Part A Geometry*

Prof Jason D. Lotay
jason.lotay@maths.ox.ac.uk

Hilary Term 2025–2026

Course overview

Geometry is a fundamental topic in mathematics with links to algebra, analysis, number theory and topology, as well as applied mathematics and theoretical physics, including classical mechanics, mathematical biology and General Relativity. This course will introduce foundational geometric concepts, such as submanifolds and manifolds, transversality and degree, which provide the essential tools for further study in many aspects of algebraic and differential geometry and topology. The course will discuss key examples of relevance both within and beyond geometry and give various applications of the theory. Finally, the course will look at important examples of geometric structures and their symmetries, namely projective and hyperbolic space. Aspects of these geometries are of central importance in many areas of mathematics, and we will discuss several of these links, including to Möbius transformations from complex analysis.

Course synopsis.

Review of derivative for functions between Euclidean spaces. Inverse and Implicit Function Theorems. Manifolds arising as submanifolds of Euclidean space and from matrix groups. Lagrange multipliers. Sard's Theorem. Transversality and intersections of submanifolds. Degree of a map with applications, including Fundamental Theorem of Algebra.
Abstract definition of manifold. Projective space and projective transformations, including intersections and Möbius transformations. Hyperbolic space and transformations, with link to Möbius transformations.

Prerequisities. There are no prerequisites for this course, other than the core courses A2.1 Metric Spaces and A2.2 Complex Analysis. Students would benefit from taking A5 Topology in conjunction with this course, but it is certainly not essential.

Recommended texts

- M. Spivak, Calculus on Manifolds, Westview Press, 1998.
- J. W. Milnor, Topology from the Differentiable Viewpoint, Princeton University Press, 1997.
- M. Reid and B. Szendröi, Geometry and Topology, Cambridge University Press, 2005
- E. G. Rees, Notes on Geometry, Springer, 2004.

*This version: February 24, 2026.

Contents

1 Introduction

Geometry is the study of shapes. A shape can be presented to us in many ways: as a physical object, in terms of the properties it possesses, or (as is often the case) as the solution to an equation. We can study shapes in many different ways and this leads us to various approaches to geometry which we will explore throughout this course.

Algebra

A simple example of geometry is the set of solutions to

$$x_1^2 + x_2^2 + x_3^2 = 1$$

for $x_1, x_2, x_3 \in \mathbb{R}$, which gives a round 2-dimensional sphere. For polynomial equations such as this we can use *algebra* to study the geometry, and this is the starting point for *algebraic geometry*. As we know, algebra is easiest over the complex numbers, so we might consider the same equation above but now for $x_1, x_2, x_3 \in \mathbb{C}$: this already gives quite a different geometric object.

Analysis

Many equations are not polynomial and then it is more appropriate to use *analysis* to study the geometry. This will allow us to discuss whether the geometry is *smooth* (or not).

Even though we gave the round 2-dimensional sphere in terms of an equation before, we can present it in many different ways using different equations, by specifying its symmetry properties, as the quotient of two groups and so on. This means it is advantageous to have an *abstract* notion of what a 2-dimensional sphere is and thus of any other smooth geometric object. This is the concept of *manifolds* and they form the foundation for *differential geometry*.

Intersections

An important but natural idea is to think about the solution to more than one equation: this corresponds to thinking about the *intersection* of two geometric objects. For example:

$$x_1^2 + x_2^2 + x_3^2 = 1 \quad \text{and} \quad x_3 = c$$

for $c \in \mathbb{R}$ fixed and $x_1, x_2, x_3 \in \mathbb{R}$. We see that we get a circle if $|c| < 1$, nothing if $|c| > 1$ and just a point if $|c| = 1$. In the last case, the two geometric objects just touch, i.e. they do not intersect *transversely*. This idea of transversality is one we will examine and see that we can ensure it almost all of the time, just like in this example.

Projective space

If we return to our discussion of algebra we can consider the equation

$$x_1^2 + x_2^2 + x_3^2 = 0$$

for $x_1, x_2, x_3 \in \mathbb{C}$. We see that this equation has the attractive feature that it is invariant under multiplication by a non-zero complex number, i.e. (x_1, x_2, x_3) is a solution if and only if $(\lambda x_1, \lambda x_2, \lambda x_3)$ is a solution for all $\lambda \in \mathbb{C}^* = \mathbb{C} \setminus \{0\}$. We can then consider the solutions of the equation up to multiplication by $\mathbb{C}^*$: this leads to the idea of *projective space*. We can also do this for other fields than $\mathbb{C}$ and it forms a rich geometry. In particular, complex projective space is one of the key players in *algebraic geometry* and has fascinating transformations.

Hyperbolic space

As you may know, in Euclidean geometry one has the *parallel postulate*: for any straight line ℓ in the plane and every point p not on the line there is a unique parallel straight line to ℓ passing through p. It was a revolution in geometry in the 19th Century to discover that there is another geometry defined on the (upper half) plane for which the parallel postulate *fails*. This is *hyperbolic geometry* and appears in many different settings, from group theory to theoretical physics. We will conclude the course by studying the fundamentals of hyperbolic space and see some of its surprising and beautiful properties.

2 A little analysis

We may have an intuition about what it means for an object living in 3 dimensions to be *smooth*. For example, it seems clear that the unit sphere in $\mathbb{R}^3$ is smooth but the unit cube in $\mathbb{R}^3$ is not. How do we formalize this notion?

To do this, we first need to remind ourselves what it means for a function to be differentiable between Euclidean spaces. Throughout we shall use $\|\cdot\|$ to denote the Euclidean norm (though, actually, any norm can be used).

Definition 2.1. A function $f\colon \mathbb{R}^n \to \mathbb{R}^m$ is *differentiable* at $a \in \mathbb{R}^n$ if there exists a (unique) linear map $\mathrm{d}f_a\colon \mathbb{R}^n \to \mathbb{R}^m$ such that

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

We call $\mathrm{d}f_a$ the *(total) derivative* of f at a. (Some authors also call $\mathrm{d}f_a$ the *differential* of f at a.) We say that f is *differentiable* on an open set $U \subseteq \mathbb{R}^n$ if it is differentiable at all $a \in U$. (We may also define f just on an open subset of $\mathbb{R}^n$ rather than all of $\mathbb{R}^n$.)

Remark. It is natural to ask: why is the derivative $\mathrm{d}f_a$ of f at a a linear map? For a function $f\colon \mathbb{R} \to \mathbb{R}$ the usual derivative $f'(a)$ at a is the gradient of the line tangent to the graph of f in $\mathbb{R}^2$ at a: it defines the "best linear approximation" to f at a. Defining the gradient of the line is the same as choosing a linear map from $x \in \mathbb{R}$ to $y \in \mathbb{R}$: it tells us how far we need to stretch in the y direction if we move in the x direction, and this must be linear because we are defining a line. For a function $f\colon \mathbb{R}^n \to \mathbb{R}^m$ therefore, we need to say how much we stretch in the $(y_1, \dots, y_m) \in \mathbb{R}^m$ directions if we move in the $(x_1, \dots, x_n) \in \mathbb{R}^n$ directions to get the "gradient" of the graph of f. Moreover, it must be linear as we want the "best linear approximation" to the graph, which will be an n-dimensional subspace in $\mathbb{R}^{n+m}$. This hopefully gives some motivation why $\mathrm{d}f_a$ is a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$.

Example. If $f : \mathbb{R} \to \mathbb{R}$ is differentiable at $a \in \mathbb{R}$ then, for $h \in \mathbb{R}$,

$$\mathrm{d}f_a(h) = f'(a)h.$$

Remark. Definition 2.1 above works perfectly well for a map between normed vector spaces. We can also see that it is equivalent to asking for the existence of a linear map $\mathrm{d}f_a$ such that

$$R_a(h) := f(a+h) - f(a) - \mathrm{d}f_a(h) = o(\|h\|) \quad \text{as } h \to 0,$$

i.e. $\|R_a(h)\|/\|h\| \to 0$ as $h \to 0$.

Example. If we take the identity map $\mathrm{id} : \mathbb{R}^n \to \mathbb{R}^n$ then $\mathrm{id}(a+h) - \mathrm{id}(a) = h$, so clearly id is differentiable with derivative

$$\mathrm{d}\,\mathrm{id}_a = \mathrm{id}$$

for all $a \in \mathbb{R}^n$. More generally, if $f\colon \mathbb{R}^n \to \mathbb{R}^m$ is a linear map then $\mathrm{d}f_a = f$ for all $a \in \mathbb{R}^n$.

Example. If $f\colon \mathbb{R}^n \to \mathbb{R}^m$ then, using the standard basis of $\mathbb{R}^m$, we can write f as a vector of functions

$$\begin{pmatrix} f_1 \\ \vdots \\ f_m \end{pmatrix}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$ for $i = 1, \ldots, m$. Then, using the standard bases of $\mathbb{R}^n$ and $\mathbb{R}^m$ and coordinates $(x_1, \ldots, x_n)$ on $\mathbb{R}^n$, the linear map $\mathrm{d}f_a$ for $a \in \mathbb{R}^n$ has matrix

$$\left(\frac{\partial f_i}{\partial x_j}(a)\right) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(a) & \ldots & \frac{\partial f_1}{\partial x_n}(a) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(a) & \ldots & \frac{\partial f_m}{\partial x_n}(a) \end{pmatrix}.$$

This is called the *Jacobian matrix* of the partial derivatives of f at a. (Note that this formula shows that $\mathrm{d}f_a$ is unique.)

Example. Let $f \colon \mathbb{C} \to \mathbb{C}$ be holomorphic. If we identify $\mathbb{C}$ with $\mathbb{R}^2$ with coordinates (x, y) we can write $f(x + iy) = u(x, y) + iv(x, y)$. Then $\mathrm{d}f_a$ has matrix

$$\begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{pmatrix} = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ -\frac{\partial u}{\partial y} & \frac{\partial u}{\partial x} \end{pmatrix},$$

by the Cauchy–Riemann equations, which has determinant

$$\left(\tfrac{\partial u}{\partial x}\right)^2 + \left(\tfrac{\partial u}{\partial y}\right)^2 \geq 0,$$

with equality if and only if f is constant. (The fact that the Jacobian determinant is non-negative is the statement that holomorphic maps "preserve orientation": we shall return to this point later.)

Although the Jacobian matrix approach might look attractive, it is not always the best option, as the next example shows.

Example. Let $M_n(\mathbb{R})$ denote the $n \times n$ real matrices, which we may identify with $\mathbb{R}^{n^2}$. Consider the determinant map $\det : M_n(\mathbb{R}) \to \mathbb{R}$. We see that if I is the identity matrix then for any $H = (h_{ij}) \in M_n(\mathbb{R})$ we have

$$\det(I + H) = \det \begin{pmatrix} 1 + h_{11} & h_{12} & \ldots & h_{1n} \\ h_{21} & 1 + h_{22} & \ldots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \ldots & 1 + h_{nn} \end{pmatrix} = 1 + h_{11} + h_{22} + \ldots + h_{nn} + Q(H)$$

where $Q(H)$ only contains terms which are quadratic or higher in the entries of H. We recognise that the linear term in the entries of H is just $\mathrm{tr}(H)$, the trace of H, and so

$$\frac{\|\det(I + H) - \det(I) - \mathrm{tr}(H)\|}{\|H\|} = \frac{\|Q(H)\|}{\|H\|} \to 0$$

as $H \to 0$. Hence, det is differentiable at I and

$$\mathrm{d}\det{}_I(H) = \mathrm{tr}(H).$$

(We can also formalize this calculation by recognizing $\det(I + H)$ as the characteristic polynomial of H evaluated at -1.)

Using the properties of the determinant, one can easily see that det is differentiable at any invertible matrix A, i.e. at any element in the *general linear group*

$$\mathrm{GL}(n, \mathbb{R}) = \{A \in M_n(\mathbb{R}) \,:\, \det A \neq 0\}$$

and compute the differential $\mathrm{d}\det{}_A$.

