recover the Euler characteristic

Theorem 15.4. *For any Morse function on a compact oriented surface S,*

$$\chi(S) = \#(\text{minima}) - \#(\text{saddles}) + \#(\text{maxima}).$$

Non-examinable Proof 1: Sketch. The high-tech proof of this, is to show that a Morse function gives rise to a cellular decomposition of S (up to homotopy[1]), with 0-cells, 1-cells, 2-cells corresponding bijectively to minima, saddles and maxima. Indeed, each critical point corresponds to the attachment of a new cell:



where $M_r = \{p \in S : f(p) \leq r\}$ is the **sublevel set**. □

Proof 2. By Theorem 14.7, we need to show $\int_S K\,dA = 2\pi(\#(\text{min and max}) - \#(\text{saddles}))$.

Since f is Morse, the critical points are isolated. Let's call $m_i, s_j, M_k \in S$ the minima, saddles, and maxima. Pick small disjoint "discs" (in a parametrization patch) around each critical point, call these discs D_i, D_j, D_k and call the anti-clockwise boundary curves $\gamma_i, \gamma_j, \gamma_k$. So by the proof of Theorem 14.5, letting ℓ run over all indices i, j, k,

$$\sum_\ell \int_{D_\ell} K\,dA = \sum_\ell \int \theta'_\ell(t)\,dt - \int_{\gamma_\ell} \kappa_g\,ds$$

where θ_ℓ is the angle between v_ℓ and γ'_ℓ, and v_ℓ is corresponds to the normalized first standard basis vector in the local coordinates.

[1]You can ignore this technical issue. Otherwise see the footnote to Theorem 4.2.

On the complement $C = S \setminus$ (those discs) we cannot use Theorem 14.5 directly as it may not lie within a parametrization patch. However, we know how to build a unit vector field:

$$v_c = \nabla f / \|\nabla f\|$$

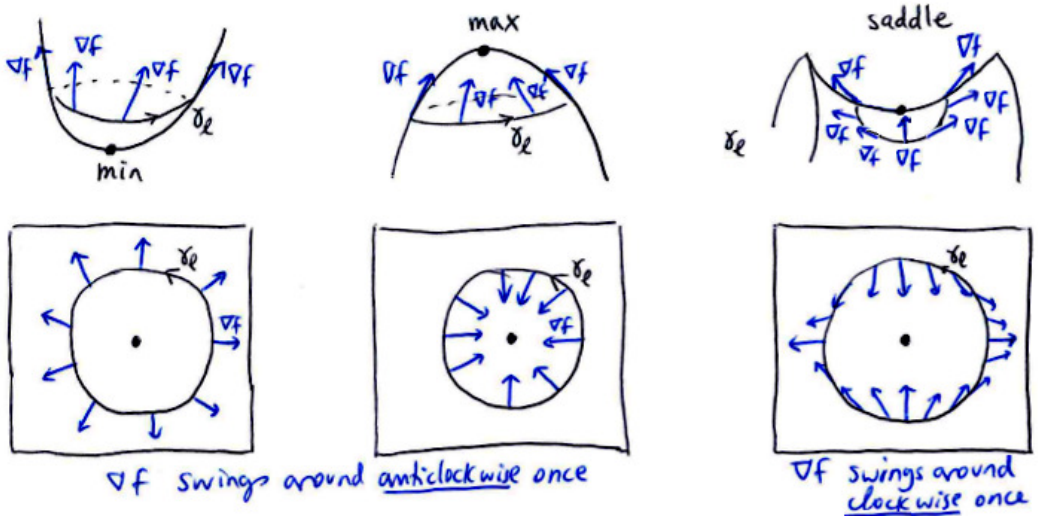
(note $\nabla f \neq 0$ on C as the critical points lie outside C). Then, as usual, define $w_c = n \times v_c$ to make v_c, w_c an orthonormal basis. The calculation in the proof of Theorem 14.5 still holds:[1]

$$\int_C K \, dA = -\sum_\ell \int \theta'_{c,\ell}(t) \, dt + \int_{\gamma_\ell} \kappa_g \, ds$$

using that the γ_ℓ are clockwise boundary curves for C, and where $\theta_{c,\ell}$ is the angle between v_c and γ'_ℓ. Summing up:

$$\int_S K \, dA = \sum_\ell \int_{D_\ell} K \, dA + \int_C K \, dA = \sum_\ell \int (\theta_\ell - \theta_{c,\ell})'(t) \, dt.$$

Observe that the angle difference $\theta_\ell - \theta_{c,\ell}$ equals the angle between v_ℓ and v_c. The total change in this angle, as you travel around γ_ℓ, is an integer multiple of 2π.



However, because we don't know the Riemannian metric (first fundamental form) we don't know what ∇f is (hence we don't know v_c), so it's hard to calculate this angle. So we use a deformation trick. If we deform γ_ℓ, then this integer must vary continuously, but being an integer it must be constant. By the same argument, we may also deform the metric: again this integer must be constant. Notice in general that if we linearly interpolate two inner products, so $t\langle\cdot,\cdot\rangle_1 + (1-t)\langle\cdot,\cdot\rangle_2$ for $0 \leq t \leq 1$, then it still satisfies bilinearity, symmetry, positive definiteness, so it is still an inner product. So we may deform the metric to the standard Euclidean metric in local coordinates. Then $(\nabla f)_{local}$ just becomes the usual Euclidean $\nabla_{Eucl} f$. Since we can choose local coordinates so that

$$f(x,y) = f(p) + \lambda_1 x^2 + \lambda_2 y^2,$$

we get $(\nabla f)_{local} = (2\lambda_1 x, 2\lambda_2 y)$ and we know $v_\ell = (1,0)$ is the first standard basis vector locally. So we just need to understand how many total rotations v_c undergoes locally (using

[1]Both signs on the right hand side are the opposite signs of those found in Step 5 of the proof of Theorem 14.5. This is because γ_ℓ is a clockwise curve bounding C, but Green's theorem requires an anti-clockwise curve. So in Step 3 of that proof, we get $\int_C K \, dA = -\int \nabla_t w \cdot v \, dt$ with a minus sign.

Euclidean eyes) as we move along γ_ℓ. This integer is called the **index of the vector field** ∇f around the zero of ∇f (the critical point of f).

We may assume that the curve γ_ℓ is locally $(x, y) = (\cos t, \sin t)$ and so along γ_ℓ:

$$v_c = \frac{1}{\sqrt{\lambda_1^2 + \lambda_2^2}} \begin{pmatrix} \lambda_1 \cos t \\ \lambda_2 \sin t \end{pmatrix}$$

If λ_1, λ_2 are both positive or both negative, then the vector v_c will swing around once anti-clockwise (notice that a global minus sign in front of the vector v_c would be the same as changing t to $t + \pi$, so it still rotates anti-clockwise). If λ_1, λ_2 have opposite signs, then changing t to $-t$ would switch the sign of $\sin t$ above and bring us back into the situation where λ_1, λ_2 have equal signs, so in this case v_c swings clockwise around once. So minima and maxima each contribute $+1$ times 2π, whereas saddles contribute -1 times 2π to $\int_S K\, dA$. □

15.3 Indices of vector fields, Poincaré-Hopf and hairy ball theorems

This Section is non-examinable.

Within the previous proof, we defined the index of the vector field ∇f, but the definition works for any vector field v on S. Given an isolated zero p of v, pick (right-handed) local coordinates near p, pick a small circular anti-clockwise path γ around p in the local coordinates, and count how many times the vector field v swings anti-clockwise around as you travel around γ (it counts as -1 times if it swings clockwise).

Remark. Let's check that the index does not depend on the observer. In another parametrization, the vector field becomes $D\tau(v)$ (with $\det D\tau > 0$ since we use right-handed parametrizations). Counting the swings in this parametrization is the same as measuring the angle in the original parametrization between v and $D\tau^{-1}(e_1)$ (rather than $e_1 = (1, 0)$). But the (non-vanishing) vector field $D\tau^{-1}(e_1)$ is defined on the whole parametrization, and we can[1] *continuously deform it into the vector field e_1. Since the index is an integer which depends continuously on the curve, the index will not change under deformations.*

Theorem 15.5 (Poincaré-Hopf theorem)**.** *Given any compact oriented surface S, and any vector field v with isolated zeros,*

$$\chi(S) = \sum_{p \in S \text{ with } v(p)=0} \text{Index}_v(p).$$

Proof. The same proof as for Theorem 15.4 applies, using the vector field v on C. □

Remark. This theorem holds for any compact oriented manifold, it even holds if the manifold has a boundary provided the vector field points orthogonally outwards along the boundary.

Example. For the torus $T^2 \cong S^1 \times S^1$, there is a non-vanishing vector field which points along the direction of one of the two circle factors (e.g. pointing in the latitudinal direction). So it has no zeros, so $\chi(T^2) = 0$.

Corollary 15.6 (hairy ball theorem)**.** *There is no nonvanishing continuous vector field on the sphere S^2. More informally: if you attempt to comb a hairy ball to make the hair flat (tangent to the surface), there will always be at least one tuft of hair somewhere.*

Proof. If the vector field had no zero, then by Poincaré-Hopf: $\chi(S^2) = 0$ which is false. □

[1] Fix a reference point, and from there move outwards radially and slowly undo the twist in $D\tau^{-1}e_1$, this will deform $D\tau^{-1}e_1$ into a constant vector field.

16. Geodesics

16.1 Differentiating vector fields defined along a curve

Let $w = w(t)$ be a vector field on S defined only along a curve $\gamma(t) \in S$, so $w(t) \in T_{\gamma(t)}S$. Recall that we defined $\nabla_t w$ to be the orthogonal projection onto TS:

$$\nabla_t w = \partial_t w - (n \cdot \partial_t w)\, n.$$

We would expect that the operator ∇_t satisfies the chain rule

$$\nabla_t = x' \nabla_x + y' \nabla_y$$

where $\gamma' = x' \partial_x F + y' \partial_y F$, using a local parametrization $F(x, y)$ (locally $\gamma'_{loc} = (x', y')$ and $\gamma' = DF(\gamma'_{loc})$). Then we can think of ∇_t as the tangential directional derivative in the direction γ', in fact one often also writes[1] $\nabla_{\gamma'}$ to mean ∇_t. However, strictly speaking $\nabla_x w$ is not defined, since w is a vector field only defined along the curve γ, so we cannot "see" how w varies in the x-direction, we only "see" how w varies in the γ' direction. The next Lemma clarifies this is not a problem (and puts it on a rigorous footing even for abstract surfaces).

Lemma 16.1 (Chain rule)**.** *Let v be any vector field on S defined in a neighbourhood of the curve $\gamma \subset S$, which extends[2] $w(t)$, i.e. satisfies $v(\gamma(t)) = w(t)$. Then*[1]

$$\boxed{\nabla_t w = \nabla_{\gamma'} v = \sum c^j \nabla_j v}$$

where $\gamma'(t) = \sum c^j X_j$ in the basis $X_j = \partial_j F$.

Proof. $\gamma'(t) = \sum c^j(t) X_j|_{\gamma(t)}$ in the basis $X_j = \partial_j F$ evaluated at $\gamma(t)$. By the chain rule,

$$\partial_t w = \partial_t(v(\gamma(t)) = Dv(\gamma'(t)) = \begin{pmatrix} \partial_1 v & \mid & \partial_2 v \end{pmatrix} \begin{pmatrix} c^1 \\ c^2 \end{pmatrix} = \sum c^j \partial_j(v).$$

So $\nabla_t w = \partial_t w - (\partial_t w \cdot n)n = \sum c^j \partial_j v - \sum c^j (\partial_j v \cdot n)\, n = \sum c^j \nabla_{X_j} v = \nabla_{\sum c^j X_j} v = \nabla_{\gamma'} v$. □

Remark 16.2. *This Lemma allows us to define $\nabla_{\gamma'}$ also for abstract smooth surfaces by turning the Lemma into a Definition. So $\nabla_t w$ or $\nabla_{\gamma'} w$ means $\nabla_{\gamma'} v$ for an extension v as above, and one checks that the choice of extension v does not matter.*[3] *Indeed, if $w(t) = \sum h^k(t)\, X_k|_{\gamma(t)}$ then the Leibniz rule holds: $\nabla_t w = \sum \partial_t(h^k)\, X_k|_{\gamma(t)} + \sum h^k \nabla_t X_k|_{\gamma(t)}$ where $\nabla_t X_k|_{\gamma(t)} = \sum c^j \nabla_j X_k|_{\gamma(t)}$ (and the latter in turn is known since $\nabla_j X_k = \sum \Gamma^i_{jk} X_i$).*

16.2 Equivalent definitions of a geodesic

Definition 16.3 (Geodesic)**.** *A smooth curve γ in S is a **geodesic** if*

$$\boxed{\nabla_{\gamma'} \gamma' = 0}$$

Recall by Section 14.1, a geodesic is a smooth curve γ in S which looks "straight" from the point of view of S, and that any of the following conditions are equivalent:

---

[1] Notice that γ' is not a local vector field since it is only defined along the curve $\gamma(t)$, so we haven't actually defined $\nabla_{\gamma'}$. However, since $\nabla_X Y$ is tensorial in X, this is not an issue. Indeed, let Z be any vector field defined near the curve which restricts to $Z|_{\gamma(t)} = \gamma'(t)$. Then we can define $\nabla_{\gamma'} Y = (\nabla_Z Y)|_{\gamma(t)}$ evaluating at $\gamma(t)$, and then we check that this equals $\sum c^j(t)\, (\nabla_{X_j} Y)|_{\gamma(t)}$ independently of the choice of Z. Indeed $(\nabla_Z Y)|_{\gamma(t)} = \nabla_{\sum Z^j(\gamma(t)) X_j} Y = \sum Z^j(\gamma(t))\, (\nabla_{X_j} Y)|_{\gamma(t)} = \sum c^j(t)\, (\nabla_{X_j} Y)|_{\gamma(t)}$.

[2] Note, by this Lemma, the tangential derivative $\nabla_{\gamma'} v$ does not depend on the choice of extension v for w.

[3] Using Christoffel symbols, recall $\nabla_i X_j = \sum \Gamma^k_{ij} X_k$, where X_k is a basis for the tangent space. Write $v = \sum f^k X_k$ in that basis, where f^k are functions. By construction, $w(t) = \sum (f^k \circ \gamma(t))\, X_k|_{\gamma(t)}$. By the Leibniz rule, $\nabla_j(f^k X_k) = f^k \nabla_j X_k + \partial_j(f^k) X_k$. By the usual chain rule, $\sum c^j \partial_j(f^k) = \partial_t(f^k \circ \gamma)$. Thus $\sum c^j \nabla_j v = \sum \partial_t(f^k \circ \gamma) X_k + \sum \sum f^k c^j \Gamma^i_{jk} X_i$ which does not depend on v.

(1) $\kappa_g = 0$: the geodesic curvature vanishes (recall $\kappa_g = \pm\|\nabla_t\gamma'\|$),
(2) $\nabla_t\gamma' = 0$: the tangential acceleration vanishes,
(3) $\nabla_{\gamma'}\gamma' = 0$ (using Lemma 16.1),
(4) $\gamma'' \cdot TS = 0$: the acceleration γ'' is normal to S (for γ parametrized by arc-length),
(5) $\gamma'' = II(\gamma', \gamma')n$ when γ is parametrized by arc-length,
(6) γ = reparametrization of a curve $\widetilde{\gamma}$ with $\widetilde{\gamma}', \widetilde{\gamma}'', n$ linearly dependent (Corollary 14.3).

Remark. Definitions (2) and (3) do not require the surface S to be embedded in $\mathbb{R}^3$: it suffices that a Riemannian metric g_{ij} is defined on S. Indeed, in Exercise Sheet 2 you found a formula for Γ^k_{ij} in terms of g_{ij}, this in turn defines $\nabla_X Y$, which in turn defines $\nabla_{\gamma'}$ and ∇_t.

Lemma 16.4. *Geodesics always have constant speed: $\|\gamma'(t)\|$ is constant.*

Proof. $\partial_t\|\gamma'(t)\|^2 = \partial_t I(\gamma', \gamma') = I(\nabla_t\gamma', \gamma') + I(\gamma', \nabla_t\gamma') = 2I(\nabla_t\gamma', \gamma') = 0.$ □

16.3 The geodesic equation in local coordinates

As in Lemma 16.1,

$$\gamma'(t) = \sum c^j(t)\, X_j|_{\gamma(t)}$$

in the basis $X_j = \partial_j F$ of TS evaluated at $\gamma(t)$. By the Leibniz rule,

$$\nabla_t(\sum c^j X_j) = \sum \partial_t(c^j) X_j + \sum c^j \nabla_t X_j.$$

By the chain rule (Lemma 16.1), summing over repeated indices,[1]

$$\nabla_t X_j = \sum c^i \nabla_i X_j = \sum c^i \Gamma^k_{ij} X_k.$$

Hence the geodesic equation $\nabla_t\gamma' = 0$ becomes

$$\partial_t c^k + \sum \Gamma^k_{ij} c^i c^j = 0.$$

Abbreviating $c^j = \dot{x}^j$, using one "dot" for each time derivative, we obtain:

Corollary 16.5. *In local coordinates $\gamma(t) = (x^1(t), x^2(t)) \in V \subset \mathbb{R}^2$ the geodesic equation is:*

$$\boxed{\ddot{x}^k + \sum \Gamma^k_{ij} \dot{x}^i \dot{x}^j = 0} \tag{16.1}$$

Exercise In local coordinates, let $A = \left(\begin{smallmatrix} e & f \\ f & g \end{smallmatrix}\right)$ be the matrix for the first fundamental form and $\gamma_{loc} = (x(t), y(t))$ a curve parametrized by arc-length, so the equation $\left(\begin{smallmatrix} x' \\ y' \end{smallmatrix}\right)^T A \left(\begin{smallmatrix} x' \\ y' \end{smallmatrix}\right) = 1$ holds, so explicitly $\boxed{ex'^2 + 2fx'y' + gy'^2 = 1}$. The geodesic equation is:

$$\frac{d}{dt}\left(A\begin{pmatrix} x' \\ y' \end{pmatrix}\right) = \frac{1}{2}\begin{pmatrix} \begin{pmatrix} x' \\ y' \end{pmatrix}^T (\partial_x A) \begin{pmatrix} x' \\ y' \end{pmatrix} \\ \begin{pmatrix} x' \\ y' \end{pmatrix}^T (\partial_y A) \begin{pmatrix} x' \\ y' \end{pmatrix} \end{pmatrix}$$

or more explicitly:

$$\boxed{\begin{array}{rcl} \frac{d}{dt}(ex' + fy') & = & \frac{1}{2}(x'^2\partial_x e + 2x'y'\partial_x f + y'^2\partial_x g) \\ \frac{d}{dt}(fx' + gy') & = & \frac{1}{2}(x'^2\partial_y e + 2x'y'\partial_y f + y'^2\partial_y g) \end{array}}$$

[1] recall that since $\nabla_i X_j \in TS$, we can write $\nabla_i X_j$ in the basis X_j and we call the coefficient functions the Christoffel symbols: $\nabla_i X_j = \sum \Gamma^k_{ij} X_k$.

Proof:[1] γ is a geodesic precisely if γ'' is normal, i.e. perpendicular to X_1, X_2. The first of the two orthogonality equations (the other is similar) is

$$0 = \gamma'' \cdot X_1 = \tfrac{d}{dt}(\gamma' \cdot X_1) - \gamma' \cdot \tfrac{d}{dt}(X_1).$$

Since $\gamma' = x'X_1 + y'X_2$, the first term is $\frac{d}{dt}(\gamma' \cdot X_1) = \frac{d}{dt}(ex' + fy')$ (the left hand side of the first equation in the above box). The second term, by the chain rule, is

$$\gamma' \cdot \tfrac{d}{dt}(X_1) = (x'X_1 + y'X_2) \cdot (x'\partial_x X_1 + y'\partial_y X_1) = (x'\partial_x F + y'\partial_y F) \cdot (x'\partial_{xx} F + y'\partial_{yx} F).$$

Now $\partial_x F \cdot \partial_{xx} F = \frac{1}{2}\partial_x e$, $\partial_x F \cdot \partial_{yx} F + \partial_y F \cdot \partial_{xx} F = \partial_x f$, $\partial_y F \cdot \partial_{yx} F = \frac{1}{2}\partial_x g$. □

Example 16.6. *When* $I = \left(\begin{smallmatrix} G & 0 \\ 0 & 1 \end{smallmatrix}\right)$, *the formulas give:* $\frac{d}{dt}(Gx') = Gx'' + \partial_x G x'^2 + \partial_y G x' y' = \frac{1}{2}x'^2 \partial_x G$ *and* $\frac{d}{dt}(y') = y'' = \frac{1}{2}x'^2\partial_y G$. *So for* $x(t) =$ *constant we can find a solution:* $y(t) = p + ct$ *for constants* p, c.

Exercise. (Harder) Using the above formulas, find the geodesics in the upper-half plane $\mathbb{H}$.

16.4 Existence and uniqueness of geodesics

Theorem 16.7. *Given a point* $p \in S$ *and a tangent direction* $v \in T_pS$, *there is a unique geodesic* $\gamma_{p,v} : (-\varepsilon, \varepsilon) \to S$, *defined for some* $\varepsilon > 0$, *through* p *with tangent vector* v *at* p:

$$\gamma_{p,v}(0) = p \qquad \text{and} \qquad \gamma'_{p,v}(0) = v.$$

Moreover, $\gamma_{p,v}$ *depends smoothly on the initial conditions* p, v.

Proof. $\ddot{x}^k + \sum \Gamma^k_{ij} \dot{x}^i \dot{x}^j = 0$ are a system of ODE's, therefore (see Analysis Handout): a solution exists, it is unique, and it depends smoothly on the initial conditions. □

Example. For $S = \mathbb{R}^2 \subset \mathbb{R}^3$, using $F(x, y) = (x, y, 0)$, we have $X_1 = (1, 0, 0)$ and $X_2 = (0, 1, 0)$, so $\nabla_i X_j = 0$, so all $\Gamma^k_{ij} = 0$, so the geodesic equation is

$$\ddot{x}^j = 0 \qquad \text{for } j = 1, 2.$$

So $(x^1, x^2) = (p^1 + v^1 t, p^2 + v^2 t) = p + tv$ is a straight line through p along v.

Example. Recall great circles in the unit sphere are geodesics. Conversely, given a point $p \in S^2$ and a direction $v \in T_pS^2$, there is a great circle through p in the direction v (namely, the intersection of the plane $\mathrm{span}(p, v)$ with S^2, or more explicitly: rotate $\mathbb{R}^3$ so that p, v become $p = (1, 0, 0)$ and $v = (0, 1, 0)$, then the geodesic is the equator). So, by uniqueness, all geodesics are great circles.

Remark. One can show that on a compact surface, the geodesic is defined for all time: $\gamma_{p,v} : \mathbb{R} \to S$. In the non-compact case, this may fail.[2]

---

[1] Exercise: Fill in the details in the following equivalent proof. For γ parametrized by arc-length, γ is a geodesic if and only if γ'' is normal, i.e. γ'' is orthogonal to $TS = \mathrm{span}(\partial_x F, \partial_y F)$. This is equivalent to:

$$DF^T \gamma'' = 0.$$

Since $\gamma' = DF\gamma'_{loc}$ and $A = DF^T DF$, show that this equation can be rewritten as:

$$\tfrac{d}{dt}(A\gamma'_{loc}) = (\tfrac{d}{dt} DF^T) DF \gamma'_{loc}.$$

Now just dot with $(1, 0)$ and $(0, 1)$ to obtain the above equations, using the chain rule $\partial_t = x'\partial_x + y'\partial_y$ and tricks like $\partial_{xy} F \cdot \partial_x F = \partial_{yx} F \cdot \partial_x F = \frac{1}{2}\partial_y(\partial_x F \cdot \partial_x F) = \frac{1}{2}\partial_y e$.

[2] For example, for the non-compact surface $\mathbb{R}^2 \setminus \{0\}$ the straight line $\gamma_{(-1,0),(1,0)}(t) = (-1 + t, 0)$ is defined for $t \in (-\infty, 1)$ but not at $t = 1$ since $(0, 0)$ does not officially belong to $\mathbb{R}^2 \setminus \{0\}$.

Theorem 16.8 (See **C3.3: Differentiable Manifolds**).

Geodesics locally minimize lengths of curves

Conversely, given a point $p \in S$, for any point q sufficiently close to p there is a unique curve γ from p to q which achieves the minimum $\inf L(\gamma)$ *(taking the infimum over smooth curves γ from p to q), and this minimizer is a geodesic.*

Example. The shortest path between two points on a sphere is the (short) arc of the great circle through those two points. Geodesics may not be global length minimizers: the full great circle, from the North Pole to the South Pole and then further to the North Pole, is a geodesic but of course the constant geodesic at the North Pole is the shortest path from the North Pole to the North Pole!

Cultural remark. *In physics, you declare that light rays move along shortest paths (locally). So light rays move along geodesics. Thus, it is no surprise that a geodesic is determined by initial position and velocity: think of pointing a flashlight from a position p in a direction v.*

16.5 Examples of geodesics via symmetries and isometries

Lemma 16.9. *Any plane of symmetry locally intersects a surface $S \subset \mathbb{R}^3$ in a geodesic.*

Proof. The surface is symmetric about the plane P, so also the normal vector must be, so $n \in P$. Let γ be a local curve where P intersects S, and parametrize γ by arc-length.[1] Then $\gamma' \in P$ and γ'' is orthogonal to γ' (differentiate $\gamma' \cdot \gamma' = 1$). As γ is invariant under the symmetry, so are γ', γ'', so $\gamma'' \in P$. So $n, \gamma'' \in P$ are both orthogonal to γ', so they are proportional. □

Example. For a surface of revolution S, any plane P through the axis of revolution is a plane of symmetry. So if the axis of revolution is the Z-axis, the curves running "vertically" are geodesics. For the sphere the curves heading "vertically" towards the North Pole are geodesics: indeed they are meridian great circles.

Theorem 16.10. *If two surfaces are locally isometric, then they have locally the same geodesics. In particular, isometries $\varphi : S_1 \to S_2$ map geodesics to geodesics:*[2]

$$\varphi \circ \gamma_{p,v} = \gamma_{\varphi(p), D\varphi(v)}.$$

Proof. The geodesic equation only depends on the tangential derivative, which only depends on the first fundamental form. □

Example. For the cylinder $X^2 + Y^2 = a^2$, $F(x, y) = (a \cos y, a \sin y, x)$ gives first fundamental form $I = dx^2 + a^2 dy^2$ which is the same as the first fundamental form for the plane $\mathbb{R}^2$ with $F(x, y) = (x, ay)$. So the cylinder is locally isometric to the plane. The geodesics for the plane are straight lines $(x(t), y(t)) = p + tv$, so each non-constant geodesic on the cylinder is a helix:

$$F(p + tv) = (a \cos(p_2 + tv_2), a \sin(p_2 + tv_2), p_1 + tv_1).$$

A very useful generalization of Lemma 16.9, but essentially based on the same idea, is:

[1] *Technical Remark.* You may worry that γ is not a regular curve, i.e. that there may be points with $\gamma' = 0$, in which case we cannot reparametrize by arc-length. Suppose $\gamma'(t_0) = 0$. Let $T = T_{\gamma(t_0)}S$ be the tangent plane at that stationary point. As S is symmetric about P, also T must be. The intersection $P \cap T$ of those two planes is a straight line. By Theorem 9.1, we can use T to build a local parametrization F for S near $\gamma(t_0)$. By picking a regular parametrization of the straight line $P \cap T \subset T$, the image via F will be a regular curve in S and we just use this curve instead of γ in the proof. (To clarify: this curve and γ have the same geometric image in S but they are parametrized differently, as we got rid of the stationary points).

[2] using the notation from Theorem 16.7.

Theorem 16.11 (Symmetry implies geodesic)**.** *If φ is a local isometry of an abstract smooth surface, such that locally the fixed points*[1] *of φ form a curve $\gamma(t)$ with $\gamma' \neq 0$ then γ becomes a geodesic after arc-length reparametrization.*

Proof. γ is fixed by φ, so $\text{span}(\gamma') \subset TS$ is fixed[2] by $D\varphi$. Since locally φ is not the identity map, the fixed locus of $D_{\gamma(t)}\varphi$ is precisely span $(\gamma'(t))$. Since the isometry φ fixes γ it follows[3] that $D\varphi(\nabla_t\gamma') = \nabla_t\gamma'$. Hence $\nabla_t\gamma' \in \text{span}(\gamma') \subset TS$. But after reparametrizing γ by arc-length, we have $\nabla_t\gamma' \cdot \gamma' = 0$ (from differentiating $\gamma' \cdot \gamma' = 1$). So $\nabla_t\gamma'$ is both proportional and perpendicular to γ' in the two-dimensional space TS, so $\nabla_t\gamma' = 0$. So γ is a geodesic. □

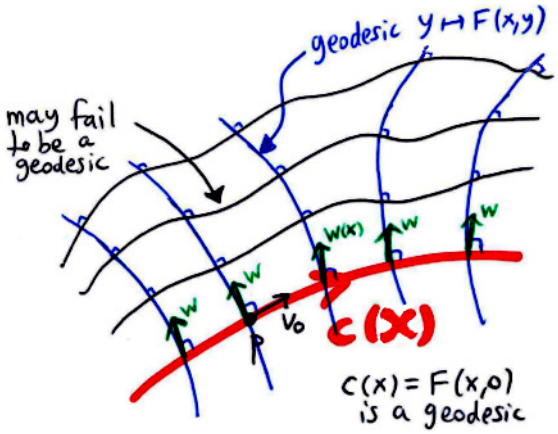
Example. For the Euclidean plane $\mathbb{R}^2$, the reflection in a straight line preserves the dot product, so it preserves the Riemannian metric. So straight lines are geodesics.

Example. For the hyperbolic plane $\mathbb{H}$, we use the hyperbolic metric $\frac{dx^2+dy^2}{y^2}$ from Section 10.9. Notice that the straight vertical line $x = 0$ (which is orthogonal to the real axis) is fixed by the reflection isometry $\varphi : \mathbb{H} \to \mathbb{H}$, $\varphi(x, y) = (-x, y)$ (indeed, put $\widetilde{x} = -x$, $\widetilde{y} = y$, then $d\widetilde{x}^2 = dx^2$, etc.). Similarly, for a straight line $x = c$, using the reflection $\varphi(x, y) = (2c - x, y)$. So vertical straight lines in $\mathbb{H}$ are hyperbolic geodesics. For straight lines which are not vertical, this does not work, since the corresponding reflections will change y and the metric will notice such changes.

17. Geodesic normal coordinates

17.1 Geodesic normal coordinates

Theorem 17.1. *Given $p \in S$, we can build a right-handed local parametrization $F(x, y)$ near p, such that $c(x) = F(x, 0)$ is a geodesic and $y \mapsto F(x, y)$ is a geodesic orthogonal to it.*



[1] so $\varphi \circ \gamma(t) = \gamma(t)$, and $\varphi(p) \neq p$ for p close to the local curve γ unless $p = \gamma(t)$ for some t. Of course, we don't want to allow the identity map $\varphi = \text{id}$ which would tell us nothing interesting. Notice that you can think of φ as the reflection in γ, locally.