A very useful tool when computing derivatives is the *Chain Rule*. We are very familiar with this from functions of one variable and there is a natural generalization as follows.

Proposition 2.2 (Chain Rule)**.** *Let $f\colon \mathbb{R}^n \to \mathbb{R}^l$ and let $g\colon \mathbb{R}^l \to \mathbb{R}^m$ be such that f is differentiable at $a \in \mathbb{R}^n$ and g is differentiable at $f(a) \in \mathbb{R}^l$. Then $g \circ f\colon \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at a and*

$$\mathrm{d}(g \circ f)_a = \mathrm{d}g_{f(a)} \circ \mathrm{d}f_a.$$

Proof. **(Not examinable.)** To help with notation we let $f(a) = b$. We also note that for a linear map L between Euclidean spaces, there exists a least constant $\|L\| \geq 0$ such that

$$\|Lx\| \leq \|L\|\|x\|$$

for all x. This constant is called the *operator norm* of L.

We know that

$$f(a+h) = f(a) + \mathrm{d}f_a(h) + Rf_a(h)$$

where $Rf_a(h) = o(\|h\|)$ as $h \to 0$. We also know that

$$g(b+k) = g(b) + \mathrm{d}g_b(k) + Rg_b(k)$$

where $Rg_b(k) = o(\|k\|)$ as $k \to 0$. Since $\mathrm{d}f_a(h)+Rf_a(h) \to 0$ as $h \to 0$, we may choose $k = \mathrm{d}f_a(h)+Rf_a(h)$ in the formula above to see that

$$\begin{aligned} g(f(a+h)) &= g(b + \mathrm{d}f_a(h) + Rf_a(h)) \\ &= g(b) + \mathrm{d}g_b(\mathrm{d}f_a(h) + Rf_a(h)) + Rg_b(\mathrm{d}f_a(h) + Rf_a(h)). \end{aligned}$$

Re-arranging, we have that (using that $\mathrm{d}g_b$ is linear)

$$g(f(a+h)) - g(f(a)) - \mathrm{d}g_b \circ \mathrm{d}f_a(h) = \mathrm{d}g_b(Rf_a(h)) + Rg_b(\mathrm{d}f_a(h) + Rf_a(h)).$$

We see that

$$\frac{\|\mathrm{d}g_b(Rf_a(h))\|}{\|h\|} \leq \|\mathrm{d}g_b\| \frac{\|Rf_a(h)\|}{\|h\|} \to 0$$

as $h \to 0$. We also have that

$$\frac{\|Rg_b(\mathrm{d}f_a(h) + Rf_a(h))\|}{\|h\|} = \frac{\|Rg_b(\mathrm{d}f_a(h) + Rf_a(h))\|}{\|\mathrm{d}f_a(h) + Rf_a(h)\|} \cdot \frac{\|\mathrm{d}f_a(h) + Rf_a(h)\|}{\|h\|}.$$

The first term on the right-hand side tends to zero as $h \to 0$ since $\mathrm{d}f_a(h) + Rf_a(h) \to 0$, and the second term is bounded as $h \to 0$ because $Rf_a(h) = o(\|h\|)$ and $\|\mathrm{d}f_a(h)\| \leq \|\mathrm{d}f_a\|\|h\|$. This completes the proof. □

Example. If we define $f\colon M_n(\mathbb{R}) \to M_n(\mathbb{R})$ by

$$f(A) = A^{\mathrm{T}}A$$

we see that for $H \in M_n(\mathbb{R})$:

$$f(A+H) - f(A) = (A+H)^{\mathrm{T}}(A+H) - A^{\mathrm{T}}A = A^{\mathrm{T}}H + H^{\mathrm{T}}A + H^{\mathrm{T}}H.$$

Therefore if we let

$$\mathrm{d}f_A(H) = A^{\mathrm{T}}H + H^{\mathrm{T}}A,$$

which is a linear map on $H \in M_n(\mathbb{R})$, then

$$f(A+H) - f(A) - \mathrm{d}f_A(H) = o(\|H\|) \quad \text{as } H \to 0.$$

Hence, f is differentiable at A and $\mathrm{d}f_A$ is the derivative of f at A. We also see that $f(I) = I$.

If we define $g : M_n(\mathbb{R}) \to \mathbb{R}$ by

$$g(A) = \det(A)$$

then we saw that g is differentiable at I with $\mathrm{d}g_I(H) = \operatorname{tr} H$. We deduce from the Chain Rule that $g \circ f(A) = \det(A^{\mathrm{T}}A)$ is differentiable at I with

$$\mathrm{d}(g \circ f)_I(H) = \mathrm{d}g_{f(I)} \circ \mathrm{d}f_I(H) = \mathrm{d}g_I(H + H^{\mathrm{T}}) = \operatorname{tr}(H + H^{\mathrm{T}}) = \operatorname{tr}(H) + \operatorname{tr}(H^{\mathrm{T}}) = 2 \operatorname{tr} H,$$

for any $H \in M_n(\mathbb{R})$. We can also see this from

$$g \circ f(A) = \det(A^{\mathrm{T}}A) = \det(A^{\mathrm{T}}) \det(A) = \det(A)^2$$

and the Chain Rule using g and the function $x \mapsto x^2$ on $\mathbb{R}$.

In geometry we often want our functions to be more than just differentiable, so we have the following notions.

Definition 2.3. A function $f \colon \mathbb{R}^n \to \mathbb{R}^m$ is C^k if all partial derivatives of f exist up to order k and are all continuous. (For example, C^1 just means all first order partial derivatives exist and are continuous.) A function $f \colon \mathbb{R}^n \to \mathbb{R}^m$ is *smooth* (or C^∞) if it is C^k for all $k \in \mathbb{N}$. We will write $C^k(\mathbb{R}^n, \mathbb{R}^m)$ and $C^\infty(\mathbb{R}^n, \mathbb{R}^m)$ for the sets of C^k and smooth functions from $\mathbb{R}^n$ to $\mathbb{R}^m$, respectively. (Again, we can replace $\mathbb{R}^n$ by an open set in $\mathbb{R}^n$.)

Remark. If we want to differentiate functions $f \colon \mathbb{C}^n \to \mathbb{C}^m$ we can do this by identifying $\mathbb{C}^n$ with $\mathbb{R}^{2n}$. We should take care because if $f \colon \mathbb{C} \to \mathbb{C}$ is differentiable then it is *not* necessarily the case that $\mathrm{d}f_a \colon \mathbb{C} \to \mathbb{C}$ is given by $\mathrm{d}f_a(h) = f'(a)h$. As we will see (on a problem sheet) this is only guaranteed if f is *holomorphic*, which is perhaps not surprising. In fact, it can lead to an alternative definition of holomorphic functions.

We know well that a smooth bijection need not have a smooth inverse. For example, $f(x) = x^3$ is a smooth bijection but its inverse is not even differentiable at 0. However, there are smooth bijections with a smooth inverse, for example the identity $f(x) = x$, and these functions will be very important.

Definition 2.4. A function $f \colon U \to V$ between open sets $U, V \subseteq \mathbb{R}^n$ is a *diffeomorphism* if it is a smooth bijection with smooth inverse.

Example. The function $\tan : (-\pi/2, \pi/2) \to \mathbb{R}$ is a diffeomorphism.

Example. Any $A \in \mathrm{GL}(n, \mathbb{R})$ defines a diffeomorphism of $\mathbb{R}^n$ since it defines an invertible linear map, and linear maps are smooth.

A nice fact that we will use is the following which relates diffeomorpishisms to the differential.

Lemma 2.5. *Let $f \colon U \to V$ be a diffeomorphism between open sets U, V in $\mathbb{R}^n$ and let $a \in U$. Then $\mathrm{d}f_a \colon \mathbb{R}^n \to \mathbb{R}^n$ is an isomorphism.*

Proof. By the Chain Rule we see that

$$\mathrm{id} = \mathrm{d}\,\mathrm{id}_a = \mathrm{d}(f^{-1} \circ f)_a = \mathrm{d}f^{-1}_{f(a)} \circ \mathrm{d}f_a.$$

Similarly,

$$\mathrm{d}f_a \circ \mathrm{d}f^{-1}_{f(a)} = \mathrm{id}\,.$$

The result follows. □

Remark. Of course, we only need f and f^{-1} to be differentiable at a and $f(a)$ respectively to obtain that $\mathrm{d}f_a$ is an isomorphism.

This relation between invertibility and the differential should remind us of the Inverse Function Theorem, which you saw in the Metric Spaces course (for the case of maps from $\mathbb{R}^2$ to itself, but the proof is the same in all dimensions – the proof is **non-examinable**.)

Theorem 2.6 (Inverse Function Theorem)**.** *Let $U \subseteq \mathbb{R}^n$ be open, let $a \in U$ and let $f \in C^1(U, \mathbb{R}^n)$. If $\mathrm{d}f_a \colon \mathbb{R}^n \to \mathbb{R}^n$ is an isomorphism then there exists an open set $V \ni a$ and an open set $W \ni f(a)$ such that $f \colon V \to W$ is a bijection with inverse $f^{-1} \in C^1(W, V)$. Moreover, for all $x \in V$,*

$$\mathrm{d}f^{-1}_{f(x)} = (\mathrm{d}f_x)^{-1}.$$

Putting the Inverse Function Theorem together with Lemma 2.5 above we arrive at the following.

Proposition 2.7. *Let $W \subseteq \mathbb{R}^n$ be open, let $a \in W$ and let $f \colon W \to \mathbb{R}^n$ be smooth. There exists an open set $U \ni a$ and an open set $V \ni f(a)$ such that $f \colon U \to V$ is a diffeomorphism (i.e. f is a local diffeomorphism at a) if and only if $\mathrm{d}f_a$ is an isomorphism.*

Example. If we take $f \colon \mathbb{C} \to \mathbb{C}$ given by

$$f(z) = e^z$$

then, because f is holomorphic,

$$\mathrm{d}f_z(w) = e^z w$$

for $w \in \mathbb{C}$, so f is a local diffeomorphism at every point $z \in \mathbb{C}$ as $e^z \neq 0$. However, f is obviously not a diffemorphism since it is not injective: $f(z + 2n\pi i) = f(z)$ for all $n \in \mathbb{Z}$. It is also not surjective, since $f(z) \neq 0$ for all $z \in \mathbb{C}$.

Diffeomorphisms will provide the natural notion of equivalence in geometry. Much of geometry involves relating local and global properties, and so local diffeomorphisms will turn out to be very useful.

3 Submanifolds

We can now give the definition of smooth object we wanted inside Euclidean space.

Definition 3.1. A subset M of $\mathbb{R}^n$ is a *k-dimensional (embedded) submanifold* if for $p \in M$ there exists an open set $U \ni p$ in $\mathbb{R}^n$ and a smooth function $f\colon U \to \mathbb{R}^{n-k}$ such that $M \cap U = f^{-1}(0)$ and 0 is a *regular value* of f, i.e. $\mathrm{d}f_a\colon \mathbb{R}^n \to \mathbb{R}^{n-k}$ is *surjective* for all $a \in f^{-1}(0)$.

Before we give the motivation for this definition, let us just check some examples. Our first is actually the local model for all submanifolds.

Example. Let $F\colon \mathbb{R}^k \to \mathbb{R}^{n-k}$ be a smooth function. Let

$$M = \mathrm{Graph}(F) = \{(x, F(x)) \,:\, x \in \mathbb{R}^k\} \subseteq \mathbb{R}^k \times \mathbb{R}^{n-k} = \mathbb{R}^n$$

be the graph of F. Then we can let $f\colon \mathbb{R}^n = \mathbb{R}^k \times \mathbb{R}^{n-k} \to \mathbb{R}^{n-k}$ be given by

$$f(x, y) = F(x) - y.$$

Then f is smooth because F is smooth, $M = f^{-1}(0)$ and, for all $(a, b) \in \mathbb{R}^k \times \mathbb{R}^{n-k}$,

$$\mathrm{d}f_{(a,b)} = (\mathrm{d}F_a \ -I)$$

where I is the identity matrix of size $n-k$. We see that $\mathrm{d}f_{(a,b)}$ has rank $n-k$ because of the presence of $-I$, so it is surjective, and thus 0 is a regular value of f. Hence M is a k-dimensional submanifold of $\mathbb{R}^n$.