[2] $\gamma' = \partial_t(\gamma) = \partial_t(\varphi \circ \gamma) = D\varphi \circ \gamma'$.

[3] Note that via an isometry, two surfaces with Riemannian metrics are to all intents and purposes identical. More pedantically: if two patches of surfaces are isometric via $\varphi : S_1 \to S_2$, then if $F : V \to S_1$ is a local parametrization, then $\varphi \circ F$ is a local parametrization for S_2, such that φ in local coordinates just becomes the identity map (indeed we use the same local coordinates x, y) and the Riemannian metric g_{ij} written locally is the same. So the curves γ in S_1 and $\varphi \circ \gamma$ in S_2 are the same curve $\gamma_{loc}(t) = (x(t), y(t))$ in local coordinates. Hence ∇_t is defined in the same way for both surfaces, and so $\nabla_t\gamma'$ is the same, using Lemma 13.1 that ∇_t only depends on the Riemannian metric (In Exercise Sheet 2 you found an explicit formula for Γ^k_{ij} in terms of the metric g_{ij}, which determines $\nabla_{X_i}X_j$, which determines ∇_t by Section 16.1).

Proof. Pick any non-zero vector $v_0 \in T_pS$ at p. Let $c(t) = \gamma_{p,v_0}(t)$ as in Theorem 16.7. Let

$$w(x) = n \times c'(x) = (c'(x) \text{ rotated by 90 degrees}).$$

We define $F(x, y)$ to be the geodesic through $c(x)$ with initial velocity $w(x)$:

$$F(x, y) = \gamma_{c(x),w(x)}(y)$$

By definition, the velocity vectors of the respective geodesics are:

$$\partial_x F|_{y=0} = c'(x) \qquad \partial_y F|_{y=0} = w(x).$$

So $\partial_x F, \partial_y F$ are linearly independent at $y = 0$ (indeed orthogonal) so they must be linearly independent also for small y close to 0. Hence F is a local parametrization for small x, y. □

Corollary 17.2. *If we pick $v_0 = c'(0)$ to have unit length, then for the above $F(x, y)$,*

$$I = G(x, y)dx^2 + dy^2$$

for the smooth function $G(x, y) = \|\partial_x F\|^2 > 0$.

Proof. If we pick v_0 to be a unit vector, then by Lemma 16.4,

$$\|c'(x)\| = \|v_0\| = 1.$$

By construction $w(x) = n \times c'(x)$ has unit length (as n, c' are orthogonal), so by Lemma 16.4,

$$\|\partial_y F(x, y)\| = \|\partial_y F(x, 0)\| = \|w(x)\| = 1.$$

We now show $\partial_x F$, $\partial_y F$ are orthogonal, i.e. that $I(\partial_x F, \partial_y F) = 0$. First we compute

$$\partial_y I(\partial_x F, \partial_y F) = I(\nabla_y \partial_x F, \partial_y F) + I(\partial_x F, \nabla_y \partial_y F).$$

Now $\nabla_y \partial_y F = 0$ since F is a geodesic in y. For the other term:

$$\begin{aligned} \nabla_y \partial_x F &= \partial_y \partial_x F - (n \cdot \partial_y \partial_x F)n \\ &= \partial_x \partial_y F - (n \cdot \partial_x \partial_y F)n \\ &= \nabla_x \partial_y F. \end{aligned}$$

So $I(\nabla_y \partial_x F, \partial_y F) = I(\nabla_x \partial_y F, \partial_y F) = \frac{1}{2}\partial_x I(\partial_y F, \partial_y F) = \frac{1}{2}\partial_x \|\partial_y F\|^2 = \frac{1}{2}\partial_x(1) = 0$. Combining the above, $\partial_y I(\partial_x F, \partial_y F) = 0$, thus $I(\partial_x F, \partial_y F)$ is constant in y and hence equal to zero (since $\partial_x F, \partial_y F$ are orthogonal at $y = 0$, by our choice of c', w). □

Warning. *$F(x, y)$ is typically not a geodesic in the x-coordinate (for $y \neq 0$), otherwise we would have $I = dx^2 + dy^2$, so S would be isometric to a plane, so $K = 0$.*

In the above proof, we showed that

$$\boxed{\nabla_y \partial_x F = \nabla_x \partial_y F}$$

This is a general symmetry property that you proved in Exercise Sheet 2: $\Gamma^k_{ij} = \Gamma^k_{ji}$.

Definition 17.3 (Geodesic normal coordinates)**.** *We call the above coordinates x, y geodesic normal coordinates (taking c' of unit length, so $I = Gdx^2 + dy^2$).*

17.2 The Gaussian curvature revisited

Theorem 17.4. *If $I = Gdx^2 + dy^2$ locally (for example in geodesic normal coordinates),*

$$K = -\frac{\partial_{yy}\sqrt{G}}{\sqrt{G}}.$$

Proof. **Claim 1.** The curves ($x =$ constant) are geodesics, so $F(x, y)$ is a geodesic in y.
Proof. See Example 16.6. We may assume $F(x, y)$ is a right-handed parametrization.
Claim 2. We can calculate the Riemann curvature in two ways.
Proof. First, by Lemma 13.3,

$$(\nabla_x\nabla_y - \nabla_y\nabla_x)\partial_y F = -K\sqrt{\det I_F}\; n \times \partial_y F = -K\sqrt{G}\; n \times \partial_y F.$$

Since $\partial_x F, \partial_y F, n$ are orthogonal, $n \times \partial_y F$ is parallel to $\partial_x F$, and taking into account orientations and lengths ($\|\partial_x F\|^2 = G$, $\|\partial_y F\|^2 = 1$):

$$n \times \partial_y F = -\partial_x F/\|\partial_x F\| = -\tfrac{1}{\sqrt{G}}\partial_x F.$$

Secondly, $\nabla_y \partial_y F = 0$ since F is a geodesic in y, and using $\nabla_x \partial_y F = \nabla_y \partial_x F$:

$$(\nabla_x\nabla_y - \nabla_y\nabla_x)\partial_y F = -\nabla_y\nabla_y\partial_x F.$$

Write $\nabla_y \partial_x F = a\partial_x F + b\partial_y F$ in the basis $\partial_x F, \partial_y F$ for TS. Using that $I = Gdx^2 + dy^2$, and that the $\partial_j F$ are orthogonal to n:

$$Ga = \partial_x F \cdot \nabla_y \partial_x F = \partial_x F \cdot \partial_y \partial_x F = \tfrac{1}{2}\partial_y(\partial_x F \cdot \partial_x F) = \tfrac{1}{2}\partial_y G,$$
$$b = \partial_y F \cdot \nabla_y \partial_x F = \partial_y F \cdot \partial_y \partial_x F = \partial_y F \cdot \partial_x \partial_y F = \tfrac{1}{2}\partial_x(\partial_y F \cdot \partial_y F) = \tfrac{1}{2}\partial_x(1) = 0.$$

Thus $\nabla_y \partial_x F = \frac{\partial_y G}{2G}\partial_x F = \frac{\partial_y \sqrt{G}}{\sqrt{G}}\partial_x F$. Combining and using the Leibniz rule:

$$\begin{aligned}
K\partial_x F \;=\; -K\sqrt{G}(-\tfrac{1}{\sqrt{G}}\partial_x F) \;&=\; -\nabla_y\nabla_y\partial_x F \\
&=\; -\nabla_y((\tfrac{\partial_y\sqrt{G}}{\sqrt{G}})\partial_x F) \\
&=\; -\partial_y(\tfrac{\partial_y\sqrt{G}}{\sqrt{G}})\partial_x F - (\tfrac{\partial_y\sqrt{G}}{\sqrt{G}})\nabla_y\partial_x F \\
&=\; [-\partial_y(\tfrac{\partial_y\sqrt{G}}{\sqrt{G}}) - (\tfrac{\partial_y\sqrt{G}}{\sqrt{G}})^2]\;\partial_x F \\
&=\; [-\tfrac{\partial_y\partial_y\sqrt{G}}{\sqrt{G}} + \tfrac{(\partial_y\sqrt{G})^2}{G} - (\tfrac{\partial_y\sqrt{G}}{\sqrt{G}})^2]\;\partial_x F \\
&=\; -\tfrac{\partial_y\partial_y\sqrt{G}}{\sqrt{G}}\;\partial_x F. \quad \square
\end{aligned}$$

Example. For the unit sphere, with $F(x, y) = (\cos x \sin y, \sin x \sin y, \cos y)$ we saw in Section 10.5 that $I = \sin^2 y\, dx^2 + dy^2$. Then, as expected:

$$K = -\frac{\partial_{yy}\sqrt{\sin^2 y}}{\sqrt{\sin^2 y}} = -\frac{\partial_{yy}\sin y}{\sin y} = \frac{\sin y}{\sin y} = 1.$$

17.3 The Gaussian curvature of the hyperbolic plane

Recall from Section 10.9 that for the hyperbolic plane $\mathbb{H} = \{z \in \mathbb{C} : \operatorname{Im} z > 0\}$, parametrizing by $F(x, y) = x + iy$, we use the Riemannian metric

$$I = \frac{1}{y^2}dx^2 + \frac{1}{y^2}dy^2 = y^{-2}dx^2 + d(\log y)^2,$$

using the trick $d\log y = (\partial_y \log y)dy = \frac{1}{y}dy$.

Corollary 17.5. *For the hyperbolic plane, $K = -1$.*

Proof. In the new coordinates $\widetilde{x} = x$, $\widetilde{y} = \log y$ the metric becomes $Gd\widetilde{x}^2 + d\widetilde{y}^2$ for

$$G(\widetilde{x}, \widetilde{y}) = y^{-2} = e^{-2\widetilde{y}}.$$

By Theorem 17.4, $K = -\frac{\partial_{\widetilde{y}\widetilde{y}}\sqrt{e^{-2\widetilde{y}}}}{\sqrt{e^{-2\widetilde{y}}}} = -\frac{\partial_{\widetilde{y}\widetilde{y}}e^{-\widetilde{y}}}{e^{-\widetilde{y}}} = -\frac{e^{-\widetilde{y}}}{e^{-\widetilde{y}}} = -1$. □

18. Surfaces of constant curvature

Lemma 18.1. *If x, y are geodesic normal coordinates, so $I = Gdx^2 + dy^2$, then*

$$G(x, 0) = 1 \qquad \text{and} \qquad \partial_y G(x, 0) = 0.$$

Proof. The curve $F(x, 0)$ is a geodesic of unit speed, so $G(x, 0) = \|\partial_x F\|^2 = 1$ for $y = 0$. Also

$$\frac{1}{2}\partial_y G = \frac{1}{2}\partial_y(\partial_x F \cdot \partial_x F) = \partial_x F \cdot \partial_y \partial_x F = \partial_x F \cdot \partial_x \partial_y F = \partial_x F \cdot \nabla_x \partial_y F = -\nabla_x \partial_x F \cdot \partial_y F$$

where the last equality follows by differentiating the orthogonality condition $\partial_x F \cdot \partial_y F = 0$. Evaluating at $y = 0$, $\nabla_x \partial_x F = 0$ since $F(x, 0)$ is a geodesic in x. So $(\partial_y G)|_{y=0} = 0$. □

Theorem 18.2.

(1) *If $K = 0$, then the surface is locally isometric to the plane,*
(2) *If $K = 1$, then the surface is locally isometric to the unit sphere,*
(3) *If $K = -1$, then the surface is locally isometric to the hyperbolic plane.*

Proof. Using geodesic normal coordinates, $I = Gdx^2 + dy^2$, by Theorem 17.4,

$$\partial_{yy}\sqrt{G} = -K\sqrt{G}$$

where $K = 0, 1, -1$ in the three respective cases.

For $K = 0$, integrating: $\sqrt{G}(x, y) = a(x)y + b(x)$ for smooth functions a, b of the other variable, so $G = (a(x)y + b(x))^2$. By Lemma 18.1, $G(x, 0) = 1$ so $b(x) = \pm 1$, and $\partial_y G(x, 0) = 0$ so $2(a(x)0 \pm 1)a(x) = 0$ so $a(x) = 0$. Thus $G = 1$, so we get the plane metric: $I = dx^2 + dy^2$.

For $K = \pm 1$ we need to solve the general equation $z'' \pm z = 0$. We know[1] the solutions: $z = a\cos t + b\sin t$ in the $+$ case, and $z = a\cosh t + b\sinh t$ in the $-$ case.[2]

So for $K = +1$, $G(x, y) = (a(x)\cos y + b(x)\sin y)^2$. By Lemma 18.1, $G(x, 0) = 1$ so $a(x) = \pm 1$, and $\partial_y G(x, 0) = 0$ so $2(a(x)1 + b(x)0)(-a(x)0 + b(x)1) = \pm 2b(x) = 0$, so $b(x) = 0$. Thus $G = \cos^2 y$, so we get the metric of a sphere: $I = \cos^2 y dx^2 + dy^2$ (by taking $\widetilde{x} = x$, $\widetilde{y} = \frac{\pi}{2} - y$ we get $\sin^2 \widetilde{y} d\widetilde{x}^2 + d\widetilde{y}^2$ as in Section 10.5).

And for $K = -1$, $G(x, y) = (a(x)\cosh y + b(x)\sinh y)^2$. By Lemma 18.1, $G(x, 0) = 1$ so $a(x) = \pm 1$, and $\partial_y G(x, 0) = 0$ so $2(a(x)1 + b(x)0)(a(x)0 + b(x)1) = \pm 2b(x) = 0$ so $b(x) = 0$. Thus $G = \cosh^2 y$, so we get the metric

$$I = \cosh^2 y\, dx^2 + dy^2.$$

We need a clever change of variables to turn this into the hyperbolic plane metric

$$\widetilde{I} = \frac{1}{\widetilde{y}^2}(d\widetilde{x}^2 + d\widetilde{y}^2).$$

---

[1] Since those are solutions, and for appropriate a, b they satisfy any initial conditions $z(0), z'(0)$, then by uniqueness of ODE solutions there are no other solutions.

[2] Refresher on hyperbolic functions: $\cosh(t) = \frac{1}{2}(e^t + e^{-t})$, $\sinh(t) = \frac{1}{2}(e^t - e^{-t})$. They satisfy the equality $\cosh^2(t) - \sinh^2(t) = 1$ and have derivatives $\cosh' = \sinh$ and $\sinh' = \cosh$.

We try:[1]

$$\widetilde{x} = e^x \tanh y \qquad \widetilde{y} = e^x \operatorname{sech} y$$

Then:

$$d\widetilde{x} = e^x \tanh y\, dx + e^x \operatorname{sech}^2 y\, dy \qquad d\widetilde{y} = e^x \operatorname{sech} y\, dx - e^x \tanh y \operatorname{sech} y\, dy$$

Squaring and adding will make the double products cancel, leaving squares:

$$\begin{aligned} d\widetilde{x}^2 + d\widetilde{y}^2 &= e^{2x}(\tanh^2 y + \operatorname{sech}^2 y)\, dx^2 + e^{2x}(\operatorname{sech}^4 y + \tanh^2 y \operatorname{sech}^2 y)\, dy^2 \\ &= e^{2x}\, dx^2 + e^{2x} \operatorname{sech}^2 y\, dy^2 \\ &= e^{2x} \operatorname{sech}^2 y\, (\cosh^2 y\, dx^2 + dy^2) \end{aligned}$$

where the coefficient $e^{2x}\operatorname{sech}^2 y = \widetilde{y}^2$ as required. □

19. Riemann surfaces: holomorphic maps and Riemann-Hurwitz

19.1 The local form of a holomorphic map between Riemann surfaces

Theorem 19.1 (Local form of a holomorphic map)**.** *For any holomorphic map $f : S \to R$ between Riemann surfaces, with $f(s) = r$, we can choose local complex coordinates around $s \in S$, $r \in R$, so that in local coordinates f is the map*[2]

$$f : D \to D,\ f(z) = z^n.$$

Proof. We can assume (by translating) that the local coordinates are chosen so that s, r correspond to $0 \in \mathbb{C}$, so in local coordinates $f(0) = 0$. The Taylor series for f is

$$f(z) = a_n z^n + a_{n+1} z^{n+1} + \cdots = a_n z^n (1 + \text{higher order terms})$$

where $a_n \neq 0$ is the first non-zero coefficient. This $n \geq 1$ is called the order of the vanishing $f(0) = 0$. A holomorphic n-th root of f is then defined[3] near 0, with $f(z)^{1/n} = a_n^{1/n} z +$ higher terms. The derivative of this n-th root at $z = 0$ is $a_n^{1/n} \neq 0$. By the inverse function theorem, there is a local holomorphic inverse $G : \mathbb{C} \to \mathbb{C}$ defined near 0, so $G(0) = 0$ and

$$f(G(z))^{1/n} = z.$$

Now we can change coordinates on the domain using the local biholomorphism G, explicitly: if $F : \mathbb{C} \to S$ (defined near $0 \in \mathbb{C}$) was the original local parametrization near $s \in S$, then the new one is $F \circ G : \mathbb{C} \to \mathbb{C} \to S$ (defined near $0 \in \mathbb{C}$). The new local expression for f becomes

$$z \mapsto f(G(z)) = z^n. \quad \square$$

In Exercise Sheet 4, you will deduce the following from Theorem 19.1:

---

[1] Second refresher on hyperbolic functions: $\tanh t = \frac{\sinh t}{\cosh t}$ and $\operatorname{sech} t = \frac{1}{\cosh t}$ have derivatives $\tanh'(t) = \operatorname{sech}^2(t)$ and $\operatorname{sech}'(t) = -\frac{\sinh t}{\cosh^2 t} = -\tanh t \operatorname{sech} t$. We will use the useful identity

$$\tanh^2 y + \operatorname{sech}^2 y = 1$$

(which follows from $\cosh^2 y - \sinh^2 y = 1$). E.g. this shows we can invert the above change of variables: $\widetilde{x}^2 + \widetilde{y}^2 = e^{2x}$ so we recover x, then we recover y.

[2] Explicitly and pedantically: there are local parametrizations $F : V \to S$, $0 \in V \subset \mathbb{R}^2$, $G : W \to R$, $0 \in W \subset \mathbb{R}^2$, $F(0) = s$, $G(0) = r$ and $f^{\text{local}}(z) = G^{-1} \circ f \circ F(z) = z^n$. We abusively just say "locally $f = \ldots$".

[3] For $a \in \mathbb{C}$, $w \in \mathbb{C}$ with $|w| < 1$, $(1 + w)^a$ has a series expansion (Newton's generalised binomial series)

$$(1 + w)^a = 1 + aw + \frac{a(a-1)}{2!} w^2 + \frac{a(a-1)(a-2)}{3!} w^3 + \cdots$$

for any $a \in \mathbb{C}$, and the series converges absolutely.

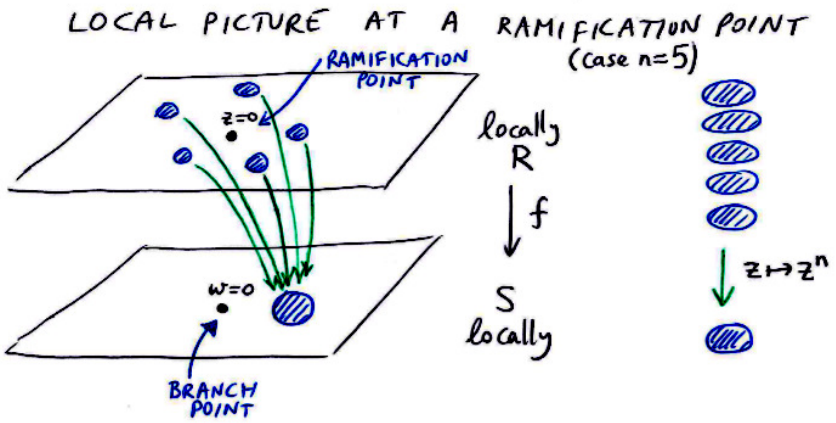
Corollary 19.2 (Open mapping theorem)**.** *A non-constant holomorphic map $f : R \to S$ between Riemann surfaces, with R connected, is an open map: f(any open set) is open.*

In Exercise 4, you'll deduce the following, for $f : R \to S$ holomorphic, R, S Riemann surfaces:

(1) If f is non-constant, R compact connected, then $f(R) \subset S$ is a connected component.
(2) If f is non-constant, R, S both compact connected, then f is surjective: $f(R) = S$.
(3) If R is compact connected, S non-compact connected, then f is constant.
(4) A holomorphic map $S \to \mathbb{C}$ on a compact connected Riemann surface is constant.
(5) Fundamental theorem of algebra: non-constant complex polynomials have a root.

So you should view the above as a powerful generalization of the fundamental theorem of algebra. The fundamental theorem of algebra in fact says that the number of roots (counting multiplicity) equals the degree of the polynomial. We now generalize this to Riemann surfaces.

19.2 Branch points and ramification points



Recall the local form Theorem 19.1: $f : R \to S$ has the local form $z \mapsto z^n$ near $p \in R$, where p corresponds to $z = 0$ locally. We call $v_f(p) = n$ the **valency** of f at p. Geometrically, it tells you how many solutions there are to the equation

$$f(z) = w$$

for small $w \neq 0$. This shows n does not depend on the choice of local coordinates near $p, f(p)$. The point $w = 0$ is bad because there are missing solutions: only $z = 0$ is a solution for small z. We call $z = 0$ a ramification point and $w = 0$ a branch point. Intuitively, a ramification point in R is where the local number of solutions of $f(z) = w$ has suddenly dropped. The value of w, when solutions are missing, is a branch point.

Definition 19.3 (Ramification point and branch point)**.** *For a point p where $v_f(p) \neq 1$:*

(1) *$p \in R$ is called* ***ramification point***
(2) *the image $f(p) \in S$ is called* ***branch point***
(3) *$v_f(p)$ is called* ***ramification index***

Equivalently:
$\diamond$ *$r \in R$ is a* ***ramification point*** $\iff$ *the derivative $f'(r) = 0$ in local coordinates,*
$\diamond$ *$s \in S$ is a* ***branch point*** $\iff$ *the preimage $f^{-1}(s) \subset R$ contains a ramification point,*
$\diamond$ *the* ***ramification index*** $= 1 +$ *number of derivatives of f vanishing at r in local coordinates.*

Lemma 19.4. *For compact R, $v_f(p) = 1$ for all except finitely many $p \in R$.*

Proof. Locally $f(z) = z^n$, so $f'(z) = nz^{n-1} \neq 0$ for $z \neq 0$ near $z = 0$. So the subset in R of points where $v_f(p) > 1$ is discrete. So it is finite when R is compact. □

19.3 The degree of a holomorphic map of compact Riemann surfaces

Definition 19.5 (Degree)**.** *The* ***degree*** *of a non-constant holomorphic map $f : R \to S$ between compact connected Riemann surfaces is*

$$\deg(f) = \sum_{r \in f^{-1}(s)} v_f(r)$$

where we fix a point $s \in S$.

Theorem 19.6. $\deg(f)$ *does not depend on the choice of point $s \in S$.*

Proof. Since R is compact, $f^{-1}(s) \subset R$ is finite (f is not[1] constantly r). Pick small disjoint discs $D_p \subset R$, one around each point $p \in f^{-1}(s)$, so that on each disc f has the local form $z \mapsto z^{v_f(p)}$. By shrinking the radii of the discs D_p, we can assume all D_p map surjectively onto the same open neighbourhood V of s. By construction, $f^{-1}(V)$ contains all the discs D_p, but may contain also other points of R. However, by further shrinking the radii of the D_p we can ensure that $f^{-1}(V) = \cup D_p$ contains nothing else.[2]

By the local model, it follows that $d = \sum_{p \in f^{-1}(s)} v_f(p)$ is the number of solutions to the equation $f(q) = w$ for $w \neq s \in V$ (in each disc D_p we find $v_f(p)$ solutions). Thus

$$|f^{-1}(w)| = \sum_{q \in f^{-1}(w)} v_f(q) = d,$$

independently of the choice of $w \neq 0 \in V$. Since S is connected, the locally constant function $w \mapsto \sum_{q \in f^{-1}(w)} v_f(q)$ on S must be constant. □

Corollary 19.7. *For all points $s \in S$ except branch points, there are precisely $\deg(f) = |f^{-1}(s)|$ points in R mapping to s.*

Example. For a complex polynomial $f(z)$ of degree d, we have $d = \deg(f)$. So there are precisely d solutions to $f(z) = 0$ unless 0 is a branch point of f. When 0 is a branch point, there are repeated roots, but if we count the roots with the correct multiplicity $v_f(p)$ then there are still d solutions.

19.4 The Riemann-Hurwitz formula

The total number of "missing solutions" that we expected for $f(r) = s$, over all $s \in S$, is:

$$b(f) = \sum_{s \in S} (\deg(f) - |f^{-1}(s)|) = \sum_{s \in S} \sum_{r \in f^{-1}(s)} (v_f(r) - 1)$$

which we call the **branching index** $b(f)$, where recall $f : R \to S$ is a holomorphic map between compact connected Riemann surfaces R, S.

---

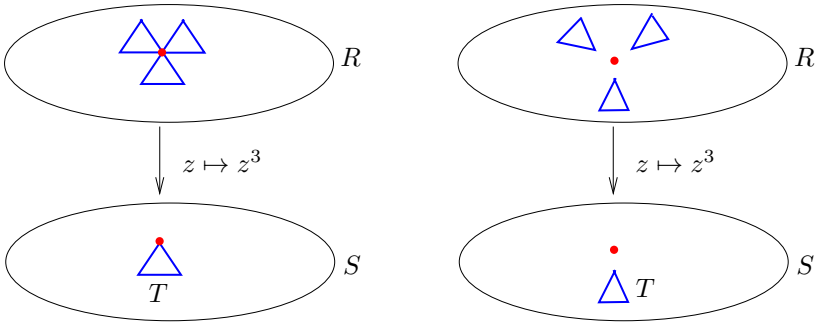
[1] If $f^{-1}(s)$ had infinitely many points, then it would have a limit point r. At this limit point r you could not have a local form of type $z \mapsto z^N$, as s has infinitely many preimages near r (in particular more than N).

[2] otherwise, by contradiction, there would be a sequence of points in R bounded away from $\cup\{p\} = f^{-1}(s)$ for which the f-values converge to s. By compactness of R a subsequence would converge to a point p' with $f(p') = s$ which we had not included in $\cup\{p\} = f^{-1}(s)$, contradiction.

Theorem 19.8 (Riemann-Hurwitz formula)**.** *For any non-constant holomorphic map $f : R \to S$ between compact connected Riemann surfaces, the Euler characteristics of R, S satisfy*

$$\chi(R) = \deg(f)\chi(S) - b(f),$$

which determines the genus $g(R)$ since $\chi(R) = 2 - 2g(R)$.



Proof. Pick a triangulation for S so that the branch points belong to the vertices of the triangulation. We want the preimage to yield a triangulation of R. So we subdivide the triangles into smaller triangles if necessary, so that each triangle $T \subset S$ lies inside an open set $V \subset S$ small enough so that $f^{-1}(V) \to V$ can be written in the usual local form on each connected component $U \subset R$ of $f^{-1}(V)$ (just as in the proof of Theorem 19.6). If the local form of $f : U \to V$ is $z \mapsto z$, so $v_f = 1$, then the preimage of T is an exact copy of that triangle, but if the local form is $z \mapsto z^n$, so $v_f = n$, then $f^{-1}(T)$ consists of n triangles (they share a vertex if a vertex of T is a branch point). The same holds for edges. However, for the vertices which are branch points we have fewer preimage vertices than expected (we lose $v_f - 1$ vertices at each branch point). So

$$V(R) = \deg(f)V(S) - b(f) \qquad E(R) = \deg(f)E(S) \qquad F(R) = \deg(f)F(S). \quad \square$$

Remark. The formula also holds when R is disconnected (apply the formula on each connected component of R). Of course S needs to be connected (can you see why?)

Example. Recall from Exercise sheet 1 that

$$R = \{(z, w) \in \mathbb{C}^2 : w^2 = (z-1)(z-2)(z-3)\} \cup \{\infty\}$$

is secretly a torus (where we compactify with a point at infinity, $(X, Y) = (0, 0)$ using $X = 1/z$, $Y = w/z^2$ at infinity, and we declare that Y is a local holomorphic coordinate - see Section 8.3). Claim: the map $f : R \to \mathbb{C}P^1$, $f(z, w) = z$ is holomorphic (in equivalent notation: the map is $f(z, w) = [1 : z]$ and $f(\infty) = [0 : 1]$).
Proof: z is a local coordinate on R when $\partial_w f \neq 0$, so when $w \neq 0$, so when $z \neq 1, 2, 3$. In that case f is locally the map $f_{loc}(z) = f(z, w) = z$, so holomorphic! When $z = 1, 2, 3$, we have $\partial_z f \neq 0$, so w is a local coordinate, so R is locally $(g(w), w)$ for a holomorphic function g, and so $f_{loc}(w) = f(g(w), w) = g(w)$ is holomorphic. Finally, at infinity, f is locally the function[1] $f_{loc}(Y) = 1/f(X, Y) = X$ so holomorphic because again R is locally a graph $X = g(Y)$ for g holomorphic, since $\partial_X f \neq 0$ at $(X, Y) = (0, 0)$. $\square$
For almost any choice of z, the number of preimages $f^{-1}(z)$ is 2 since there will be two choices of square root w. So $\deg(f) = 2$. The only times when the number of preimages drops, is if $(z-1)(z-2)(z-3) = 0$, so $z = 1, 2, 3$, and possibly at infinity. Since we only add one point

[1]More precisely: $f = [1 : f(z, w)] = [1 : z] \in \mathbb{C}P^1$ so near $z = \infty$ we must write $f = [\frac{1}{f(z,w)} : 1] \in \mathbb{C}P^1$ so $f_{loc}(Y)$ is the first entry, since that is the local coordinate that we use near the North Pole $[0 : 1] \in \mathbb{C}P^1$.

at infinity, $f^{-1}(\infty)$ only contains one point instead of two.[1] Thus $z = 1, 2, 3, \infty$ are the branch points, $(1, 0)$, $(2, 0)$, $(3, 0)$, ∞ are the ramification points, the ramification indices are all $v_f = 2$, and the branching index is $b(f) = 4 \cdot (2 - 1) = 4$. By Riemann-Hurwitz,

$$\chi(R) = \deg(f) \cdot \chi(\mathbb{C}P^1) - b(f) = 2 \cdot 2 - 4 = 0,$$

so the genus is $g(R) = 1$ so R is a torus.

Cultural Remark: Why is the Riemann-Hurwitz theorem spectacular? *A lot of mathematics you have seen so far, at least in geometry, has involved checking things that are obviously true, or at least intuitively obvious (e.g. the Jordan curve theorem). But Riemann-Hurwitz is different: you are not verifying something that you already "know" is true. You are proving a global result, namely computing the genus of the surface, by simply doing some basic local calculations (order of vanishing of local derivatives). This is geometry at its best: when you prove something geometrical that you cannot "see".*

20. Riemann surfaces: Meromorphic functions

20.1 Definition and examples

We can study Riemann surfaces S by investigating holomorphic maps from S into some test Riemann surface. Using $\mathbb{C}$ as test space is essentially useless:

Theorem 20.1 (See Section 19.1)**.** *If S is a compact Riemann surface, then holomorphic maps $S \to \mathbb{C}$ are constant on each connected component.*

You can also prove this via the maximum modulus principle,[2] and notice it generalizes[3] **Liouville's theorem** that a bounded holomorphic function $\mathbb{C} \to \mathbb{C}$ is constant.