A very special case is to take $F = 0$ which shows that the plane

$$\{(x_1, \ldots, x_n) \in \mathbb{R}^n \,:\, x_{k+1} = \ldots = x_n = 0\}$$

is a k-dimensional submanifold. In this case f is (minus) the projection map to $\mathbb{R}^{n-k}$.

We now show that the sphere is indeed a submanifold.

Example. Let $\mathcal{S}^n$ be the unit sphere in $\mathbb{R}^{n+1}$:

$$\mathcal{S}^n = \{x = (x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \,:\, \|x\|^2 = \sum_{i=1}^{n+1} x_i^2 = 1\}.$$

If we define $f\colon \mathbb{R}^{n+1} \to \mathbb{R}$ by

$$f(x_1, \ldots, x_{n+1}) = \sum_{i=1}^{n+1} x_i^2 - 1$$

then f is smooth because it is polynomial and $\mathcal{S}^n = f^{-1}(0)$. We then see that

$$\mathrm{d}f_a = 2a^{\mathrm{T}},$$

i.e. $\mathrm{d}f_a(h)$ is 2 times the dot product of a with h. We see that $\mathrm{d}f_a\colon \mathbb{R}^{n+1} \to \mathbb{R}$ is surjective except when $a = 0$: for $c \in \mathbb{R}$ and $a \neq 0$ take

$$h = c\frac{a}{2\|a\|^2} \quad \Rightarrow \quad \mathrm{d}f_a(h) = 2a^{\mathrm{T}}h = 2c\frac{a^{\mathrm{T}}a}{\|a\|^2} = c.$$

However, $0 \notin \mathcal{S}^n$, so 0 is a regular value of f and $\mathcal{S}^n$ is an n-dimensional submanifold of $\mathbb{R}^{n+1}$.

To be slightly more sophisticated, we show that a certain matrix group is a submanifold.

Example. Let

$$\mathrm{SL}(n, \mathbb{R}) = \{A \in M_n(\mathbb{R}) \,:\, \det A = 1\}$$

be the *special linear group*. If we let $f\colon M_n(\mathbb{R}) \to \mathbb{R}$ be

$$f(A) = \det A - 1$$

then f is smooth, as it is polynomial in the entries of A, and $f^{-1}(0) = \mathrm{SL}(n,\mathbb{R})$. We can compute that, at any A with $\det A = 1$, we have

$$\mathrm{d}f_A(H) = \mathrm{tr}(A^{-1}H),$$

noting that A must be invertible. We want to show that $\mathrm{d}f_A\colon M_n(\mathbb{R}) \to \mathbb{R}$ is surjective. If we are given $c \in \mathbb{R}$ then we can take

$$H = \frac{c}{n}A \quad \Rightarrow \quad \mathrm{d}f_A(\tfrac{c}{n}A) = \tfrac{c}{n}\,\mathrm{tr}(I) = c$$

since $\mathrm{tr}\, I = n$. Hence, 0 is a regular value and $\mathrm{SL}(n,\mathbb{R})$ is an n^2-1-dimensional submanifold of $M_n(\mathbb{R})$.

We will see that many matrix groups will be submanifolds. Let us do one more example.

Example. Let

$$\mathrm{O}(n) = \{A \in M_n(\mathbb{R}) \,:\, A^{\mathrm{T}}A = I\}$$

be the *orthogonal group*. If we let $f\colon M_n(\mathbb{R}) \to M_n(\mathbb{R})$ be given by

$$f(A) = A^{\mathrm{T}}A - I$$

then we already saw that f is differentiable and

$$\mathrm{d}f_A(H) = A^{\mathrm{T}}H + H^{\mathrm{T}}A.$$

We see that $f^{-1}(0) = \mathrm{O}(n)$ so our goal should be show that 0 is a regular value.

However, we quickly realise that there is a problem: if this were the case, then $\mathrm{O}(n)$ would be 0-dimensional, which it obviously isn't. We see that $\mathrm{d}f_A$ is in fact *not* surjective onto $M_n(\mathbb{R})$. How do we fix this?

The idea is to realise that f actually maps into a *subspace* of $M_n(\mathbb{R})$, i.e. $\mathrm{Sym}_n(\mathbb{R})$, the *symmetric* matrices, since $f(A)^{\mathrm{T}} = f(A)$. Now our goal is to show that for all $C \in \mathrm{Sym}_n(\mathbb{R})$ and $A \in \mathrm{O}(n)$ there is some $H \in M_n(\mathbb{R})$ such that

$$\mathrm{d}f_A(H) = C.$$

This we can achieve by choosing $H = \frac{1}{2}AC$:

$$\begin{aligned}\mathrm{d}f_A(\tfrac{1}{2}AC) &= \tfrac{1}{2}(AC)^{\mathrm{T}}A + \tfrac{1}{2}A^{\mathrm{T}}AC\\ &= \tfrac{1}{2}C^{\mathrm{T}}A^{\mathrm{T}}A + \tfrac{1}{2}C\\ &= \tfrac{1}{2}C^{\mathrm{T}} + \tfrac{1}{2}C = C,\end{aligned}$$

using the facts that $A^{\mathrm{T}}A = I$ and $C^{\mathrm{T}} = C$. Thus, 0 is a regular value of $f\colon M_n(\mathbb{R}) \to \mathrm{Sym}_n(\mathbb{R})$ and, since $\dim \mathrm{Sym}_n(\mathbb{R}) = \frac{1}{2}n(n+1)$ we have $\mathrm{Sym}_n(\mathbb{R}) = \mathbb{R}^{\frac{1}{2}n(n+1)}$ and deduce that $\mathrm{O}(n)$ is a submanifold of $M_n(\mathbb{R}) = \mathbb{R}^{n^2}$ with

$$\dim \mathrm{O}(n) = \dim M_n(\mathbb{R}) - \dim \mathrm{Sym}_n(\mathbb{R}) = n^2 - \tfrac{1}{2}n(n+1) = \tfrac{1}{2}n(n-1).$$

Example. For the *special orthogonal group*

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

we notice if $A \in \mathrm{O}(n)$ then $\det A \in \{-1,+1\}$, so

$$\mathrm{SO}(n) = \mathrm{O}(n) \cap \{A \in M_n(\mathbb{R}) \,:\, \det A > 0\}.$$

Since the set of matrices with positive determinant is open, we deduce that $\mathrm{SO}(n)$ is a submanifold of $M_n(\mathbb{R})$ of the same dimension as $\mathrm{O}(n)$, i.e. $\frac{1}{2}n(n-1)$.

Another way to see that $\mathrm{SO}(n)$ is a submanifold of the same dimension as $\mathrm{O}(n)$ is to use a bit of (metric space) topology: we realise that $\det : \mathrm{O}(n) \to \{\pm 1\}$ is continuous and surjective, so $\mathrm{SO}(n)$ is just

the connected component of $\mathrm{O}(n)$ containing I (i.e. it is a closed and open subset of $\mathrm{O}(n)$ containing I), and hence is a submanifold of the same dimension $\frac{1}{2}n(n-1)$.

We now have a classic warning example.

Example. Define $f\colon \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x,y) = x^3 - y^3.$$

Then we compute that

$$\mathrm{d}f_{(a,b)} = (3a^2, -3b^2)$$

which is zero at $(a,b) = (0,0)$ and hence 0 is *not* a regular value of f. However, we see that

$$f^{-1}(0) = \{(x,y) \in \mathbb{R}^2 \,:\, x^3 = y^3\} = \{(x,y) \in \mathbb{R}^2 \,:\, x = y\}$$

is definitely an embedded 1-dimensional submanifold of $\mathbb{R}^2$ (for example because it is the graph of the identity map on $\mathbb{R}$).

Hence, the regular value condition on $f\colon \mathbb{R}^n \to \mathbb{R}^{n-k}$ at 0 is sufficient but not necessary to guarantee that $f^{-1}(0)$ is a k-dimensional submanifold.

Now that we have seen the definition seems to give us some good examples we can answer where the definition comes from: the answer is the Implicit Function Theorem.

Theorem 3.2 (Implicit Function Theorem)**.** *Let $f\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ and $(a,b) \in \mathbb{R}^n \times \mathbb{R}^m$ such that f is C^1 in an open set $U \ni (a,b)$ and $f(a,b) = 0$. Write*

$$\mathrm{d}f_{(a,b)} = (A\ B)$$

for an $m \times n$ matrix A and $B \in M_m(\mathbb{R})$, i.e.

$$B = (b_{ij}) = \left(\frac{\partial f_i}{\partial x_{n+j}}\right).$$

If $\det B \neq 0$, then there exists an open set $V \ni a$ in $\mathbb{R}^n$, an open set $W \ni b$ in $\mathbb{R}^m$ and $g \in C^1(V,W)$ such that

$$f^{-1}(0) \cap (V \times W) = \{(x, g(x)) \,:\, x \in V\}.$$

Remarks.

(a) The condition on f in the Implicit Function Theorem implies that $\mathrm{d}f_{(a,b)}\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ is surjective, which shows the link to our embedded submanifold definition (Definition 3.1).

(b) Informally, the conclusion of the Implicit Function Theorem says that the zero set of f, near (a,b), is the graph of g: this shows the connection to the graph example that we gave earlier.

(c) The restriction to $f(a,b) = 0$ is just for convenience: we can take $f(a,b) = c$ for any fixed $c \in \mathbb{R}^m$.

(d) If f is assumed to be C^k (for $k \geq 1$) or C^∞ on U in the Implicit Function Theorem then g will also be C^k or C^∞.

As we will now see, the main idea behind the proof of the Implicit Function Theorem is the Inverse Function Theorem we saw earlier.

Proof of Implicit Function Theorem. Define $F\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n \times \mathbb{R}^m$ by

$$F(x,y) = (x, f(x,y)).$$

Then, writing in block notation as in the statement of the theorem (Theorem 3.2),

$$\mathrm{d}F_{(a,b)} = \begin{pmatrix} I & 0 \\ A & B \end{pmatrix}.$$

Therefore $\det \mathrm{d}F_{(a,b)} = \det B \neq 0$, so $\mathrm{d}F_{(a,b)}$ is invertible.

By the Inverse Function Theorem (Theorem 2.6) there exists an open set $\tilde{V} \ni (a,b)$ and an open set $\tilde{W} \ni F(a,b) = (a,0)$ such that $f\colon \tilde{V} \to \tilde{W}$ is invertible with C^1 inverse. We may choose open sets $V \ni a$ and $W \ni b$ such that $V \times W \subseteq \tilde{V}$ and make $\tilde{W}$ smaller if necessary (i.e. replace $\tilde{W}$ by $F(V \times W)$) so that $F\colon V \times W \to \tilde{W}$ is invertible with C^1 inverse $F^{-1}\colon \tilde{W} \to V \times W$.

Since $F(x,y) = (x, f(x,y))$, there exists a surjective C^1 function $G\colon \tilde{W} \to W$ such that

$$F^{-1}(x,y) = (x, G(x,y)).$$

Let $\pi\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be the projection map. Then $\pi \circ F = f$ and

$$f(x, G(x,y)) = f \circ F^{-1}(x,y) = (\pi \circ F) \circ F^{-1}(x,y) = \pi \circ (F \circ F^{-1})(x,y) = \pi(x,y) = y$$

for all $y \in W$. Thus,

$$f(x, G(x,0)) = 0$$

and if we define $g\colon V \to W$ by

$$g(x) = G(x,0)$$

then the result follows. □

Remark. Since the function g in the Implicit Function Theorem satisfies $f(x, g(x)) = 0$ for all x, by differentiating this equation and using the Chain Rule one can quite easily compute the derivative of g, which can help to find what g is.

Let us look at a simple example to see the Implicit Function Theorem in action.