So the simplest interesting example of maps from a Riemann surface S into a test space, is to consider the test space $\mathbb{C}P^1$:

Definition 20.2. *A **meromorphic function** is a holomorphic map $S \to \mathbb{C}P^1$, which is not identically equal to infinity on any connected component of S.*[4]

Let's unpack the definition, in local coordinates. Recall you can view

$$\mathbb{C}P^1 = \mathbb{C} \cup \{\infty\}$$

where you use the local coordinate Z for $\mathbb{C}$ (corresponding to $[Z_0 : Z_1] = [1 : Z] \in \mathbb{C}P^1$), and you use the local coordinate $W = 1/Z$ near ∞ (corresponding to $[Z_0 : Z_1] = [W : 1] \in \mathbb{C}P^1$).

Consider a point $p \in S$, and pick a local holomorphic coordinate z near p. Then

(1) If $f(p) \neq \infty$, then by continuity $f \neq \infty$ for points of S close to p, so locally

$$Z = f(z) \text{ is a holomorphic function in } z.$$

(2) If $f(p) = \infty$, then near $f(p) = \infty \in \mathbb{C}P^1$ we must use $W = 1/Z$, so

$$W = \frac{1}{f(z)} \text{ is a holomorphic function in } z.$$

[1] You could also check the local model explicitly: At infinity, R is locally $Y^2 = X(1-X)(1-2X)(1-3X)$, so for $X = 0$ there is also a drop.

[2] The continuous function $|f| : S \to \mathbb{R}$ attains a maximum at some point $p \in S$, but then in local coordinates $|f|$ would have a maximum at p so it would need to be constant near p.

[3] If $f : \mathbb{C} \to \mathbb{C}$ is bounded and holomorphic, then ∞ is a removable singularity, so f extends to $f : \mathbb{C}P^1 \to \mathbb{C}$.

[4] We exclude the constant function ∞ because we want meromorphic functions on a connected Riemann surface to form a field, so that function would be problematic. When S is disconnected, we require that meromorphic functions are not identically equal to ∞ on any connected component.

Example. Meromorphic functions on $\mathbb{C}$ are functions $f : \mathbb{C} \to \mathbb{C} \cup \{\infty\}$ such that

(1) $f(z)$ is holomorphic in z near points where $f(z) \neq \infty$,
(2) $1/f(z)$ is holomorphic in z near points where $f(z) = \infty$,

and f is not constantly ∞. The following are meromorphic:

$$f(z) = \frac{z^2 - 3z + 1}{z^3 + z^2 - z - 1} \quad f(z) = \frac{az + b}{cz + d} \quad f(z) = e^z \quad f(z) = \frac{1}{\sin z} \quad f(z) = \frac{\text{holo function of } z}{\text{holo function of } z}$$

On the other hand $f(z) = e^{1/z}$, with $f(0) = \infty$, is not meromorphic[1] near $z = 0$ because $1/f(z) = e^{-1/z} = 1 - \frac{1}{z} + \frac{1}{2!}(\frac{1}{z})^2 + (\text{higher order in } \frac{1}{z})$ is not holomorphic in z. Notice that $z = 0$ is an **essential singularity**.

Intuitively, you can think of a meromorphic function as being locally always a quotient of two holomorphic functions (see Exercise sheet 4).

Theorem 20.3. *Meromorphic functions on a Riemann surface S are precisely holomorphic functions $S \setminus P \to \mathbb{C}$, for some discrete set $P \subset S$, such that f has poles at the points in P.*

Proof. Suppose f is meromorphic. In local coordinates, suppose $f(z) = \infty$ at $z = 0$. Expand the holomorphic function $1/f(z)$ near $z = 0$ in a Taylor series:

$$\frac{1}{f(z)} = a_n z^n + a_{n+1} z^{n+1} + \cdots = a_n z^n \cdot (1 + \text{higher order } z \text{ terms})$$

Taking the reciprocal,[2] we obtain a Laurent series:

$$f(z) = a_n^{-1} \frac{1}{z^n} \cdot (1 - \text{higher order } z \text{ terms}).$$

So $f(z)$ has a **pole** at $z = 0$ of the same order as the vanishing of $1/f(z)$. The claim follows since poles are isolated (this follows from the above local expression: $f(z) \neq \infty$ except at $z = 0$). Conversely, given $f : S \setminus P \to \mathbb{C}$ as in the claim, extend by $f : S \to \mathbb{C}P^1$, $f(p) = \infty$ for all $p \in P$, then f is meromorphic since $1/f$ is holomorphic at points of P. □

Corollary 20.4. *The set of meromorphic functions on a connected[3] Riemann surface S is a field, under pointwise addition and multiplication of functions, called the* ***function field*** *$K(S)$.*

Proof. This follows by expanding $f + \lambda g$, $f \cdot g$, $1/f$ as Laurent series in local coordinates near any given point, and using Theorem 20.3. □

Cultural remark. (For those of you who like algebra and Galois theory) Studying compact connected Riemann surfaces is in fact equivalent to studying function fields $K(S)$ which are algebraic extensions of $\mathbb{C}$ of transcendence degree 1 (a purely algebraic problem). This $K(S)$ arises as the field of functions of the smooth projective curve corresponding to S (see **B3.3**).

Corollary 20.5. *For any meromorphic function on a compact Riemann surface, the number of zeros equals the number of poles (counted with multiplicity).*

Proof. $\deg(f) = \#\text{poles} = \#\text{zeros}$, counted with multiplicities v_f. □

Example. The example at the end of Sec.2.5 shows that $1/z$ is a meromorphic function on $\mathbb{C}P^1$. Meromorphic functions on $\mathbb{C}P^1$ are functions $f : \mathbb{C} \to \mathbb{C} \cup \{\infty\}$ satisfying (1) and (2) above, and

(3) if $f(\infty) \neq \infty$, $f(\frac{1}{w})$ is holomorphic in w near $w = 0$,

---

[1] in fact, it is not even continuous: consider $z = ir$ for reals $r \to 0$.

[2] Recall $\frac{1}{1+w} = 1 - w + w^2 - w^3 + \cdots$ is holomorphic in $w \in \mathbb{C}$, for $|w| < 1$.

[3] In the disconnected case, we cannot get a field as there are zero divisors: if $S = S_1 \sqcup S_2$, take $f_j = 1$ on S_j and 0 otherwise, then $f_1 \cdot f_2 = 0$.

(4) if $f(\infty)=\infty$, $1/f(\frac{1}{w})$ is holomorphic in w near $w=0$.

where $w=1/z$ is the local coordinate near $\infty\in\mathbb{C}P^1$. The following are meromorphic:

$$f(z)=\frac{z^2-3z+1}{z^3+z^2-z-1}\qquad f(z)=\frac{az+b}{cz+d}\qquad f(z)=(\text{rational function in } z)=\frac{\text{polynomial in } z}{\text{polynomial in } z}$$

However $f(z)=\frac{1}{\sin z}$ and $f(z)=e^z$ are not meromorphic at infinity. More generally if $f(z)=a_0+a_1z+a_2z^2+\cdots$ for large z, then in the local coordinate $w=1/z$: $f(\frac{1}{w})=a_0+a_1\frac{1}{w}+a_2(\frac{1}{w})^2+\cdots$ so $w=0$ is a pole (or removable) $\iff$ only finitely many a_j are non-zero (otherwise $w=0$ is an essential singularity).

Theorem 20.6. *All meromorphic functions on $\mathbb{C}P^1$ are **rational functions**.*

Proof. Recall that for a holomorphic function, zeros are isolated unless the function is identically zero (this is the **Identity theorem** – see the Analysis handout). Since $\mathbb{C}P^1$ is compact, it follows that a meromorphic function $f:\mathbb{C}P^1\to\mathbb{C}P^1$ can have only finitely many zeros $z_1,\dots,z_n\in\mathbb{C}$ (unless f is identically zero) and only finitely many poles $p_1,\dots,p_m\in\mathbb{C}$ (since those are isolated zeros of $1/f$). Let $a_1,\dots,a_n$ and $b_1,\dots,b_m$ be the orders of the zeros and of the poles. Let

$$g(z)=\prod(z-z_j)^{a_j}/\prod(z-p_k)^{b_k}.$$

Then f/g is meromorphic, and it no longer has zeros or poles in $\mathbb{C}$ (they are removable singularities). If at infinity it also does not have a pole, then it is a holomorphic map $f/g:\mathbb{C}P^1\to\mathbb{C}$ and hence is constant by Theorem 20.1. If it has a pole at infinity, then consider the reciprocal $g/f:\mathbb{C}P^1\to\mathbb{C}$, which again must be constant (here we used that f/g has no zeros in $\mathbb{C}$ so g/f has no poles in $\mathbb{C}$), so actually there was no pole. Thus $f=\text{constant}\cdot g$, so f is a quotient of polynomials. □

Notice above it is easy to verify Corollary 20.5, since $g(z)$ and its reciprocal have no pole at infinity, so $\sum a_j\le\sum b_k$ and $\sum a_j\ge\sum b_k$, hence equality.

Corollary 20.7. *The biholomorphisms $\mathbb{C}P^1\to\mathbb{C}P^1$ are precisely the Möbius maps $z\mapsto\frac{az+b}{cz+d}$ for $a,b,c,d\in\mathbb{C}$, $ad-bc\neq0$. So the group of automorphisms of $\mathbb{C}P^1$ is*

$$PSL(2,\mathbb{C})=\{\left(\begin{smallmatrix}a&b\\c&d\end{smallmatrix}\right):a,b,c,d\in\mathbb{C},ad-bc=1\}/\pm I$$

Proof. By the previous Theorem, it has to be a rational function. Having cancelled common factors, the polynomial at the numerator and that at the denominator are each allowed to have at most one root (a bijective map $\mathbb{C}P^1\to\mathbb{C}P^1$ has precisely one zero and one pole). □

20.2 Elliptic functions: meromorphic functions on the tori $\mathbb{C}$/lattice

Recall we mentioned that smooth functions $f:S^1\times S^1=\mathbb{R}^2/\mathbb{Z}^2\to\mathbb{R}$ on the torus are precisely smooth functions $\mathbb{R}^2\to\mathbb{R}$ which are 1-periodic in each entry: $f(x+n,y+m)=f(x,y)$. Analogously, given a meromorphic function on an elliptic curve $\mathbb{C}/(\mathbb{Z}\omega_1+\mathbb{Z}\omega_2)$, we patch together the local expressions for f to obtain a function

$$f:\mathbb{C}\to\mathbb{C}P^1$$

respecting the equivalence relation, so $f(z+\text{lattice point})=f(z)$. Thus

$$f(z+\omega)=f(z)\text{ for all }\omega=n\omega_1+m\omega_2\text{ where }n,m\in\mathbb{Z}.$$

Definition 20.8 (Elliptic functions)**.** *An elliptic function is a meromorphic function on $\mathbb{C}$ which is doubly periodic, that is periodic in two $\mathbb{R}$-linearly independent directions $\omega_1,\omega_2\in\mathbb{C}$.*

For this reason, $\Lambda=\mathbb{Z}\omega_1+\mathbb{Z}\omega_2$ is often called the **lattice of periods** of f.

Example. The series

$$f(z) = \sum_{\omega \in \Lambda} \frac{1}{(z-\omega)^3}$$

is an absolutely convergent series[1] at $z \notin \Lambda$. By some basic complex analysis which we recall below, the series converges absolutely and uniformly to a holomorphic function on any given compact set avoiding the poles $\omega \in \Lambda$, the series can be differentiated term by term, and the order of summation does not matter. In fact, given any compact set $K \subset \mathbb{C}$, if we omit the finitely many terms of the above series that involve poles ω that lie in K, we deduce that the rest of the series converges holomorphically, therefore the whole series has precisely the poles determined by those finitely many terms. To prove periodicity, we just need to reorder the series via $\omega \mapsto \omega - \omega'$ to deduce that $f(z) = f(z+\omega')$ for any $\omega' \in \Lambda$. At points $z = \alpha \in \Lambda$ the same argument holds if you omit $\frac{1}{(z-\alpha)^3}$ from the sum, so f has a pole of order 3 at α. So f is an elliptic function.

Some Complex Analysis Background. Suppose $f_n : U \to \mathbb{C}$ are continuous functions defined on an open set $U \subset \mathbb{C}$. **Weierstrass's M-test** says that *if* $|f_n| \leq M_n$ *on* U *with* $\sum M_n < \infty$*, then* $\sum f_n$ *converges absolutely and uniformly.* So the limit is continuous (uniform limits of continuous functions are continuous). Suppose that f_n is also holomorphic, so the partial sums $S_n(z) = \sum_{j=1}^{n} f_j(z)$ are holomorphic. We claim that $f = \lim S_n$ is holomorphic. **Morera's theorem** says that *a function* $f : U \to \mathbb{C}$ *is holomorphic if and only if* $\int_{\partial T} f(z)\,dz = 0$ *for all boundaries* ∂T *of solid triangles* $T \subset U$. Applying this to our setup:

$$\int_{\partial T} f(z)\,dz = \int_{\partial T} \lim S_n(z)\,dz = \lim \int_{\partial T} S_n(z)\,dz = 0$$

where we use a basic fact about integration: for a sequence of uniformly convergent functions, the limit commutes with the integral.[2] Finally we claim that we can differentiate the series term by term. **Cauchy's integral formula** says *if* f *is holomorphic, and* w *lies in the interior of a closed disc*[3] *entirely contained in* U*, then* $f(w) = \frac{1}{2\pi i}\int_\gamma \frac{f(z)}{z-w}\,dz$ *where* γ *is the anti-clockwise curve bounding the disc.* It is relatively straightforward to show that one can keep differentiating in w to obtain formulas for the derivatives as well, in particular:

$$f'(w) = \frac{1}{2\pi i}\int_\gamma \frac{f(z)}{(z-w)^2}\,dz.$$

Applied to our setup, $S_n'(w) = \frac{1}{2\pi i}\int \frac{S_n(z)}{(z-w)^2}\,dz$, and using again the above theorem about commuting limits and integrals, $\lim S_n'(w) = \frac{1}{2\pi i}\int_\gamma \frac{\lim S_n(z)}{(z-w)^2}\,dz = \frac{1}{2\pi i}\int_\gamma \frac{f(z)}{(z-w)^2}\,dz = f'(w)$. As an exercise,[4] show that the derivatives of the partial sums converge uniformly to f'.

20.3 The Weierstrass $\wp$-function

The **Weierstrass P-function** (or **Weierstrass's elliptic function**) is

$$\boxed{\wp(z) = \frac{1}{z^2} + \sum_{0 \neq \omega \in \Lambda} \left(\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2} \right)}$$

[1] Try checking this by hand, comparing with the convergent series $\sum_{n=1}^{\infty} \frac{1}{n^2}$ (not $\frac{1}{n^3}$). Alternatively, you can do this by an integral test, comparing with $\int (x^2+y^2)^{-3/2}\,dx\,dy$ (pass to polar coordinates).

[2] provided we integrate over a set of finite volume, in our case that is the length of ∂T which is finite.

[3] It doesn't have to be a disc, it can be any simply connected domain contained in U whose boundary is a piecewise smooth curve.

[4] Hint. Consider $|f'(w) - S_n'(w)|$ by estimating the integral, using the trick $|\int g| \leq \int |g|$.

Motivation: you cannot find a meromorphic function on $\mathbb{C}/\Lambda$ with only one simple pole since otherwise $\mathbb{C}/\Lambda$ would be biholomorphic to $\mathbb{C}P^1$ by Exercise Sheet 4, but we know that the torus and the sphere are not homeomorphic. So you need at least a pole of order 2. So you look at $1/z^2$. To make that doubly periodic, you would consider $\sum_{\omega\in\Lambda} 1/(z-\omega)^2$, but that is unfortunately divergent.[1] Intuitively, $\sum_{0\neq\omega\in\Lambda} 1/\omega^2$ ought to have a lot of cancellations (for the lattice $\mathbb{Z}\cdot 1+\mathbb{Z}\cdot i$ we expect zero due to cancellations in pairs via the symmetry $\omega\mapsto i\omega$), but sadly that series diverges for most choices of ordering of the sum. The function $\wp$ is the difference of these two divergent series: miraculously it solves all convergence issues! This difference is also a natural choice because then near $z=0$ we have $\wp(z)=\frac{1}{z^2}+g(z)$ with g holomorphic and $g(0)=0$.

Lemma 20.9. *$\wp(z)$ converges to an elliptic function, in the sense that it absolutely converges on any compact set $K\subset\mathbb{C}$ once we omit the finitely many terms with poles in K.*

Proof. Let's first prove the interesting bit, which is periodicity, assuming convergence.

Remark. The brute force approach to proving $\wp(z+\omega')=\wp(z)$ requires a clever reordering argument, because you cannot break up the series into two divergent series.

The better approach is as follows. By absolute convergence, you may differentiate the series term by term. Thus:

$$\wp'(z)=-2\sum_{\omega\in\Lambda}\frac{1}{(z-\omega)^3}$$

which is elliptic by the previous example, with order 3 poles at lattice points. Now the trick:

$$\frac{d}{dz}(\wp(z+\omega')-\wp(z))=\wp'(z+\omega')-\wp'(z)=0$$

by periodicity. So $\wp(z+\omega')-\wp(z)$ is constant. To determine that the constant is zero, plug in $z=-\omega'/2$ and use that $\wp$ is an even function: $\wp(-z)=\wp(z)$ (reorder the sum via $\omega\mapsto-\omega$).

We now prove convergence. Compute

$$\frac{1}{(z-\omega)^2}-\frac{1}{\omega^2}=\frac{\omega^2-(z-\omega)^2}{\omega^2(z-\omega)^2}=\frac{-z^2+2z\omega}{\omega^2(z-\omega)^2}=\frac{z(2\omega-z)}{\omega^2(z-\omega)^2}$$

For $|\omega|>2|z|$, the absolute value of the above is bounded by $\frac{|z|\frac{5}{2}|\omega|}{|\omega|^2\frac{1}{4}|\omega|^2}=\frac{10}{|\omega|^3}|z|$. Now z is fixed, and the condition $|\omega|>2|z|$ just omits finitely many terms from the sum. Moreover $\sum_{0\neq\omega\in\Lambda}\frac{1}{|\omega|^3}$ converges (as in the previous example, by comparing with $\sum_{n=1}^{\infty}\frac{1}{n^2}$). So, apart from finitely many terms in the sum, we deduce absolute convergence. The finitely many terms we omitted are holomorphic except for finitely many poles, so the claim follows. □

Lemma 20.10. *$\wp(z)$ has the following properties*

1. *$\wp(z)$ is an even function: $\wp(-z)=\wp(z)$,*
2. *within the fundamental parallelogram $\{s\omega_1+t\omega_2 : s,t\in[0,1]\}\subset\mathbb{C}$ modulo edge-identifications, $\wp$ has one pole at 0 of order 2 (it is harder to describe the two zeros),*
3. *$\deg(\wp)=2$, viewing $\wp$ as a map $\mathbb{C}/\Lambda\to\mathbb{C}P^1$,*
4. *$\wp'(z)=0$ at half-lattice points $\frac{\omega}{2}$, where[2] $\omega\in\Lambda$.*
5. *$\wp$ has ramification points at half-lattice points and at lattice points, so $\wp$ has precisely four distinct ramification points within the fundamental parallelogram (modulo Λ):*

$$0,\ \tfrac{1}{2}\omega_1,\ \tfrac{1}{2}\omega_2,\ \tfrac{1}{2}(\omega_1+\omega_2).$$

[1] Near a circle of radius r, centre 0, you have roughly $2\pi r$ terms, each of size roughly $\frac{1}{r^2}$, so the sum grows roughly like $2\pi\sum\frac{1}{r}$, which grows like a logarithm and thus diverges.

[2] Stricly speaking $\wp'(\omega)$ is not defined, as there is a pole, but using the local coordinate $1/(z-w)$ one would also find $\wp'=0$ (indeed those poles are ramification points).

(6) *The valencies at ramification points are $v_\wp = 2$, the branching index $b(\wp) = 4$.*

Proof. (1) follows by replacing $z \mapsto -z$ and reordering the sum via $\omega \mapsto -\omega$. (2) follows by the proof of Lemma 20.9: for z close to 0, the sum in $\wp(z)$ is holomorphic once $\frac{1}{z^2}$ is omitted. Note that the other poles $\omega_1, \omega_2, \omega_1 + \omega_2$ in the parallelogram are all identified with 0 under the translation group Λ. (3) follows by (2) since $\wp^{-1}(\infty) = 0 \in \mathbb{C}/\Lambda$ with valency $v_\wp(0) = 2$. For (4), differentiate the equation $\wp(-z) = \wp(z)$ from (1):

$$\wp'(z) = -\wp'(-z)$$

(so $\wp'$ is an odd function). But $\wp'$ is also doubly periodic, so if z satisfies $z = -z + \omega$ in the quotient $\mathbb{C}/\Lambda$ then $\wp'$ must vanish at z. Solving that equation we get $z = \frac{1}{2}\omega$. So half-lattice points satisfy $\wp'(z) = 0$ and are therefore ramification points, and since the poles have order 2 they are also ramification points ($1/\wp$ has a zero of order 2). Since valencies at ramification points satisfy $1 < v_\wp \leq \deg(\wp) = 2$, they must all be 2.

How do we know that we have found all ramification points? We use Riemann-Hurwitz:

$$0 = \chi(\text{torus}) = \chi(\mathbb{C}/\Lambda) = 2\chi(\mathbb{C}P^1) - b(\wp) = 4 - b(\wp),$$

so the branching index $b(\wp) = 4$, so all ramification points are already accounted for. □

The branch points of $\wp$ are denoted:

$$e_1 = \wp(\tfrac{1}{2}\omega_1), \quad e_2 = \wp(\tfrac{1}{2}\omega_2), \quad e_3 = \wp(\tfrac{1}{2}(\omega_1 + \omega_2)), \quad \infty = \wp(0).$$

In Exercise sheet 4 you will prove:

Theorem 20.11. *The following is a biholomorphism:*

$$\begin{array}{rcl} \mathbb{C}/\Lambda & \to & \{(Z, W) \in \mathbb{C}^2 : W^2 = 4(Z - e_1)(Z - e_2)(Z - e_3)\} \cup \{\infty\} \\ z & \mapsto & (\wp(z), \wp'(z)) \end{array}$$

where on the right we compactify as shown in Section 8.3 (compare Exercise Sheets 1 & 2).

Cultural Remark. *The function field of all meromorphic functions on an elliptic curve turns out to be $\mathbb{C}(\wp, \wp') =$ (rational functions[1] in the variables $\wp$, $\wp'$). The key trick in the proof is the fact that if you kill all the poles of a meromorpic function (e.g. by rescaling with polys in $\wp$, $\wp'$), then the result has degree 0 so it is a constant function. The **B3.3 Algebraic Curves** course studies more generally the function field of any algebraic curve.*

21. Hyperbolic Geometry: an introduction

21.1 Refresher about Möbius maps

Möbius maps for hyperbolic geometry are as important as rotations and translations are in Euclidean geometry. Indeed, for the hyperbolic plane $\mathbb{H} = \{z \in \mathbb{C} : \operatorname{Im} z > 0\}$ and the hyperbolic disc $D = \{z \in \mathbb{C} : |z| < 1\}$, a subgroup of the Möbius maps will turn out to be the group of all isometries. Recall that we found isometries in Section 10.9, between D and $\mathbb{H}$:

$$D \to \mathbb{H},\ z \mapsto \tau(z) = \frac{iz + i}{-z + 1} \qquad \mathbb{H} \to D,\ z \mapsto \tau^{-1}(z) = \frac{z - i}{z + i}.$$

Recall Corollary 20.7:

[1] i.e. ratios of polynomials.

Corollary 21.1. *The biholomorphisms $\varphi : \mathbb{C}P^1 \to \mathbb{C}P^1$ are precisely the Möbius maps*

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

for $a, b, c, d \in \mathbb{C}$ with $ad - bc \neq 0$ (we often rescale numerator and denominator so that $ad - bc = 1$). In particular, $\varphi(\infty) = a/c$, $\varphi(-d/c) = \infty$. These form a group isomorphic to

$$PSL(2,\mathbb{C}) = \{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) : a, b, c, d \in \mathbb{C}, ad - bc = 1\}/ \pm \text{Identity}.$$

We recall some useful properties about Möbius maps:

Lemma 21.2.

(1) *Möbius maps preserve angles*
(2) *Möbius maps are generated by translations $\varphi(z) = z + b$, dilations $\varphi(z) = az$ (for $a \neq 0$), and inversions $\varphi(z) = 1/z$ (which corresponds to inversion in the unit circle followed by reflection in the real axis, since $\varphi(re^{i\theta}) = \frac{1}{r}e^{-i\theta}$).*
(3) *Möbius maps send circles to circles (where we allow straight lines, thought of as circles of infinite radius).*
(4) *Given any three distinct points $z_0, z_1, z_2 \in \mathbb{C}P^1$, there is a Möbius map φ such that $\varphi(z_0) = 0$, $\varphi(z_1) = 1$, $\varphi(z_2) = \infty$.*
(5) *A Möbius map is uniquely determined by where it sends three points, for example it is determined by the values $\varphi(0)$, $\varphi(1)$, $\varphi(\infty)$.*
(6) *The Möbius maps with $\varphi(\mathbb{H}) = \mathbb{H}$, with $ad - bc = 1$, are those with a, b, c, d all real:*

$$\boxed{M\ddot{o}b(\mathbb{H}) = \left\{\frac{az+b}{cz+d} : a, b, c, d \in \mathbb{R}, ad - bc = 1\right\} \cong \{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) : a, b, c, d \in \mathbb{R}, \det = 1\}/ \pm Id = PSL(2,\mathbb{R})}$$

(7) *Given $z \in \mathbb{H}$ there is a Möbius map φ with $\varphi(\mathbb{H}) = \mathbb{H}$ and $\varphi(i) = z$.*

Proof. For (1): Möbius maps φ are holomorphic, so the derivative $D_z\varphi$ is a composition of a scaling and a rotation (see Analysis handout), so it preserves angles.

For (2), when $c \neq 0$,

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

so it is a composition of translations, dilations, inversions. The case $c = 0$ is even easier. Conversely, translations, dilations and inversions are Möbius maps.

For (3): it is enough by (2) to check that translations, dilations and inversions send circles to circles, and an easy check shows that they do.

For (4), we can even write an explicit formula:

$$\varphi(z) = \frac{(z-z_0)(z_1-z_2)}{(z-z_2)(z_1-z_0)}$$

For (5): suppose a Möbius map ψ sends distinct points w_0, w_1, w_2 to z_0, z_1, z_2. Let φ be a map as in (4). Then $\varphi \circ \psi$ is a Möbius map which sends w_0, w_1, w_2 to $0, 1, \infty$. Let α be the inverse of a map as in (4) for the points w_0, w_1, w_2, so α sends $0, 1, \infty$ to w_0, w_1, w_2. Then $\varphi \circ \psi \circ \alpha$ is a Möbius map which fixes $0, 1, \infty$. By looking at the equations this implies for the constants a, b, c, d which define the map, you easily deduce that: $b = 0$ (fixes 0), $c = 0$ (fixes ∞) and so $a/d = 1$ (fixes 1), so this map is the identity. So $\psi = \varphi^{-1}\alpha^{-1}$ is determined.

For (6): since Möbius maps are biholomorphisms, if $\varphi(\mathbb{H}) = \mathbb{H}$ then the restriction $\varphi : \mathbb{H} \to \mathbb{H}$ is a biholomorphism as well. By continuity, the boundary $\mathbb{R}$ of $\mathbb{H}$ has to be mapped into itself. As $\varphi(0) = r_0$, $\varphi(1) = r_1$, $\varphi(\infty) = r_2$ are real numbers, by the above formula we have $\varphi^{-1}(z) = \frac{(z-r_0)(r_1-r_2)}{(z-r_2)(r_1-r_0)}$. This is in $PSL(2,\mathbb{R})$ so its inverse φ is also in $PSL(2,\mathbb{R})$. Conversely,

if a, b, c, d are real, then $\varphi(\mathbb{R}) = \mathbb{R}$, hence φ permutes the two connected components of $\mathbb{C} \setminus \mathbb{R}$ (since φ is a biholomorphism). So $\varphi(\mathbb{H}) = \mathbb{H}$ precisely if $\varphi(i) \in \mathbb{H}$. So we check the sign: $\operatorname{sign} \operatorname{Im} \varphi(i) = \operatorname{sign}(ad - bc)$. So $\varphi(\mathbb{H}) = \mathbb{H}$ precisely if $\det > 0$.

For (7): $z = b + ia$ in terms of real and imaginary parts $b, a \in \mathbb{R}$, then $\varphi(z) = az + b$ works (so take $c = 0$ and $d = 1$). □

Exercise. *Recall the hyperbolic disc $D = \{z \in \mathbb{C} : |z| < 1\}$ is isometric to $\mathbb{H}$ (using the hyperbolic metric $\frac{4|dz|^2}{(1-|z|^2)^2}$ on D). Show that the Möbius maps for which $\varphi(D) = D$ are:*[1]

$$\varphi(z) = \frac{az + b}{\overline{b}z + \overline{a}}$$

with $a, b \in \mathbb{C}$ and $|a|^2 - |b|^2 = 1$. So:[2]

$$\boxed{M\ddot{o}b(D) = \left\{ \frac{az+b}{\overline{b}z+\overline{a}} : a, b \in \mathbb{C}, |a|^2 - |b|^2 = 1 \right\} \cong \{ \left(\begin{smallmatrix} a & b \\ \overline{b} & \overline{a} \end{smallmatrix}\right) : a, b \in \mathbb{C}, \det = 1 \} / \{e^{i\theta} \mathrm{Id}\} = PSU(1,1)}$$

Notice that $a \neq 0$, so you can rescale numerator and denominator so that $|a| = 1$, so you can replace $a = e^{i\theta/2}$. Then φ becomes:

$$\varphi(z) = e^{i\theta} \frac{z + b}{\overline{b}z + 1}.$$

with $b \in \mathbb{C}$ and $|b| < 1$. In particular, then $\varphi(0) = e^{i\theta} b$ so picking $b > 0 \in \mathbb{R}$ shows that there is a Möbius map with $\varphi(D) = D$ and $\varphi(0) =$ some chosen point in D.

21.2 Isometries of the hyperbolic disc D and the hyperbolic plane $\mathbb{H}$

Theorem 21.3. *The group of orientation-preserving isometries of $\mathbb{H}$ contains $M\ddot{o}b(\mathbb{H})$. The group of all isometries of $\mathbb{H}$ contains $M\ddot{o}b(\mathbb{H})$ and the reflection $z \mapsto -\overline{z}$ (so it contains the orientation-reversing isometries $\psi(z) = \frac{-a\overline{z}+b}{-c\overline{z}+d}$).*
Remark. Later we show that there are no other isometries.