Example. Let $f\colon \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ be given by

$$f(x,y) = \|x\|^2 + y^2 - 1.$$

Then f is smooth as it is a polynomial and given $(a,b) \in \mathbb{R}^n \times \mathbb{R}$ with $f(a,b) = 0$ then

$$\mathrm{d}f_{(a,b)} = 2(a^{\mathrm{T}} \ b).$$

We therefore see that the Implicit Function Theorem applies if $b \neq 0$. Suppose that $b > 0$. In this case we can of course find g by solving $f(x,y) = 0$:

$$g(x) = \sqrt{1 - \|x\|^2}.$$

(We would have taken the negative square root if $b < 0$.) Here V is the open unit ball in $\mathbb{R}^n$ and WR, which contains b. We therefore see that $g\colon V \to W$ is indeed a smooth map, but it is not smooth (or even differentiable) in any neighbourhood of any x with $\|x\| = 1$.

There is a useful generalisation of the Implicit Function Theorem which we also give.

Theorem 3.3. *Let $f\colon \mathbb{R}^n \to \mathbb{R}^{n-k}$ and $a \in \mathbb{R}^n$ such that f is C^1 in an open set $U \ni a$ in $\mathbb{R}^n$ and $f(a) = 0$. If $\mathrm{d}f_a\colon \mathbb{R}^n \to \mathbb{R}^{n-k}$ is surjective then there exists an open set $V \ni a$ contained in U, an open set $W \ni 0$ in $\mathbb{R}^n$ and $\psi \in C^1(W,V)$ a bijection with C^1 inverse such that $\psi(0) = a$ and*

$$f \circ \psi(x_1, \ldots, x_n) = (x_{k+1}, \ldots, x_n).$$

Proof. Since $\mathrm{d}f_a$ is surjective we can permute the coordinates on $\mathbb{R}^n$ so that the last $n-k$ columns of $\mathrm{d}f_a$ form an invertible matrix, as in the statement of the Implicit Function Theorem (Theorem 3.2). If σ denotes this permutation then $f \circ \sigma$ now satisfies the conditions of the Implicit Function Theorem, so taking F essentially as in the proof of that theorem we have that

$$(f \circ \sigma) \circ F^{-1}(x_1, \ldots, x_n) = (x_{k+1}, \ldots, x_n).$$

Taking $\psi = \sigma \circ F^{-1}$ gives the result. □

Remark. Theorem 3.3 implies that if we have a k-dimensional submanifold M of $\mathbb{R}^n$, then given any $p \in M$ we can choose coordinates (given by the diffeomorphism ψ) on a neighbourhood $U \ni p$ in $\mathbb{R}^n$ such that $M \cap U$ is just the plane where $x_{k+1} = \ldots = x_n = 0$. More formally, for all $p \in M$ there exists an open set $U \ni p$ in $\mathbb{R}^n$, an open set $V \ni 0$ in $\mathbb{R}^k$, an open set $W \ni 0$ in $\mathbb{R}^{n-k}$ and a diffeomorphism $\psi : V \times W \to U$ such that

$$\psi(0) = p \quad \text{and} \quad M \cap U = \psi(V \times \{0\}).$$

As a result, every point in $M \cap U$ can be written uniquely as $\psi(x, 0)$ for $x \in V$ and we can define $\varphi : M \cap U \to V$ by

$$\varphi(\psi(x, 0)) = x$$

for all $x \in U$, which implies that p is the unique point in U such that $\varphi(p) = 0$. Then $\varphi : M \cap U \to V \subseteq \mathbb{R}^k$ is a diffeomorphism.

Let us look at a slightly more sophisticated example.

Example. Given $c \in \mathbb{R}$ let us consider the map $f \colon \mathbb{R}^3 \to \mathbb{R}$ given by

$$f(x_1, x_2, x_3) = x_1^2 + x_2^2 - x_3^2 - c.$$

We see that f is smooth (as it is a polynomial) and

$$\mathrm{d}f_{(a_1,a_2,a_3)} = 2(a_1 \;\; a_2 \;\; -a_3).$$

If $f(a) = 0$ then we see that $\mathrm{d}f_a$ is surjective as long as $a \neq 0$.

If $c < 0$ and $f(a_1, a_2, a_3) = 0$ then $a_3 \neq 0$ and so we are in the setting of the usual Implicit Function Theorem. We then find that $f^{-1}(0)$ is locally the graph of

$$g(x_1, x_2) = \pm\sqrt{x_1^2 + x_2^2 - c},$$

which is smooth everywhere as $c < 0$ (where the sign is determined by the sign of a_3). This makes sense as in this case $f^{-1}(0)$ is a hyperboloid of two sheets, and so is the graph of two functions (whose images are disjoint).

If $c > 0$ and $f(a_1, a_2, a_3) = 0$ then now a_3 can vanish, but then $a_1^2 + a_2^2 = c > 0$ so one of a_1, a_2 is non-zero. If $a_2 \neq 0$, say, then we get that $f^{-1}(0)$ is locally the graph of

$$g(x_1, x_3) = \sqrt{c + x_3^2 - x_1^2}.$$

We see that we have to change coordinates to describe the local graph and also that $f^{-1}(0)$ is no longer a global graph: this makes sense again, as it is a hyperboloid of one sheet.

Finally, if $c = 0$ then $0 \in f^{-1}(0)$ and none of our theorems apply. We see that the type of implicit function we would be led to consider would be, for example,

$$g(x_1, x_2) = \sqrt{x_1^2 + x_2^2}$$

which is definitely not smooth near 0. One can use this to argue that $f^{-1}(0)$ is, in fact, not a submanifold of $\mathbb{R}^3$, but we shall see a different argument later.

As we have seen, k-dimensional submanifolds of $\mathbb{R}^n$ are locally graphs over k-dimensional planes in $\mathbb{R}^n$, but what are these planes geometrically? The answer is that they are the *tangent spaces* to the submanifold.

Definition 3.4. Let $M \subseteq \mathbb{R}^n$ be a k-dimensional submanifold and let $p \in M$.

Let $\alpha \colon (-\epsilon, \epsilon) \to M$ be a curve with $\alpha(0) = p$. We say that $\alpha'(0) \in \mathbb{R}^n$ is the *tangent vector* to α at p. We let T_pM be the set of all tangent vectors to curves $\alpha \colon (-\epsilon, \epsilon) \to M$ with $\alpha(0) = p$. Then T_pM is the *tangent space* to M at p.

Let us start with the simplest possible example, though it actually will be very useful.

Example. Let

$$M = \{(x_1, \dots, x_n) \in \mathbb{R}^n \,:\, x_{k+1} = \dots = x_n = 0\}.$$

Then any curve α in M will have $\alpha'(0) \in M$. Therefore, $T_pM \subseteq M$. Moreover, given any $p, v \in M$ we have that $\alpha(t) = p + tv$ is a curve in M with $\alpha(0) = p$ and $\alpha'(0) = v$, so $v \in T_pM$. Hence,

$$T_pM = M$$

for all $p \in M$ and is, in particular, a k-dimensional vector space.

We can now show how to find all tangent spaces to submanifolds.

Proposition 3.5. *Let M be a k-dimensional submanifold of $\mathbb{R}^n$ and let $p \in M$. Then the tangent space T_pM to M at p is a k-dimensional vector space and if $f\colon U \subseteq \mathbb{R}^n \to \mathbb{R}^{n-k}$, where $U \ni p$ is open, is smooth with $M \cap U = f^{-1}(0)$ and 0 is a regular value of f, then*

$$T_pM = \ker \mathrm{d}f_p$$

Proof. Let α be any curve in M with $\alpha(0) = p$. We know from our generalisation (Theorem 3.3) of the Implicit Function Theorem that there exist open sets $V \ni p$ and $W \ni 0$ in $\mathbb{R}^n$ and a diffeomorphism $\psi\colon W \to V$ such that $\psi(0) = p$ and, if $\pi\colon \mathbb{R}^n = \mathbb{R}^k \times \mathbb{R}^{n-k} \to \mathbb{R}^{n-k}$ denotes the projection onto the last $n-k$ coordinates:

$$f \circ \psi = \pi.$$

Let $P = \ker \pi$ (which is just the plane from the example above). Then the formula above says that

$$M \cap V = \psi(P \cap W).$$

Therefore, as ψ is a diffeomorphism, there exists a curve β in $W \cap P$ with $\beta(0) = 0$ and $\psi \circ \beta = \alpha$ on $M \cap V \ni p$. Therefore, by the Chain Rule,

$$\alpha'(0) = (\psi \circ \beta)'(0) = \mathrm{d}\psi_{\beta(0)} \circ \beta'(0) = \mathrm{d}\psi_0\big(\beta'(0)\big).$$

Since $\mathrm{d}\psi_0$ is an isomorphism (Lemma 2.5) and using the example above, we deduce that

$$\mathrm{d}\psi_0(P) = T_pM$$

and hence is a k-dimensional vector space.

Now we observe that $f \circ \alpha = 0$ so by the Chain Rule:

$$0 = (f \circ \alpha)'(0) = \mathrm{d}f_{\alpha(0)} \circ \alpha'(0) = \mathrm{d}f_p\big(\alpha'(0)\big).$$

Hence, T_pM is contained in $\ker \mathrm{d}f_p$, which is a k-dimensional vector space as well (by Rank-Nullity), giving the result. □

Let us apply this result to a simple example to see that it matches our intuition.

Example. Let us consider the unit sphere $\mathcal{S}^n$ in $\mathbb{R}^{n+1}$. Then we know that $\mathcal{S}^n = f^{-1}(0)$ where

$$f(x) = \|x\|^2 - 1$$

and

$$\mathrm{d}f_a = 2a^{\mathrm{T}}.$$

Therefore, by our result,

$$T_pM = \ker \mathrm{d}f_p = \{y \in \mathbb{R}^{n+1} \,:\, p^{\mathrm{T}}y = 0\} = \langle p \rangle^{\perp},$$

the vector space orthogonal to the span of p. This is clear geometrically.

We now revisit a previous example to see why it is not a submanifold.

Example. If we return to the example of $f\colon \mathbb{R}^3 \to \mathbb{R}$ given by

$$f(x_1, x_2, x_3) = x_1^2 + x_2^2 - x_3^2$$

one can see that if we take a curve α in $f^{-1}(0)$ with $\alpha(0) = 0$ then $\alpha'(0) \in f^{-1}(0)$. Moreover, every $v \in f^{-1}(0)$ is $\alpha'(0)$ for some curve α in $f^{-1}(0)$: just take $\alpha(t) = tv$. However, $f^{-1}(0)$ is not a vector space and so $f^{-1}(0)$ cannot be a submanifold of $\mathbb{R}^3$.

We now look at some examples of matrix groups which we showed were submanifolds.

Example. If we consider $\mathrm{SL}(n, \mathbb{R})$ then by our earlier calculations we find that

$$T_I \,\mathrm{SL}(n, \mathbb{R}) = \{B \in M_n(\mathbb{R}) \,:\, \mathrm{tr}\, B = 0\},$$

the trace-free matrices, which is a vector space of dimension $n^2 - 1$.

Example. If we instead consider $\mathrm{O}(n)$ (or $\mathrm{SO}(n)$) then again by our earlier calculations we find that

$$T_I \,\mathrm{O}(n) = \{B \in M_n(\mathbb{R}) \,:\, B^{\mathrm{T}} + B = 0\},$$

the skew-symmetric matrices, which is a vector space of dimension $\frac{1}{2}n(n-1)$.

Remark. (Not examinable). The tangent space at the identity of matrix groups like this, which are submanifolds, is particularly interesting as it is related to *Lie algebras*.

We now consider variational problems on submanifolds, that is, trying to minimize/maximize a function on the submanifold. This gives the theory of *Lagrange multipliers*.

We first start with an elementary observation.