Proof. We start by checking that Möb($\mathbb{H}$) are isometries. We run the same calculation as at the end of Section 10.9: we need to show

$$\frac{|dz|^2}{(\operatorname{Im} z)^2} = \frac{|d(\varphi(z))|^2}{(\operatorname{Im} \varphi(z))^2}.$$

First, $d\varphi(z) = \varphi'(z)\, dz$ where, having normalized: $ad - bc = 1$,

$$\varphi'(z) = \frac{a(cz + d) - (az + b)c}{(cz + d)^2} = \frac{1}{(cz + d)^2}.$$

Secondly,

$$\operatorname{Im} \varphi(z) = \operatorname{Im} \frac{(az + b)(c\overline{z} + d)}{|cz + d|^2} = \operatorname{Im} \frac{(ax + iay + b)(cx - icy + d)}{|cz + d|^2} = \frac{(ad - bc)\, y}{|cz + d|^2} = \frac{\operatorname{Im} z}{|cz + d|^2}.$$

The required equality above then follows.

For the last part, notice that $|d(-\overline{z})|^2 = (-d\overline{z})(-dz) = |dz|^2$, and $\operatorname{Im}(-\overline{z}) = \operatorname{Im}(z)$. □

[1] You could repeat the proof for $\mathbb{H}$ for D, so asking yourself which Möbius maps send ∂D to ∂D. The shortcut is to observe that isometries $D \to D$ arise from $D \to \mathbb{H} \to \mathbb{H} \to D$ where $\mathbb{H} \to \mathbb{H}$ are the isometries we found above, and the maps $D \to \mathbb{H}$ (and back) are the isometries τ, τ^{-1} mentioned at the start of Section 21.1. You can calculate compositions of Möbius maps by multiplying the corresponding matrices.

[2] For $SU(2)$ the $(2, 1)$ entry of the matrix would need to be $-\overline{b}$, so that $\det = |a|^2 + |b|^2$. For $SU(1, 1)$ the signature of the quadratic form has one $+$ and one $-$ sign: $\det = +|a|^2 - |b|^2$.

Exercise. *Show that Möb(D) are orientation-preserving isometries of D, and that the reflection $z \mapsto \overline{z}$ is an orientation-reversing isometry of D. Notice in particular that the rotations $z \mapsto e^{i\theta}z$ are isometries of D, and that the reflection in the line with angle θ to the real axis is $z \mapsto e^{2i\theta}\overline{z} = e^{i\theta}\overline{e^{-i\theta}z}$ so also an (orientation-reversing) isometry.*

Now the issue is: how can we show that the above are all isometries? How can we be sure we have not omitted any? The easiest route to prove this, is to use geodesics, as follows.

21.3 Geodesics of $\mathbb{H}$

Theorem 21.4. *The geodesics of the hyperbolic disc D are circles (including straight lines) which are orthogonal to the boundary $S^1 = \partial D$. The geodesics of $\mathbb{H}$ are circles (including straight lines) which are orthogonal to the boundary $\mathbb{R} = \partial\mathbb{H}$.*

Proof. By Theorem 16.11, the straight line $t \mapsto tv$ through 0 with direction $v \in \mathbb{R}^2 = T_0D$ is a geodesics in D because the (Euclidean) reflection in that straight line is an isometry of D. So these are all the geodesics $\gamma_{0,v}$ for $v \in T_0D$ by Theorem 16.7. By Lemma 21.2 (and the fact that $\mathbb{H} \cong D$ are isometric), there is a Möbius isometry $\varphi : D \to D$ which sends 0 to any chosen point $p \in D$. But $D\varphi : T_0D \to T_pD$ is bijective.[1] So for any $w \in T_pD$, $\gamma_{p,w} = \varphi \circ \gamma_{0,v}$ taking $v = D\varphi^{-1}w$. So we have obtained all geodesics in D. By Lemma 21.2, $\varphi \circ \gamma_{0,v}$ is a circle since $\gamma_{0,v}$ is a circle (line), and $\varphi \circ \gamma_{0,v}$ is orthogonal to ∂D because $\gamma_{0,v}$ is orthogonal to ∂D (using that Möbius maps preserve angles, and that $\varphi(\partial D) = \partial D$). Since $\mathbb{H} \cong D$ are isometric via a Möbius map, the claim for $\mathbb{H}$ follows by Theorem 16.10 (again using that circles map to circles via Möbius maps). □

Exercise. *Use the explicit geodesic equation, at the end of Section 16.3, to obtain explicit solutions yielding the geodesics claimed above.*

Hints. For $\mathbb{H}$, one equation becomes $\frac{d}{dt}(x'/y^2) = 0$, and the condition of being parametrized by arc-length becomes $(x'^2 + y'^2)/y^2 = 1$. You want an equation involving only x, y (not t), so calculate $dy/dx = y'/x' = \cdots$.

Corollary 21.5. *There are no other isometries for $D, \mathbb{H}$ beyond those found in Section 21.2.*

Proof. Suppose T is an isometry. Pick a $\varphi \in$ Möb(D) with $\varphi(0) = T(0)$. Then $\varphi^{-1} \circ T$ is an isometry fixing 0. Isometries map geodesics to geodesics (Theorem 16.10), and the geodesics in D through 0 are straight lines, so T permutes the straight lines through 0. Say it maps the geodesic line segment $[0, 1)$ to $e^{i\theta}[0, 1)$. Then, since isometries fix lengths, $e^{-i\theta}\varphi^{-1} \circ T$ fixes $[0, 1)$. Finally, isometries also preserve angles up to sign,[2] so $e^{-i\theta}\varphi^{-1} \circ T$ must either fix all straight lines, or it must reflect $z \mapsto \overline{z}$. The two cases are distinguished by whether or not T is orientation-preserving. This proves the statement for D. The statement for $\mathbb{H}$ follows[3] by using the Möbius isometry $D \to \mathbb{H}$. □

21.4 Hyperbolic lengths and hyperbolic angles

Theorem 21.6. *Hyperbolic angles are equal to Euclidean angles.*

Proof. Recall $I = \frac{1}{y^2}(dx^2 + dy^2)$ in the usual parametrization $F(x, y) = x + iy$ for $\mathbb{H}$, so the matrix entries of I satisfy the conditions $e = g$ and $f = 0$ required by Exercise Sheet 2

[1] By the chain rule the derivative of a diffeomorphism is bijective: $(D\varphi)^{-1} = D(\varphi^{-1})$.

[2] they preserve the Riemannian metric, so they preserve cos(angles), but $\cos\alpha = \cos(-\alpha)$.

[3] For the orientation-preserving ones, $D \to \mathbb{H} \to \mathbb{H} \to D$ is a composition of Möbius maps, but we already found all Möbius isometries of D; for the orientation-reversing ones just compose with $\mathbb{H} \to \mathbb{H}$, $z \mapsto -\overline{z}$ to reduce to the orientation-preserving case.

to guarantee that the parametrization F is conformal (i.e. angle-preserving). So hyperbolic angles equal Euclidean angles. □

Theorem 21.7. *In the hyperbolic disc D, the distance* $\operatorname{dist}_D(0,z) = 2\tanh^{-1}|z|$. *Indeed*

$$\operatorname{dist}_D(p,q) = 2\tanh^{-1}\left|\frac{q-p}{1-\overline{p}q}\right| \qquad \operatorname{dist}_{\mathbb{H}}(p,q) = 2\tanh^{-1}\left|\frac{q-p}{q-\overline{p}}\right|.$$

Proof. Recall $I = \frac{4(dx^2+dy^2)}{(1-x^2-y^2)^2}$ for D and $F(x,y) = x+iy$. The curve $\gamma(t) = (t,0)$ for $0 \le t \le x$ has $\gamma'(t) = (1,0)$, so $I(\gamma'(t),\gamma'(t)) = \frac{4}{(1-t^2)^2}$. Thus[1]

$$L(\gamma) = \int_0^x \sqrt{I(\gamma',\gamma')}\,dt = \int_0^x \frac{2}{1-t^2}\,dt = 2\tanh^{-1}x.$$

Since rotations are isometries, the formula for $\operatorname{dist}_D(0,z)$ is the same with $x = |z|$ in place of x. Given general points $p,q \in D$ we simply apply an isometry φ to move p to 0 and then use the above formula with $z = \varphi(q)$. Now $\varphi(z) = \frac{z-p}{-\overline{p}z+1}$ works (see Section 21.1). The formula for $\mathbb{H}$ follows by using the isometry $\mathbb{H} \to D$, $z \mapsto \tau^{-1}(z) = \frac{z-i}{z+i}$:

$$\operatorname{dist}_{\mathbb{H}}(p,q) = 2\tanh^{-1}\left|\frac{\frac{q-i}{q+i} - \frac{p-i}{p+i}}{1 - \frac{\overline{p}+i}{\overline{p}-i}\frac{q-i}{q+i}}\right| = 2\tanh^{-1}\left|\frac{\overline{p+i}}{p+i}\frac{p-q}{q-\overline{p}}\right| = 2\tanh^{-1}\left|\frac{q-p}{q-\overline{p}}\right|. \quad \square$$

Exercise. *Show by direct calculation that the length in $\mathbb{H}$ of the segment $(1,y)i$ on the imaginary axis is* $\log y$. *In principle you could now find a Möbius isometry sending general points $p,q \in \mathbb{H}$ to i, iy to obtain* $\operatorname{dist}_{\mathbb{H}}(p,q)$, *although this is not as easy as it was for D.*

21.5 Areas of triangles, and limits

We call **hyperbolic triangle** a geodesic triangle in the hyperbolic metric. We will call A,B,C the vertices, α,β,γ the internal angles, a,b,c the hyperbolic lenghts of the sides opposite to the vertices A,B,C, so

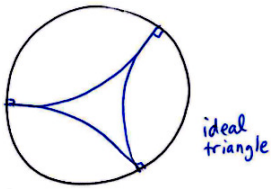
$$a = \operatorname{dist}(B,C) \qquad b = \operatorname{dist}(A,C) \qquad c = \operatorname{dist}(A,B).$$

Recall that geodesics in D passing through 0 are straight line segments, so geodesics passing through points close to 0 are almost Euclidean-straight, so small geodesic triangles near 0 are almost Euclidean. Recall that by Gauss-Bonnet (using that $K = -1$),

$$\alpha + \beta + \gamma = \pi - \operatorname{Area}(ABC)$$

so if the triangle is small, then the area is small, so the sum of the angles is almost π as expected in Euclidean geometry.

All formulas you write down in the hyperbolic world should, in the small limit, resemble formulas for Euclidean geometry. In the large limit, if we let the vertices A,B,C limit to the boundary ∂D, then the angles α,β,γ converge to zero because the geodesics are perpendicular to ∂D. Geodesic triangles with $A,B,C \in \partial D$ are called **ideal triangles**, and by Gauss-Bonnet they have area π.



[1] $\frac{d}{dx}\tanh^{-1}(x) = 1/(\tanh'(\tanh^{-1}x)) = 1/(\operatorname{sech}^2(\tanh^{-1}x)) = 1/(1-\tanh^2(\tanh^{-1}(x)) = 1/(1-x^2)$.

21.6 Hyperbolic geometry and Euclid's axioms

Non-examinable. *For the purposes of this course, the following is just a historical curiosity.*

Recall the axioms (postulates) of Euclid are:

(1) There is a unique straight line through any two distinct points,
(2) Any straight line segment can be extended to a straight line,
(3) Given a straight line segment AB, there is a unique circle with centre A passing through B,
(4) All right angles are congruent,
(5) Given a straight line ℓ and a point p outside of ℓ, there is a unique straight line through p which does not intersect ℓ.

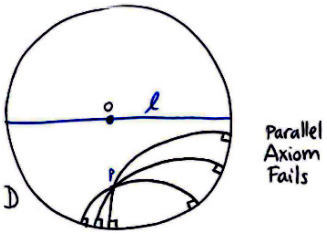
The **parallel axiom** (5) as above is called **Playfair's axiom**. Euclid's original axiom (5) is: Given a straight line ℓ, and two straight lines L_1, L_2 forming internal angles α_1, α_2 with $\alpha_1 + \alpha_2 < \pi =$ (two right angles), then L_1 and L_2 intersect. These two formulations of (5) are equivalent. There are other interesting equivalent formulations of axiom (5):

- The sum of the angles in every triangle is π (Saccheri and Legendre).
- There exists a triangle whose angles add up to π (Saccheri and Legendre).
- The sum of the angles is the same for every triangle.
- There exist two triangles which are similar but not congruent (Saccheri).
- Given a triangle, one can construct a similar triangle of any size (Wallis).
- There exists a triangle of arbitrarily large area (Gauss).

Example. In spherical geometry, that is the unit sphere S^2 with the chordal metric as in Exercise Sheet 2, axioms (2), (3) and (4) hold. Axiom (1) fails: there are infinitely many geodesics joining the North Pole to the South Pole, and axiom (5) fails: any two geodesics will intersect. We can make axiom (1) work, by using $\mathbb{R}P^2$ instead of S^2: that is, we identify antipodal points. Then only axiom (5) fails. This is called **elliptic geometry**, and is a non-Euclidean geometry. It should be noted that the Saccheri-Legendre theorem (the sum of the angles in a triangle is at most π) holds in hyperbolic geometry, but not in elliptic geometry, because parallel lines do not exist in elliptic geometry. The construction of parallel lines in geometry uses the fact that a straight line divides the plane into two connected components, but the geodesic from the North to the South Pole in $\mathbb{R}P^2$ does not disconnect $\mathbb{R}P^2$. The proof of the existence of a parallel line in Euclidean geometry tacitly assumes the **axiom of order**: given three points A, B, C on a straight line, one and only one point is "in between" the other two points. This fails for $\mathbb{R}P^2$: can you see why?

Theorem 21.8. *The hyperbolic disc D satisfies all of Euclid's axioms (including the axiom of order) except for axiom (5).*

The only thing we still need to prove is axiom (3), which follows from the next Lemma.



Lemma 21.9. *Given any two points $A, B \in D$, the hyperbolic circle[1] with centre A passing through B, is a Euclidean circle (whose Euclidean-centre is typically not A).*

Proof. Möbius maps send Euclidean circles to Euclidean circles (allowing straight lines), and hyperbolic isometries send hyperbolic circles to hyperbolic circles. So by applying a Möbius isometry we may assume $A = 0 \in D$. Since rotations about 0 are hyperbolic isometries, the

[1]that is the set of points $\{q \in D : \mathrm{dist}_D(A, q) = r\}$ where $r = \mathrm{dist}_D(A, B)$.

Euclidean circle with centre 0 passing through B coincides with the hyperbolic circle with centre 0 passing through B.
Remark. The Möbius isometry will typically not map Euclidean centres to Euclidean centres, indeed the hyperbolic centre will lie closer to ∂D than the Euclidean centre, because short Euclidean distances near ∂D are actually very long hyperbolic distances. □

21.7 The cosine rule and the sine rule

Non-examinable. *For the purposes of this course, the following is just a historical curiosity.*

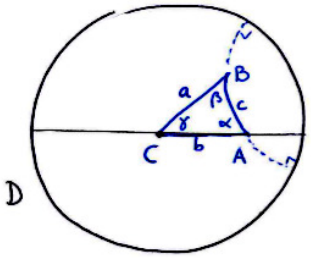
Recall that in Euclidean geometry, the cosine rule states: $c^2 = a^2 + b^2 - 2ab\cos\gamma$ (which is the generalization of Pythagoras's theorem, which is the case $\gamma = \pi/2$).