Lemma 3.6. *Let $g\colon \mathbb{R}^n \to \mathbb{R}$ be differentiable with a local minimum (or maximum) at $a \in \mathbb{R}^n$, i.e. there exists some open set $U \ni a$ such that for all $x \in U$ we have $g(x) \geq g(a)$. Then*

$$\mathrm{d}g_a = 0.$$

Proof. Suppose, for a contradiction, that $\mathrm{d}g_a \neq 0$. Then there exists $v \in \mathbb{R}^n$ with $\|v\| = 1$ such that

$$\mathrm{d}g_a(v) > 0.$$

Now, by the definition of the differential, we have for all $t \in (-\epsilon, \epsilon)$ such that $a + tv \in U$,

$$g(a + tv) = g(a) + t\mathrm{d}g_a(v) + t\eta(t)$$

for a function $\eta(t)$ with $\eta(t) \to 0$ as $t \to 0$. Hence, by making ϵ smaller we can ensure that

$$|\eta(t)| < \frac{1}{2}\mathrm{d}g_a(v)$$

for all t. This means that

$$0 \leq g(a + tv) - g(a) = t(\mathrm{d}g_a(v) + \eta(t))$$

which is impossible for $t < 0$. □

We can now prove our result.

Theorem 3.7 (Lagrange multipliers)**.** *Let M be a k-dimensional submanifold of $\mathbb{R}^n$ and let $p \in M$. Let $f \in C^1(U, \mathbb{R}^{n-k})$, where $U \ni p$ is open in $\mathbb{R}^n$, be such that $f^{-1}(0) = M \cap U$ and 0 is a regular value of f. Suppose that $g\colon \mathbb{R}^n \to \mathbb{R}$ is C^1 and has a local minimum (or maximum) on M at p. Then if $f = (f^1, \dots, f^{n-k})$ there exist $\lambda_1, \dots, \lambda_{n-k} \in \mathbb{R}$ (called Lagrange multipliers) such that*

$$\mathrm{d}g_p = \sum_{i=1}^{n-k} \lambda_i \mathrm{d}f^i_p.$$

Proof. Suppose first that

$$f(x_1, \dots, x_n) = (x_{k+1}, \dots, x_n).$$

Then

$$\mathrm{d}f^i_p = e_i,$$

the unit vector in the x_i direction. We are then asking that g has a local minimum on $\mathbb{R}^k = \{x_{k+1} = \dots = x_n = 0\}$, so $\mathrm{d}g_p$ must vanish on directions in $\mathbb{R}^k$ by Lemma 3.6. This then means that $\mathrm{d}g_p$ lies in the $\mathbb{R}^{n-k}$ and hence in the span of the $\mathrm{d}f^i_p$.

In general, we know from our generalisation (Theorem 3.3) of the Implicit Function Theorem that there exists a diffeomorphism ψ from some open set containing 0 to an open set in U containing p such that

$$f \circ \psi(x_1, \dots, x_n) = (x_{k+1}, \dots, x_n).$$

Therefore, by considering $g \circ \psi$ instead, which now has a local minimum at 0 in $\mathbb{R}^k$, everything now follows because $\mathrm{d}\psi_0$ is an isomorphism (Lemma 2.5). □

Remarks.

(a) The result says that $\mathrm{d}g_p$ vanishes on $T_pM = \ker \mathrm{d}f_p$, which means that, viewed as a vector in $\mathbb{R}^n$, it is *normal* to T_pM.

(b) Solving for the Lagrange multipliers finds a critical point for g on M, but it does not determine what type of critical point this is. We usual need a separate argument (often involving bounds on f, g) to show that the critical point is actually a minimum or maxium.

Example. Suppose we want to find the closest point to the origin on a given surface. For example, let us take $r > 0$ and consider

$$M = \{(x_1, x_2, x_3) \in \mathbb{R}^3 \,:\, x_1^2 + x_2^2 - x_3^2 = -r^2,\ x_3 > 0\}.$$

We then have that

$$f(x_1, x_2, x_3) = x_1^2 + x_2^2 - x_3^2 + r^2 \quad \text{and} \quad g(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2.$$

Then from our earlier calculations and Theorem 3.7, we have that at such a closest point (a_1, a_2, a_3) on M there must be a constant $\lambda \in \mathbb{R}$ such that

$$\mathrm{d}g_a = 2(a_1 \ \ a_2 \ \ a_3) = \lambda \mathrm{d}f_a = 2\lambda(a_1 \ \ a_2 \ \ -a_3)$$

Now since $a_3 \neq 0$ this forces $\lambda = -1$ and $a_1 = a_2 = 0$, so the only extremum of g on M is

$$(0, 0, r).$$

Indeed, since there is only one critical point and the function g must have a minimum, we see that it is indeed the closest point (as we already knew).

To finish this section we look at level sets of functions $f\colon M \to N$ between submanifolds, i.e. the sets $f^{-1}(c)$ where $c \in N$. We first need a definition.

Definition 3.8. Let $M \subseteq \mathbb{R}^m$ and $N \subseteq \mathbb{R}^n$ be submanifolds. A function $f\colon M \to N$ is *smooth* if for all $p \in M$ there is an open set $U \ni p$ in $\mathbb{R}^m$ such that f on $U \cap M$ is the restriction of a smooth map from U to $\mathbb{R}^n$.

We now make an observation.

Lemma 3.9. *Let M, N be submanifolds and let $f\colon M \to N$ be smooth. Then f induces a linear map*

$$\mathrm{d}f_p\colon T_pM \to T_{f(p)}N.$$

Proof. Let $\alpha\colon (-\epsilon, \epsilon) \to M$ be a curve with $\alpha(0) = p$. Then $f \circ \alpha$ is a curve in N with $f \circ \alpha(0) = f(p)$. By the Chain Rule,

$$(f \circ \alpha)'(0) = \mathrm{d}f_{\alpha(0)} \circ \alpha'(0) = \mathrm{d}f_p(\alpha'(0)).$$

Since $(f \circ \alpha)'(0) \in T_{f(p)}N$ the result follows. □

Remark. To compute the differential of $f\colon M \to N$ at p we take the smooth map F defined on the open set $U \subseteq \mathbb{R}^m$ containing p which restricts to f, compute $\mathrm{d}F_p\colon \mathbb{R}^m \to \mathbb{R}^n$ as usual and then restrict the differential to T_pM to find $\mathrm{d}f_p$. We will see this in practice below.

We can now understand when level sets of functions between submanifolds are submanifolds.

Theorem 3.10 (Regular Value Theorem)**.** *Let $M \subseteq \mathbb{R}^m$ and $N \subseteq \mathbb{R}^n$ be submanifolds of dimensions k and l respectively and let $f\colon M \to N$ be smooth. If $c \in f(M) \subseteq N$ is a regular value of f, i.e. $\mathrm{d}f_p : T_pM \to T_cN$ is surjective for all $p \in f^{-1}(c)$, then $f^{-1}(c)$ is a submanifold of dimension $k - l$ and*

$$T_pf^{-1}(c) = \ker \mathrm{d}f_p.$$

Proof. **(Not examinable).** Let $p \in f^{-1}(c)$. By definition, we want to show that there is a smooth function g defined on some open neighbourhood $U \ni p$ in $\mathbb{R}^m$ into $\mathbb{R}^{m-(k-l)} = \mathbb{R}^{m-k} \times \mathbb{R}^l$ such that $g^{-1}(0) = f^{-1}(c) \cap U$ and 0 is a regular value of g.

Since M is a k-dimensional submanifold, there exists an open set $U_1 \ni p$ in $\mathbb{R}^m$ and a smooth function $g_1 : U_1 \to \mathbb{R}^{m-k}$ such that $g_1^{-1}(0) = M \cap U_1$ and 0 is a regular value of g_1. This gives us the first component of g.

To obtain the second component, as discussed after Theorem 3.3, we note that there exists an open set $U_2 \ni c$ in $\mathbb{R}^n$, an open set $V_2 \ni 0$ in $\mathbb{R}^l$ and a diffeomorphism $\varphi_2 : N \cap U_2 \to V_2$ with $\varphi_2(c) = 0$. Note, in particular, $c \in N \cap U_2$ is the unique point such that $\varphi_2(c) = 0$.

Finally, we note that by the definition of smooth function, we can choose $U_1 \ni p$ so that there exists a smooth function $F : U_1 \to \mathbb{R}^n$ such that $F|_{M \cap U_1} = f$.

Choosing $U \ni p$ in $\mathbb{R}^m$ such that $U \subseteq U_1$ and $F(U) \subseteq U_2$, we can then define $g : U \to \mathbb{R}^{m-k} \times \mathbb{R}^l$ by

$$g(x) = \big(g_1(x), \varphi_2 \circ F(x)\big) \quad \text{for all } x \in U.$$

Then g is smooth since g_1, φ_2, F are smooth.

We see that $g(x) = 0$ forces $g_1(x) = 0$ which occurs if and only if $x \in M \cap U$. For $x \in M \cap U$, $F(x) = f(x)$ and $\varphi_2 \circ f(x) = 0$ if and only if $f(x) = c$, i.e. $x \in f^{-1}(c)$. Hence, $g^{-1}(0) = f^{-1}(c) \cap U$.

Finally, we need to see that 0 is a regular value of g and $\ker \mathrm{d}g_p = \ker \mathrm{d}f_p$, which then completes the proof. If $a \in g^{-1}(0) = f^{-1}(c) \cap U$ and $v \in \mathbb{R}^n$,

$$\mathrm{d}g_a(v) = \big(\mathrm{d}(g_1)_a(v), \mathrm{d}(\varphi_2)_c \circ \mathrm{d}F_a(v)\big).$$

First $v \in \ker \mathrm{d}g_a$ forces $v \in \ker \mathrm{d}(g_1)_a = T_aM$. Then $\mathrm{d}F_a(v) = \mathrm{d}f_a(v)$. Since $\mathrm{d}(\varphi_2)_c$ is an isomorphism (Lemma 2.5) we then deduce that $v \in \ker \mathrm{d}g_a$ if and only if $v \in \ker \mathrm{d}f_a$. Since c is a regular value of f, $\ker \mathrm{d}f_a = \ker \mathrm{d}g_a$ is $(k-l)$-dimensional for all $a \in g^{-1}(0)$, so 0 is a regular value of g and $\ker \mathrm{d}f_p = \ker \mathrm{d}g_p$ as desired. □

Example. Let $f\colon \mathcal{S}^2 \to \mathbb{R}$ be given by

$$f(x_1, x_2, x_3) = x_3.$$

Then f is smooth, as the restriction of a smooth map (which we also call f for simplicity) on $\mathbb{R}^3$ which has differential

$$\mathrm{d}f_{(a_1,a_2,a_3)} = (0\ 0\ 1)$$

in $\mathbb{R}^3$. However we need to restrict the differential to $T_p\mathcal{S}^2$, so we see that if we let $p = (0, 0, \pm 1)$ then

$$T_p\mathcal{S}^2 = \{(x_1, x_2, 0) \,:\, x_1, x_2 \in \mathbb{R}^2\}$$

and so $\mathrm{d}f_p : T_p\mathcal{S}^2 \to \mathbb{R}$ is the zero map: this is exactly when $f(p) = \pm 1$. Otherwise, the differential $\mathrm{d}f_p$ is surjective (since there will be a vector in $T_p\mathcal{S}^2$ with a component in the x_3 direction).

The Regular Value Theorem then implies that $f^{-1}(c)$ is a 1-dimensional submanifold for $c \in f(\mathcal{S}^2) = [-1, 1]$ except when $c = \pm 1$. This makes sense as these are just lines of latitude when $c \in (-1, 1)$.

Just to finish this section we explain why we have the word "embedded" highlighted in the definition of submanifold by talking about three important classes of maps.

Definition 3.11. Let M, N be submanifolds and let $f\colon M \to N$ be smooth.

We say that f is an *immersion at* $p \in M$ if $\mathrm{d}f_p : T_pM \to T_{f(p)}N$ is injective, and say that f is an *immersion* if it is an immersion at all $p \in M$.

We say that an immersion f is an *embedding* if $f\colon M \to f(M)$ is a homeomorphism, i.e. a continuous map with continuous inverse. An important fact is that if M is compact (which is the same as closed and bounded) then an injective immersion will be an embedding.

We say that f is a *submersion at* $p \in M$ if $\mathrm{d}f_p : T_pM \to T_{f(p)}N$ is surjective, and say that f is a *submersion* if it is a submersion at all $p \in M$.

Example. Let $f\colon \mathbb{R} \to \mathbb{R}^2$ be given by

$$f(\theta) = (\cos\theta, \sin\theta).$$

Then f is smooth and

$$\mathrm{d}f_\theta = \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix}$$

which is clearly an injective map from $\mathbb{R}$ to $\mathbb{R}^2$. Hence f is an immersion, but it is not an embedding since it is not injective: $f(\theta + 2\pi) = f(\theta)$ for all θ.

Example. If M is a k-dimensional submanifold of $\mathbb{R}^n$ then the inclusion map $\iota\colon M \to \mathbb{R}^n$ is an embedding. More generally, if $f\colon M \to \mathbb{R}^n$ is an embedding then $f(M)$ is also a k-dimensional submanifold.

Example. Let $f\colon \mathbb{R}^{n+1} \to \mathbb{R}$ be given by projection onto the last coordinate:

$$f(x_1, \ldots, x_{n+1}) = x_{n+1}.$$

Then f is smooth and

$$\mathrm{d}f_{(x_1,x_2)} = (0\ \ldots\ 0\ 1),$$

which is definitely surjective, so f is a submersion. However, if we restrict f to the sphere then $f\colon \mathcal{S}^n \to \mathbb{R}$ is not a submersion at $x_{n+1} = \pm 1$ (as we saw).

This shows the relationships between submersions, projections and regular values.

Example. Let $f\colon \mathbb{R} \times \{0\} \cup \mathbb{R} \times \{1\} \to \mathbb{R}^2$ be given by

$$f(x, 0) = (x, 0) \quad \text{and} \quad f(x, 1) = (0, x).$$

Then f is smooth, it is an immersion (as it is effectively the identity on each copy of $\mathbb{R}$) but it is not injective as $f(0,0) = f(0,1) = (0,0)$. The image is two straight lines meeting at just one point: this is a good local model for thinking about immersions which are not embeddings.

Remark. (Not examinable). The previous example and an example on a problem sheet hint at a subtle issue we have avoided, namely the issue of *immersed submanifolds*, which are the images of immersions. Many things are still true for immersed submanifolds as for embedded submanifolds, but not everything, so this is why we restricted ourselves to the embedded ones.

4 Transversality

Our next goal is to try to answer the following question: when is the intersection of two submanifolds still a submanifold (of the dimension we think it should be)?

To begin with we have a fundamental and important result in analysis which plays a crucial role in geometry and topology: *Sard's Theorem.*

Theorem 4.1 (Sard's Theorem)**.** *Let U be an open set in $\mathbb{R}^n$ and let $f\colon U \to \mathbb{R}^m$ be a smooth map. Let C be the set of critical points of f, i.e.*

$$C = \{a \in U \,:\, \operatorname{rank} \mathrm{d}f_a < m\}.$$

Then $f(C)$, the set of critical values of f, has Lebesgue measure zero in $\mathbb{R}^m$.

Remark. If you are not taking Integration, then $f(C)$ having Lebesgue measure zero just means that for all $\epsilon > 0$ there exists a collection $\{I_j \,:\, j \in \mathbb{N}\}$ of cubes in $\mathbb{R}^m$ such that $f(C) \subseteq \bigcup_j I_j$ and the sum of the volumes of the I_j is strictly less than ϵ: $\sum_j \operatorname{vol}(I_j) < \epsilon$.

A corollary of Sard's Theorem is the following, which is used frequently.

Corollary 4.2. *Let $f\colon M \to N$ be a smooth map between submanifolds. Then the set of regular values of f is dense in N.*

Proof. **(Not examinable).** We can cover the k-dimensional submanifold M with a countable collection of open sets $\{V_j \,:\, j \in \mathbb{N}\}$ where M is given by the graph of a function on an open set U_j in $\mathbb{R}^k$. Using this, we can consider f locally (i.e. on each V_j) as a function on an open set (namely U_j) in $\mathbb{R}^k$. The result then follows immediately from Sard's Theorem. □

Let us try to understand what Sard's Theorem is saying in an example.

Example. Let $f\colon \mathbb{R} \to \mathbb{R}$ be smooth. The condition that $\operatorname{rank} \mathrm{d}f_a < 1$ is just $f'(a) = 0$, so in this case C is just the critical points of f in the sense we already knew. Geometrically, these give the points on the graph where it becomes horizontal. Sard's Theorem then says that almost every horizontal line in $\mathbb{R}^2$ which meets the graph will meet it transversely.

This clearly generalizes to $f\colon \mathbb{R}^n \to \mathbb{R}$, where now Sard's Theorem says that almost every hyperplane in $\mathbb{R}^{n+1}$ which meets the graph of f will do so transversely. As an explicit example, take

$$f(x_1, x_2) = x_1^2 + x_2^2.$$

Then $C = \{(0,0)\}$, $f(C) = \{(0,0)\}$ (which clearly has measure zero) and the intersection of $\{x_3 = c\}$ in $\mathbb{R}^3$ with the graph of f will be either empty if $c < 0$, a point if $c = 0$ or a circle if $c > 0$.

A nice application of Sard's Theorem involves the closed unit ball

$$\overline{B}^n = \{x \in \mathbb{R}^n \,:\, \|x\| \leq 1\}.$$

This is not quite a submanifold of $\mathbb{R}^n$, but instead is a *submanifold with boundary.* To define this we let $\overline{H}^n$ be the (closed) *upper half-space*

$$\overline{H}^n = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \,:\, x_n \geq 0\}$$

and let its *boundary* be

$$\partial\overline{H}^n = \{(x_1, \ldots, x_n) \in \overline{H}^n \,:\, x_n = 0\}.$$

Definition 4.3. A subset M of $\mathbb{R}^n$ is a *k-dimensional (embedded) submanifold with boundary* if for all $p \in M$ there exists an open set $U \ni p$ in $\mathbb{R}^n$, an open set $V \subseteq \overline{H}^k$ and a diffeomorphism $\varphi : U \cap M \to V$. The *boundary* ∂M of M is the set of points $p \in M$ such that $\varphi(p) \in \partial\overline{H}^k$ for some φ as above.

Remarks.

(a) By our generalisation (Theorem 3.3) of the Implicit Function Theorem, this agrees with our previous definition of submanifold if $\partial M = \emptyset$, i.e. if each open set V as above is actually contained in the open upper half-plane $H^k = \overline{H}^k \setminus \partial \overline{H}^k$.

(b) We see that $M \setminus \partial M$ is a k-dimensional submanifold and ∂M is a $(k-1)$-dimensional submanifold.

Example. The closed unit ball $\overline{B}^n \subseteq \mathbb{R}^n$ is an n-dimensional submanifold with boundary $\partial \overline{B}^n = \mathcal{S}^{n-1}$.

We will utilize several times the following important fact.

Lemma 4.4. *Let M be a compact 1-dimensional submanifold with boundary. Then ∂M is a (finite) even number of points.*

Proof. It is a non-trivial fact the only compact 1-dimensional submanifolds with boundary are finite disjoint unions of arcs of finite length. These arcs are then either loops (and so have no boundary) or have two endpoints. □

There is a slight extension of the Regular Value Theorem (Theorem 3.10) for submanifolds with boundary that we shall use and so we will state it here (again, without proof).

Theorem 4.5. *Let $M \subseteq \mathbb{R}^m$ be a k-dimensional submanifold with boundary, let $N \subseteq \mathbb{R}^n$ be an l-dimensional submanifold and let $f \colon M \to N$ be smooth. If $c \in f(M) \subseteq N$ is a regular value of f and of $f|_{\partial M}$, i.e. both $\mathrm{d}f_p : T_pM \to T_cN$ is surjective for all $p \in f^{-1}(c)$ and $\mathrm{d}f_q : T_q\partial M \to T_cN$ is surjective for all $q \in f^{-1}(c) \cap \partial M$, then $f^{-1}(c) \subseteq \mathbb{R}^m$ is a $(k-l)$-dimensional submanifold with boundary and*

$$\partial f^{-1}(c) = f^{-1}(c) \cap \partial M.$$

Given this, we can now prove the key lemma that will give our nice application.

Lemma 4.6. *There is no smooth map $f \colon \overline{B}^n \to \mathcal{S}^{n-1}$ such that $f|_{\mathcal{S}^{n-1}} = \mathrm{id}$.*

Proof. Suppose for a contradiction that f exists. Let $y \in \mathcal{S}^{n-1}$ be a regular value of f, which must exist by Sard's Theorem. Then certainly y is a regular value for $f|_{\mathcal{S}^{n-1}} = \mathrm{id}$ as well. Therefore, by the extension of the Regular Value Theorem (Theorem 4.5), $f^{-1}(y)$ is a 1-dimensional submanifold with boundary and

$$\partial f^{-1}(y) = \partial B^n \cap f^{-1}(y) = \mathcal{S}^{n-1} \cap f^{-1}(y) = \{y\}.$$

Moreover, $f^{-1}(y)$ is compact. (To see this: if (x_j) is a sequence in $f^{-1}(y)$ converging to x then we see that $f(x_j) \to f(x)$ as $j \to \infty$ since f is continuous, so $x \in f^{-1}(y)$ and thus $f^{-1}(y)$ is closed; and $f^{-1}(y)$ is bounded in $\mathbb{R}^n$ because it is contained in $\overline{B}^n$ which is bounded.) Hence, by Lemma 4.4, $\partial f^{-1}(y)$ must be an even number of points, which is a contradiction. □

Theorem 4.7 (Brouwer's Fixed Point Theorem)**.** *Every continuous map $F \colon \overline{B}^n \to \overline{B}^n$ has a fixed point.*

Proof. We will do the proof just in the case when F is smooth. When F is not smooth, one just approximates it by a smooth (in fact polynomial) map G using the Weierstrass Approximation Theorem (and performs a suitable rescaling so that G maps $\overline{B}^n$ to $\overline{B}^n$ still).

Suppose for a contradiction that F has no fixed point. For $x \in \overline{B}^n$ we let ℓ_x be the straight line through x and $F(x)$, which meets $\mathcal{S}^{n-1}$ in two points. We let $f(x)$ be the point in $\ell_x \cap S^{n-1}$ which is closer to x than $F(x)$. Then $f \colon \overline{B}^n \to \mathcal{S}^{n-1}$ is smooth and $f(x) = x$ for all $x \in \mathcal{S}^{n-1}$, which is impossible by Lemma 4.6. □

We will now explain how to prove Sard's Theorem (in a special case).

Proof of Sard's Theorem (special case). **(Not examinable).** We will only do the special case where $n = m$, which is the "easy" case. The general case can be tackled using an induction argument involving the number of partial derivatives which vanish at a given point in C. This is subtle and in fact shows that the smoothness of f is important.

We know that U can be written as the countable union of compact cubes in $\mathbb{R}^n$ (Analysis I). We also know that a countable union of sets with measure zero still has measure zero: this is straightforward from the definition. We can therefore restrict to f defined on a compact cube $K \subseteq U$ with sides of length l.