Lemma 21.10 (Cosine Rule)**.** *For a hyperbolic triangle in D,*

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

Remark. *For small x, $\cosh x = \frac{e^x+e^{-x}}{2} \sim \frac{1+x+\frac{1}{2}x^2+1-x+\frac{1}{2}x^2}{2} = 1 + \frac{1}{2}x^2$ and $\sinh x = \frac{e^x-e^{-x}}{2} \sim \frac{1+x-1+x}{2} = x$, so for small triangles the above cosine rule becomes $1 + \frac{c^2}{2} \sim (1 + \frac{a^2}{2})(1 + \frac{b^2}{2}) - ab\cos\gamma$, which is the Euclidean cosine rule when dropping order 4 terms.*

Proof. Using isometries, we may assume $C = 0$ and then rotating we may assume $A > 0 \in \mathbb{R}$.



Then $b = \mathrm{dist}(A, C) = 2\tanh^{-1}(A)$ by Theorem 21.7. So $A = \tanh\frac{b}{2}$. Rotating by $e^{-i\gamma}$ would make B real positive, so the same formula would apply, so $B = e^{i\gamma}\tanh\frac{a}{2}$. By Theorem 21.7,

$$\tanh\frac{c}{2} = \left|\frac{B-A}{1-\overline{A}B}\right|.$$

Therefore:[1]

$$\cosh c = \frac{1+\tanh^2\frac{c}{2}}{1-\tanh^2\frac{c}{2}} = \frac{|1-\overline{A}B|^2+|B-A|^2}{|1-\overline{A}B|^2-|B-A|^2} = \frac{(1+|A|^2)(1+|B|^2)-2(\overline{A}B+A\overline{B})}{(1-|A|^2)(1-|B|^2)}$$

Similarly, $\cosh b = \frac{1+\tanh^2\frac{b}{2}}{1-\tanh^2\frac{b}{2}} = \frac{1+|A|^2}{1-|A|^2}$ and $\cosh a = \frac{1+|B|^2}{1-|B|^2}$. The claim now follows from the final calculation:

$$\begin{aligned}\frac{2(\overline{A}B+A\overline{B})}{(1-|A|^2)(1-|B|^2)} &= \frac{2\tanh\frac{b}{2}\tanh\frac{a}{2}(e^{i\gamma}+e^{-i\gamma})}{\mathrm{sech}^2\frac{b}{2}\,\mathrm{sech}^2\frac{a}{2}} = 2\sinh\tfrac{a}{2}\cosh\tfrac{a}{2}\sinh\tfrac{b}{2}\cosh\tfrac{a}{2}2\cos\gamma \\ &= \sinh a\sinh b\cos\gamma.\end{aligned}$$

□

[1] Refresher: $\sinh(ix) = i\sin(x)$ and $\cosh(ix) = \cos(x)$ so any formula involving sines and cosines gives a corresponding formula for hyperbolic functions provided you replace sin by $i\sinh$. So you replace $\cos^2$, $\sin^2$, $\tan^2$ by $\cosh^2$, $-\sinh^2$, $-\tanh^2$. So the formula $\cos^2 x = \cos^2 x - \sin^2 x = \frac{\cos^2 x-\sin^2 x}{\cos^2 x+\sin^2 x} = \frac{1-\tan^2 x}{1+\tan^2 x}$ becomes

$$\cosh x = \frac{1+\tanh^2 x}{1-\tanh^2 x}.$$

Exercise. *Show that also another cosine rule holds for D:*

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

Exercise. *Show that in spherical geometry, so for the unit sphere with the chordal metric (Exercise Sheet 2), the cosine rules become*

$$\begin{aligned}\cos c &= \cos a\cos b + \sin a\sin b\cos\gamma\\ \cos\gamma &= -\cos\alpha\cos\beta + \sin\alpha\sin\beta\cos c.\end{aligned}$$

Lemma 21.11 (Sine Rule)**.** *For a hyperbolic triangle in D,*

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

Proof. By the Cosine rule,

$$\begin{aligned}\sinh^2 a\sinh^2 b\cos^2\gamma &= (\cosh c - \cosh a\cosh b)^2\\ &= \cosh^2 c + \cosh^2 a\cosh^2 b - 2\cosh a\cosh b\cosh c.\end{aligned}$$

Expanding $\cos^2\gamma = 1-\sin^2\gamma$ and $\sinh^2 = \cosh^2 - 1$, then rearranging terms:

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

Since the right hand side is symmetric in a, b, c, we deduce symmetries for the left hand side:

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

22. Appendix: Classification of Riemann surfaces

This Appendix is non-examinable.

22.1 Conformal structure of a Riemann surface

Recall that for a Riemann surface R the transition maps τ are holomorphic, so their derivatives $D\tau \equiv \tau'(z)$ are a composition of scaling and rotation, so $D\tau$ preserves angles. So Euclidean angles defined in local parametrizations are independent of the observer. This is called the **conformal structure** of R.

It therefore makes sense to restrict one's attention to only those Riemannian metrics on R for which the angles measured using the metric[1] agree with the above well-defined angles. Since in a local holomorphic coordinate

$$z = x + iy$$

the standard basis vectors $e_1 = \partial_x$ and $e_2 = \partial_y$ form a right angle, such metrics have no diagonal terms $dx\,dy$, so:

$$I = f(x,y)\,(dx^2 + dy^2) = f(z)\,|dz|^2,$$

where of course $f(x,y) > 0$ is a positive smooth function (positive definiteness of I).

Example. For $\mathbb{H}$, $f(x,y) = \frac{1}{y^2}$ gives the standard hyperbolic metric.

The above is just a local expression, so for such I to yield a well-defined Riemannian metric, you need the local functions f to be compatible with changes of coordinates.[2]

The equivalence class of all such metrics is called the **conformal class** of R. (Two metrics are equivalent if they only differ by a positive scaling function).

Recall from Section 5.3:

[1] Recall $\cos\theta = \frac{I(v,w)}{\sqrt{I(v,v)}\sqrt{I(w,w)}}$ for vectors $v, w \neq 0 \in T_pS$ defines an angle $\pm\theta$ between v, w.

[2] So $\widetilde{f}(\widetilde{z})\,|d\widetilde{z}|^2 = f(z)\,|dz|^2$ for a holomorphic transition $\widetilde{z} = \tau(z)$, so $\widetilde{f}(\tau(z))\,|\tau'(z)|^2 = f(z)$.

Theorem 22.1 (Riemann mapping theorem)**.**
Every simply-connected[1] *Riemann surface is biholomorphic to either* $\mathbb{C}P^1$, $\mathbb{C}$ *or* $\mathbb{H}$.

A special feature of complex analysis, is that a holomorphic homeomorphism is automatically a biholomorphism (unlike in real analysis, where $\mathbb{R} \to \mathbb{R}$, $x \mapsto x^3$ is a smooth homeomorphism, but its inverse $x \mapsto x^{1/3}$ is not smooth). Indeed, since it is a homeomorphism, the local form of the map is $z \mapsto z$, so the inverse function theorem guarantees a holomorphic inverse.

Theorem 22.2 (Uniformization theorem due to Poincaré and Koebe)**.**
Every compact Riemann surface R *has a metric of constant Gaussian curvature within its conformal class. In particular, by Gauss-Bonnet, it follows that if* $K > 0$ *then* $\chi(R) = 2$ *so* R *is topologically a sphere, if* $K = 0$ *then* $\chi(R) = 0$ *so* R *is topologically a torus, and if* $K < 0$ *then* $\chi(R) < 0$ *so* R *is topologically a surface of genus* $g \geq 2$.

We will sketch the proof of Theorem 22.2. For this, we need a topological preliminary.

22.2 Universal covering space

As shown in the course **B3.5: Topology and Groups**, for a reasonable[2] connected and path-connected topological space R (for example, for any connected topological surface or manifold) one can form a **universal covering space** $\widetilde{R}$. Explicitly, $\widetilde{R}$ can be constructed as the space of equivalence classes of continuous paths $\gamma : [0,1] \to R$ in R starting from a fixed base-point $\gamma(0) = p_0$. Two paths γ_1, γ_2 are defined to be equivalent if the endpoints agree, $\gamma_1(1) = \gamma_2(1)$, and if γ_1 can be continuously deformed into γ_2 whilst keeping fixed both the initial point p_0 and the end-point.

The properties required for $\widetilde{R}$ to be a **universal cover** are:

(1) $\widetilde{R}$ is a simply-connected topological space (in particular connected)
(2) there is a continuous surjective map $\pi : \widetilde{R} \to R$, called **projection map**, which is a local homeomorphism,
(3) locally π looks like a "stack of pancakes": namely, around any point $p \in R$ we can find a small open neighbourhood V such that $\pi^{-1}(V) = \sqcup U_j$ is a disjoint union of open sets U_j, called **sheets**, each homeomorphic to V via $\pi : U_j \to V$

In the explicit construction of $\widetilde{R}$ mentioned above, the map π is $\pi[\gamma] = \gamma(1)$: just map the path to its endpoint. One can define a group G, called **deck group**, consisting of homeomorphisms $\varphi : \widetilde{R} \to \widetilde{R}$ compatible with the projection: $\pi \circ \varphi = \pi$. So locally φ permutes the sheets. One can easily check that if φ has a fixed point ($\varphi(\widetilde{r}) = \widetilde{r}$) then φ is the identity map. It turns out that R can be identified with the quotient $R = \widetilde{R}/G$ that is, the points of R can be viewed as the orbits of the G-action on $\widetilde{R}$.

Example. The universal cover of the circle S^1 is the real line $\mathbb{R}$, with projection $\pi(r) = e^{2\pi i r}$. The deck group G is all integer translations $r \mapsto r + n$, $n \in \mathbb{Z}$, so G as a group is isomorphic to $\mathbb{Z}$ with addition. Notice that indeed $S^1 = \mathbb{R}/\mathbb{Z}$.

The universal cover satisfies the following universality property: given any two universal covers $\widetilde{R}_1$, $\widetilde{R}_2$ of R, there is a homeomorphism $\psi : \widetilde{R}_1 \to \widetilde{R}_2$ compatible with the projection maps: $\pi_2 \circ \psi = \pi_1$. It follows that the deck groups are isomorphic: $G_1 \cong G_2$.

Example. Another universal cover for S^1 is $\mathbb{R}$, with projection $\pi_2(r) = e^{ir}$. The deck group G_2

[1] Simply-connected means: connected, and every continuous loop can be continuously shrunk to a point (every continuous map $S^1 \to S$ can be extended to a continuous map $\mathbb{D} \to S$ on the closed unit disc).

[2] One needs a technical condition: semi-locally simply-connected. This means every point p has some neighbourhood V such that loops in V can be contracted within V to a point.

is all translations $r \mapsto r + 2\pi n$, $n \in \mathbb{Z}$, so again $G_2 \cong (\mathbb{Z}, +)$. The homeomorphic identification with the previous example is $\mathbb{R} \to \mathbb{R}$, $r \mapsto 2\pi r$.

22.3 Sketch proof of the uniformization theorem

Sketch proof of Theorem 22.2. Given a compact Riemann surface R, consider its universal cover $\widetilde{R}$. As $\widetilde{R}$ is locally homeomorphic to R, we can use the same local holomorphic coordinate on $\widetilde{R}$ as on R via this local identification. This makes $\widetilde{R}$ into a simply-connected Riemann surface, and the projection map $\pi : \widetilde{R} \to R$ is automatically holomorphic. Since π is a local homeomorphism (the local model is $z \mapsto z$), π has no branch points. The deck group automatically consists of holomorphic homeomorphisms $\widetilde{R} \to \widetilde{R}$, so they are biholomorphisms.

By the Riemann-mapping theorem 22.1, $\widetilde{R}$ is biholomorphic to $\mathbb{C}P^1$, $\mathbb{C}$ or $\mathbb{H}$.

If $\widetilde{R} \cong \mathbb{C}P^1$, then $\widetilde{R}, R$ are both compact so there are only finitely many sheets namely $\deg \pi$ sheets. Since π has no branch points, by Riemann-Hurwitz

$$\chi(\widetilde{R}) = \deg(\pi)\chi(R).$$

But $\chi(\widetilde{R}) = \chi(\mathbb{C}P^1) = 2$, and $\chi(R) = 2 - 2g \in \{2, 0, -2, -4, \ldots\}$ (since Riemann surfaces are orientable), which forces $\chi(R) = 2$ and so $\deg \pi = 1$, so π is a biholomorphism. So R is biholomorphic to $\mathbb{C}P^1$. So we can give R the usual metric on $\mathbb{C}P^1 \cong S^2 \subset \mathbb{R}^3$ with $K = 1$.

If $\widetilde{R} \cong \mathbb{C}$, then the deck group G consists of a subgroup of the biholomorphisms $f : \mathbb{C} \to \mathbb{C}$. But such biholomorphisms have[1] the form $f(z) = az + b$ for $a \neq 0, b \in \mathbb{C}$. Since the non-identity deck group transformations have no fixed points, $az + b = z$ must not have any solution, which forces $a - 1 = 0$ and $b \neq 0$. So $f(z) = z + b$ are the translations. These preserve the flat metric $dx^2 + dy^2 = |dz|^2$ on $\widetilde{R}$, so the quotient $R = \widetilde{R}/G$ can be given the flat metric, so $K = 0$.

If $\widetilde{R} \cong \mathbb{H}$, we claim that the deck group preserves the hyperbolic metric, so we can give $R = \widetilde{R}/G$ the hyperbolic metric, so $K = -1$. This claim follows by Lemma 22.4. □

Lemma 22.3 (Schwartz's Lemma)**.** *Any holomorphic map $f : D \to D$ with $f(0) = 0$ satisfies*

$$|f(z)| \leq |z| \text{ for all } z \in D.$$

Proof. $f(0) = 0$, so $g(z) = f(z)/z$ is holomorphic with a removable singularity at 0, which is removed by defining $g(0) = f'(0)$. The maximum modulus principle applied to the disc of radius r implies that $|g(z)| \leq \frac{1}{r}$ for $|z| \leq r$ (using that $|f| \leq 1$, and that $|z| = r$ on the boundary of that disc). Let $r \to 1$ from below, then $|g(z)| \leq 1$ for all $z \in D$, so $|f(z)| \leq |z|$. □

Lemma 22.4. *Any biholomorphism $f : D \to D$ of the hyperbolic disc preserves the hyperbolic metric.*

Proof. Composing with a Möbius isometry, we may assume $f(0) = 0$. By Schwartz's Lemma, $|f(z)| \leq |z|$. Since f is invertible, also $|f^{-1}(z)| \leq |z|$ so (replacing z by $f(z)$) we get $|z| \leq |f(z)|$. Hence equality holds: $|f(z)| = |z|$. Recall by the previous proof that $g(z) = f(z)/z$ is holomorphic. Since $|g(z)| = 1$, the maximum modulus is attained at an interior point, therefore the maximum modulus principle implies that g is constant, say $g(z) = e^{i\theta}$. Thus $f(z) = e^{i\theta}z$. We know this is an isometry of D, so the claim follows. □

[1] Indeed, the holomorphic function $g(z) = f(1/z)$ defined on the punctured unit disc $D\backslash\{0\}$ has a singularity at 0 which is either removable, a pole or an essential singularity. If it was an essential singularity then by the Casorati-Weierstrass theorem[2], g on D must attain values arbitrarily close to $f(0) \in \mathbb{C}$. This contradicts that f is injective, since $g(w) = f(1/w)$ for $|1/w| > 1$ would attain a value that $f(z)$ attains for $|z| < 1$ (close to $f(0)$). So $g(w)$ has a pole at 0, say with principal part $b_n w^{-n} + \cdots + b_1 w^{-1}$. Then $f - b_n z^n - \cdots - b_1 z : \mathbb{C}P^1 \to \mathbb{C}$ is a holomorphic function (we got rid of the pole at infinity) so it is constant. So f is a polynomial. By injectivity of f, we deduce $n = 1$, so $f(z) = az + b$ as required.