Let $a \in C \cap K$. Then $\mathrm{d}f_a(\mathbb{R}^n)$ is a proper subspace of $\mathbb{R}^n$ (as $\operatorname{rank} \mathrm{d}f_a < n = m$) so there exists a non-zero linear map $\pi : \mathbb{R}^n \to \mathbb{R}$ such that

$$\pi \circ \mathrm{d}f_a = 0.$$

Let H be the following hyperplane containing $f(a)$:

$$H = \{x \in \mathbb{R}^n \,:\, \pi(x - f(a)) = 0\}.$$

Then $f(a) + \mathrm{d}f_a(y) \in H$ for all $y \in \mathbb{R}^n$ and so we must have that

$$d(f(x), H) \leq \|f(x) - f(a) - \mathrm{d}f_a(x - a)\|.$$

Since f is smooth and K is compact, we may apply Taylor's Theorem to deduce that there is a constant $\lambda > 0$ such that for all $x, y \in K$:

$$\|f(x) - f(y) - \mathrm{d}f_y(x - y)\| \leq \lambda \|x - y\|^2.$$

We deduce that

$$d(f(x), H) \leq \lambda \|x - a\|^2.$$

Let $\epsilon > 0$ and let

$$\mu = \sup\{\|\mathrm{d}f_x\| \,:\, x \in K\},$$

which is finite as K is compact. If $\|x - a\| \leq \epsilon$ then $d(f(x), H) \leq \lambda\epsilon^2$ and by the Mean Value Theorem

$$\|f(x) - f(a)\| \leq \mu \|x - a\| \leq \mu\epsilon.$$

Hence,

$$\{f(x) \,:\, x \in K, \|x - a\| \leq \epsilon\}$$

is contained in a neighbourhood of width $2\lambda\epsilon^2$ around H and a cube of side length $2\mu\epsilon$, which has volume

$$v(\epsilon) \leq 2\lambda\epsilon^2(2\mu\epsilon)^{n-1} = 2^n \lambda \mu^{n-1} \epsilon^{n+1}.$$

If we now subdivide K into N^n cubes of side length l/N, we can repeat the argument above for each of these cubes. In each of these small cubes $\tilde{K}$ we will have

$$\|x - y\| \leq \frac{l\sqrt{n}}{N}$$

and so if $C \cap \tilde{K} \neq \emptyset$ then

$$\{f(x) \,:\, x \in \tilde{K}\}$$

will be contained in a set of volume (taking $\epsilon = \frac{l\sqrt{n}}{N}$)

$$V(N) \leq v\left(\frac{l\sqrt{n}}{N}\right) \leq 2^n \lambda \mu^{n-1} \left(\frac{l\sqrt{n}}{N}\right)^{n+1}.$$

There are at most N^n such contributions to the total volume, so we have that $f(C \cap K)$ is certainly contained in a set of volume

$$V \leq N^n V(N) \leq \frac{2^n \lambda \mu^{n-1} l^{n+1} n^{\frac{n+1}{2}}}{N}.$$

Since this tends to zero as $N \to \infty$, we deduce that $f(C \cap K)$ has measure zero, which gives the result. $\square$

As we saw, Sard's Theorem implied a transversality result for graphs of functions and hyperplanes. We want to formalize now what we mean by transversality.

Definition 4.8. Let M_1, M_2 be submanifolds of $\mathbb{R}^n$. We say that M_1, M_2 *intersect transversely* if for all $p \in M_1 \cap M_2$ we have

$$\mathbb{R}^n = T_pM_1 + T_pM_2.$$

If this holds we may write $M_1 \pitchfork M_2$.

Example. If we take M_1, M_2 to be curves in $\mathbb{R}^2$ then they intersect transversely if and only if their tangent spaces are distinct at any intersection point. For example, any two distinct straight lines in the plane through 0 must intersect transversely, but two distinct circles which touch at one point do not.

Remark. If $\dim M_1 + \dim M_2 = n$ then for $M_1 \pitchfork M_2$ to hold we must have

$$\mathbb{R}^n = T_pM_1 \oplus T_pM_2$$

at all intersection points $p \in M_1 \cap M_2$. If instead $\dim M_1 + \dim M_2 > n$ then the sum can never be direct.

We can generalize the notion of transversality further to talk about maps.

Definition 4.9. Let M, N be submanifolds and let $f\colon M \to N$ be smooth. Let $L \subseteq N$ be another submanifold and let $K \subseteq M$. We say that f is *tranverse to L along K*, and write $f \pitchfork_K L$, if for all $p \in K$ such that $f(p) \in L$ we have that

$$T_{f(p)}N = \mathrm{d}f_p(T_pM) + T_{f(p)}L.$$

If $K = M$ we just write $f \pitchfork L$ and say f is *transverse to L*.

Remark. Note that K is just a *subset* of M: it need not be a submanifold.

Example. Let us see how this relates to our previous definition (Definition 4.8). Let $M = M_1$ and $L = M_2$ be submanifolds of $N = \mathbb{R}^n$, let $f\colon M \to \mathbb{R}^n$ be the inclusion map and let $K = M \cap L = M_1 \cap M_2$. We see that $f(p) = p \in L$ for all $p \in K$ and $\mathrm{d}f_p : T_pM \to \mathbb{R}^n$ is the inclusion map. Therefore f is transverse to L along K if and only if

$$\mathbb{R}^n = T_{f(p)}N = T_pM + T_pL = T_pM_1 + T_pM_2.$$

This says that the two submanifolds M_1, M_2 intersect transversely. In fact, we can just take $K = M = M_1$ and get that $f \pitchfork L$.

Remark. Roughly speaking, the more general notion of transversality in Definition 4.9 means we are thinking about the intersection between $f(M)$ and L in N at the points in $f(K) \cap L$.

Using the notion of transversality in Definition 4.9 we can generalize the Regular Value Theorem (Theorem 3.10) even further.

Theorem 4.10. *Let M, N be submanifolds, let $f \in C^\infty(M, N)$ and let $L \subseteq N$ be a submanifold. If f is transverse to L ($f \pitchfork L$) then $f^{-1}(L)$ is a submanifold. Moreover, the codimension of $f^{-1}(L)$ in M is*

$$\dim M - \dim f^{-1}(L) = \dim N - \dim L.$$

Proof. Suppose first that $N = U \times V$ where $U \subseteq \mathbb{R}^k$, $V \subseteq \mathbb{R}^l$ are open sets containing 0, and $L = U \times \{0\}$. Let $\pi : \mathbb{R}^k \times \mathbb{R}^l \to \mathbb{R}^l$ be the projection map and let $g = \pi \circ f\colon M \to \mathbb{R}^l$ so that

$$f^{-1}(L) = f^{-1}(U \times \{0\}) = g^{-1}(0).$$

Then $f\colon M \to U \times V$ is transverse to $U \times \{0\}$ if and only if

$$\mathrm{d}f_p(T_pM) + \mathbb{R}^k = \mathbb{R}^k \times \mathbb{R}^l.$$

Since π is a linear map we have $\mathrm{d}\pi_a = \pi$ and therefore

$$\mathrm{d}g_p = \mathrm{d}\pi_{f(p)} \circ \mathrm{d}f_p = \pi \circ \mathrm{d}f_p.$$

We deduce that $f \pitchfork U \times \{0\}$ if and only if 0 is a regular value of g. This shows that $f^{-1}(U \times \{0\})$ is a submanifold of codimension l.

By the Implicit Function Theorem, we can always locally (i.e. in an open neighbourhood of any $p \in L$) reduce to the case above, which then gives the result as the property of being a submanifold is a local condition (see Definition 3.1). □

We can now answer the question about intersections of submanifolds that we posed at the start of this section.

Corollary 4.11. *If two submanifolds M_1, M_2 in $\mathbb{R}^n$ intersect transversely ($M_1 \pitchfork M_2$) and $M_1 \cap M_2 \neq \emptyset$, then $M_1 \cap M_2$ is a submanifold of $\mathbb{R}^n$ of dimension* $\dim M_1 + \dim M_2 - n$.

Proof. As in the example we take $M = M_1$, $L = M_2$, $K = M_1 \cap M_2$, $N = \mathbb{R}^n$ and $f\colon M \to \mathbb{R}^n$ is the inclusion map. Then, as we saw in the example, $f \pitchfork L$. Applying Theorem 4.10, we see that $f^{-1}(L) = K = M_1 \cap M_2$ is a submanifold and

$$\dim M_1 - \dim M_1 \cap M_2 = \dim M - \dim f^{-1}(L) = \dim N - \dim L = n - \dim M_2.$$

Re-arranging gives the result. □

Example. If M_1, M_2 are submanifolds in $\mathbb{R}^n$ with $\dim M_1 + \dim M_2 = n$ and they intersect transversely then $M_1 \cap M_2$ consists of a (possibly empty) collection of isolated points.

Suppose we have two circles C_1, C_2 in the plane. It could happen that $C_1 \cap C_2$ consists of one point and the intersection is not transverse. However, if we translate C_2 slightly then the intersection will become empty (so trivially transverse) or two points and transverse. Thus we might expect that we can achieve transversality most of the time. This in fact turns out to be true: we shall omit the proof as it is a bit technically challenging, but a key ingredient is Sard's Theorem.

Theorem 4.12 (Transversality Theorem)**.** *Let L, M, N be submanifolds with $L \subseteq N$.*

(a) The set

$$\{f \in C^\infty(M, N) : f \pitchfork L\}$$

is dense in $C^\infty(M, N)$.

(b) Let L be closed and $K \subseteq M$ be compact. Then the set

$$\{f \in C^\infty(M, N) : f \pitchfork_K L\}$$

is open and dense in $C^\infty(M, N)$.

Remark. (Not examinable). We will not give the details of what being open in $C^\infty(M, N)$ means, but we can say that $f_1, f_2 \in C^\infty(M, N)$ are "close" if all the derivatives of f_1, f_2 are close to each other. Moreover, we can do better in (a) and say that the set is *residual*, i.e. contains the intersection of countably many open dense sets.

The application of the Transversality Theorem we care about is for intersections of submanifolds.

Corollary 4.13. *Let M_1, M_2 be submanifolds of $\mathbb{R}^n$ and let $\iota\colon M_1 \to \mathbb{R}^n$ be the inclusion map. Given any open set $U \ni \iota$ in $C^\infty(M_1, \mathbb{R}^n)$ there exists $f \in U$ such that f is an embedding and*

$$M_2 \pitchfork f(M_1).$$

Proof. Being an embedding is an open condition on $f \in C^\infty(M_1, \mathbb{R}^n)$. The Transversality Theorem says that the set of such f so that $f \pitchfork M_1$ is dense in $C^\infty(M_1, \mathbb{R}^n)$, so we are done. □

Remark. Corollary 4.13 says that any two submanifolds M_1, M_2 of dimensions k, l in $\mathbb{R}^n$ can be put in *general position*, i.e. we can perturb one of them (e.g. M_1) by an arbitrarily small amount so that they now intersect transversely. In particular, $M_1 \cap M_2$ is then a submanifold of $\mathbb{R}^n$ of dimension $k + l - n$, as one might expect.

Example. If M_1, M_2 are submanifolds of $\mathbb{R}^n$ of dimensions k, l such that $k + l < n$ then Corollary 4.13 says that, after a small perturbation, we can ensure that $M_1 \cap M_2 = \emptyset$. In particular, if we add enough dimensions to $\mathbb{R}^n$ we can always separate two submanifolds, which is intuitively clear: e.g. take two circles in the plane which intersect, but if we view them as in $\mathbb{R}^3$ we can just move one circle vertically out of the plane.

Example. One of the most interesting examples is if we take M_1, M_2 to be submanifolds in $\mathbb{R}^n$ of dimensions k, l such that $k + l = n$. In this case, the transversality result (Corollary 4.13) shows that (after a small perturbation) $M_1 \cap M_2$ is a 0-dimensional submanifold. If M_1, M_2 are compact, this will be a finite number of points.

We saw that the number of intersection points we get if M_1, M_2 are both circles in $\mathbb{R}^2$ intersecting transversely is 2 (or 0). If we instead take a circle and a sphere in $\mathbb{R}^3$ the answer is again 2 (or 0): we can see this by taking a planar circle that intersects the equator transversely. We shall see, by ideas in the next section, that the answer always has to be even.

Example. Another interesting case is when we take $M_1 = M_2 = M$. We then see that we can perturb $M_1 = M$ to $f(M)$ and talk about the *self-intersection* of M as the transverse intersection $M \cap f(M)$. If $\dim M = n$, M is compact and $M \subseteq \mathbb{R}^{2n}$, then this intersection is a finite number of points. Again, we will see that the number of such points has to be even.

Example. If we take two spheres in $\mathbb{R}^3$ that intersect transversely then if they intersect their intersection is a circle and thus a $(2+2-3=1)$-dimensional submanifold. More generally, a k-sphere and an l-sphere in $\mathbb{R}^n$ will meet in a $(k + l - n)$-sphere if $k + l \geq n$, if they meet at all: note that this still makes sense for $k + l = n$ since a 0-sphere is just two points.

5 Degree

We now want to move onto a different application of Sard's Theorem, namely to the *degree* of a map between submanifolds. For this we need to think about M, N being of the *same* dimension and M being compact, i.e. a closed and bounded subset of $\mathbb{R}^n$. The first thing we notice is the following.

Lemma 5.1. *Let M, N both be k-dimensional submanifolds and suppose that M is compact. Let $f\colon M \to N$ be smooth and let x be a regular value of f. Then $f^{-1}(x)$ is finite and there exists an open set $U \ni x$ such that, for all $y \in U$,*

$$|f^{-1}(x)| = |f^{-1}(y)|.$$

Proof. We know that $f^{-1}(x)$ is a 0-dimensional manifold by the Regular Value Theorem (Theorem 3.10). It is also a closed set (we saw this in the proof of Lemma 4.6) and M is compact (and thus bounded) so $f^{-1}(x)$ is compact (as it is closed and bounded), which means it is a finite number of isolated points $\{p_1, \dots, p_m\}$. Since $\mathrm{d}f_{p_j} : T_{p_j}M \to T_xN$ is an isomorphism for each p_j we know that there exists a small open set $W \ni x$ and disjoint open sets $V_j \ni p_j$ for $j = 1, \dots, m$ so that $f\colon V_j \to W$ is a diffeomorphism for all j. Then we can take

$$U = W \setminus f(M \setminus \bigcup_{j=1}^{m} V_j),$$

which is open. □

Now that we know the set $|f^{-1}(x)|$ is finite for a regular value as above and that it is locally constant, it is tempting to ask whether it is actually constant. This leads us to the following notion, where we recall that $N \subseteq \mathbb{R}^n$ is *connected* if there do not exist open sets U, V in $\mathbb{R}^n$ such that $U \cap N \neq \emptyset$, $V \cap N \neq \emptyset$ and

$$N \subseteq U \cup V \quad \text{and} \quad U \cap V \cap N = \emptyset.$$

Very informally, this means that N is made up of only "one piece".

Definition 5.2. Let M, N be k-dimensional submanifolds with M compact and N connected and let $f\colon M \to N$ be smooth. We define the *mod 2 degree* of f to be

$$\deg_2(f) = |f^{-1}(x)| \pmod 2,$$

where $x \in N$ is a regular value of f (which must exist by Sard's Theorem).

Remark. The reason to impose N connected is to avoid silly situations. For example, we could take L, M to be disjoint compact k-dimensional submanifolds in $\mathbb{R}^n$, let $N = L \cup M$ and let $f\colon M \to N$ be the identity map on M. Then N is not connected, f is smooth, and if we take $x \in N$ then the number of points in $f^{-1}(x)$ is 1 for $x \in M$ but 0 for $x \in L$, which means the mod 2 degree could never make sense.

As it stands, the notion of mod 2 degree might not be well-defined, but this is exactly what we now want to show. To do this we need to introduce an equivalence relation on maps known as *homotopy*.

Definition 5.3. Let M, N be submanifolds and let $f, g : M \to N$ be smooth. Then f is *smoothly homotopic* to g if and only if there exists a smooth map $h : M \times [0,1] \to N$ such that

$$h(p,0) = f(p) \quad \text{and} \quad h(p,1) = g(p)$$

for all $p \in M$. (In other words, we can interpolate smoothly from the map f to the map g.) The map h is then called a *smooth homotopy* (from f to g) and we write $f \sim_h g$ or just $f \sim g$. Note that being smoothly homotopic is an equivalence relation (see problem sheet).

Definition 5.3 leads us to the following lemma, which shows the relation between smooth homotopy and mod 2 degree. For the purposes of notation we let

$$\deg_2(f, x) = |f^{-1}(x)| \pmod 2.$$

Lemma 5.4. *Let M, N be k-dimensional submanifolds, let M be compact and let $f, g : M \to N$ be smoothly homotopic. If $x \in N$ is a regular value of f and g then*

$$\deg_2(f, x) = \deg_2(g, x).$$

Proof. Let $f \sim_h g$. Using Sard's Theorem (Theorem 4.1) and the fact that $|f^{-1}(x)|$ and $|g^{-1}(x)|$ are locally constant (Lemma 5.1) we can assume, possibly by perturbing x slightly, that x is also a regular value of h. Then by the extension of the Regular Value Theorem (Theorem 4.5), $h^{-1}(x)$ is a compact 1-dimensional submanifold with boundary given by

$$\partial h^{-1}(x) = f^{-1}(x) \times \{0\} \cup g^{-1}(x) \times \{1\}.$$

Since the number of boundary points of a compact 1-dimensional submanifold with boundary has to be even (Lemma 4.4) we see that

$$|f^{-1}(x)| + |g^{-1}(x)| = 0 \pmod 2,$$

which gives the result. □

We can now prove that the mod 2 degree is indeed well-defined.

Theorem 5.5. *Let M, N be k-dimensional submanifolds with M compact and N connected. Let $f \colon M \to N$ be smooth. The mod 2 degree $\deg_2 f$ of f is well-defined and if $f \sim g$ then $\deg_2 f = \deg_2 g$*

Proof. Suppose that $x, y \in N$ are two distinct regular values of f. Then I claim there exists a smooth map $H : N \times [0, 1] \to N$ such that, for all $p \in N$,

$$H(p, 0) = p, \quad H(x, 1) = y \quad \text{and} \quad H : N \times \{t\} \to N \text{ is a diffeomorphism for all } t \in [0, 1].$$

We will not prove this, but it uses the connectedness of N. Let $F \colon N \to N$ be given by $F = H|_{N \times \{1\}}$. The last condition on H means that it is a *smooth isotopy* from id to F, and we say that F is *smoothly isotopic* to the identity id.

We then define $g \colon M \to N$ by

$$g = F \circ f.$$

Then g is smooth and

$$g^{-1}(y) = f^{-1}(F^{-1}(y)) = f^{-1}(x).$$

Therefore, for all $a \in g^{-1}(y)$ we have by the Chain Rule that

$$\mathrm{d}g_a = \mathrm{d}(F \circ f)_a = \mathrm{d}F_{f(a)} \circ \mathrm{d}f_a$$

is surjective as $\mathrm{d}f_a$ is surjective (as $a \in f^{-1}(x)$ and x is a regular value of f) and $\mathrm{d}F_{f(a)}$ is an isomorphism by Lemma 2.5 as F is a diffeomorphism. Therefore, y is a regular value of g and $f \sim_h g$ where

$$h(p, t) = H(f(p), t).$$

Hence, by Lemma 5.4 and the fact that $g^{-1}(y) = f^{-1}(x)$, we have

$$\deg_2(f, y) = \deg_2(g, y) = \deg_2(f, x).$$

Thus $\deg_2 f$ is well-defined (it is independent of the choice of regular value x).

Now let $f \sim g$ (this is any g, not just the specific one above). By Sard's Theorem there exists $x \in N$ which is a regular value of both f and g. The result now follows by Lemma 5.4. □

Example. Let M be a compact connected submanifold of dimension at least 1.

Let $c \in M$ and let $f \colon M \to M$ be $f(p) = c$ for all $p \in M$. Then f has mod 2 degree 0, just by taking any $x \neq c$ in M (such x exists as $\dim M \geq 1$) which is then trivially a regular value.

The identity map $\mathrm{id} : M \to M$ has mod 2 degree 1 since $\mathrm{id}(p) = p$ for all $p \in M$. Hence, id is not smoothly homotopic to the constant map f.

We deduce that there does not exist a smooth map $f\colon B^{n+1} \to \mathcal{S}^n$ such that $f|_{\mathcal{S}^n} = \mathrm{id}$ (as we already saw in Lemma 4.6): if there were such a map then we could define $h\colon \mathcal{S}^n \times [0,1] \to \mathcal{S}^n$ by

$$h(p,t) = f(tp)$$

which would then give a smooth homotopy between the constant map where $c = f(0)$ and the identity map, which is a contradiction.

We can also refine the notion of degree to give an integer rather than an integer mod 2, but for that we need to introduce the important idea of *orientation*.

Definition 5.6. Let V be an n-dimensional real vector space. Suppose that we have two ordered bases $\{u_1, \dots, u_n\}$ and $\{v_1, \dots, v_n\}$ of V. Then there exists a unique matrix $A = (a_{ij}) \in \mathrm{GL}(n, \mathbb{R})$ such that

$$v_i = \sum_{j=1}^{n} a_{ij} u_j \quad \text{for } i = 1, \dots, n.$$

We say that the bases define the *same orientation* if $\det A > 0$ and the *opposite orientation* if $\det A < 0$.

The property of defining the same orientation is an equivalence relation on ordered bases, so an *orientation* on V is a choice of equivalence class of ordered bases. We see that V always has exactly two possible orientations (since either $\det A > 0$ or $\det A < 0$). A vector space endowed with an orientation is called *oriented*. An ordered basis of an oriented vector space will then be *positively oriented* if it lies in the orientation, and *negatively oriented* otherwise.

Remark. In practice we define an orientation on a vector space V by choosing one ordered basis for V and then the orientation is just its equivalence class.

Example. On $\mathbb{R}^n$ the standard ordered basis $\{e_1, \dots, e_n\}$ where e_i is 1 in the ith entry and 0 otherwise defines the standard orientation. If we permute the elements of the basis by $\sigma \in S_n$ then this new basis will be positively/negatively oriented when σ is even/odd.

We now have distinguished maps between oriented vector spaces.

Definition 5.7. Let $T : V \to W$ be an isomorphism between oriented vector spaces with positively oriented bases $\{v_1, \dots, v_n\}$ and $\{w_1, \dots, w_n\}$. Let $A \in \mathrm{GL}(n, \mathbb{R})$ be the matrix of T with respect to these bases. Then T is *orientation preserving* if $\det A > 0$ (i.e. $\operatorname{sign} \det A = +1$) and *orientation reversing* if $\det A < 0$ (i.e. $\operatorname{sign} \det A = -1$). This notion does not depend on the choice of positively oriented bases.

Example. The elements of the orthogonal group $\mathrm{O}(n)$ which are orientation preserving on $\mathbb{R}^n$ are precisely those in $\mathrm{SO}(n)$. This fits with our intuition that rotations preserve orientation whereas reflections are orientation reversing.

Now that we have orientations on vector spaces we want to define orientations on submanifolds in $\mathbb{R}^n$, including those with boundary as this will be important for our applications.

Definition 5.8. Let M be a k-dimensional submanifold (with boundary) in $\mathbb{R}^n$. An *orientation* on M is a continuous choice of orientation on each tangent space T_pM, given by an ordered basis $\{v_1, \dots, v_k\}$, for all $p \in M$ so that we can choose a diffeomorphism $\varphi : U \cap M \to V$, where $U \ni p$ is open in $\mathbb{R}^n$, V is open in $\mathbb{R}^k$ (or $\overline{H}^k$) and $\mathrm{d}\varphi_p(v_i) = e_i$ for $i = 1, \dots, k$ (i.e. $\mathrm{d}\varphi_p$ is orientation preserving).

If there exists an orientation on M we say it is *orientable* (and *non-orientable* if not) and if it is endowed with an orientation it is called *oriented*.

Remark. Strictly speaking we have not defined T_pM for k-dimensional submanifold $M \subseteq \mathbb{R}^n$ with $\partial M \neq \emptyset$ if $p \in \partial M$. If we take the tangent vectors to all curves in M at p this will form a half-space (a copy of H^k) in $\mathbb{R}^n$, but we define T_pM to be the vector space generated by vectors in that half-space. We therefore get three kinds of tangent vectors in T_pM: those which are tangent to ∂M, so lie in $T_p\partial M$; those which are "inward pointing", so there is a curve α in M with $\alpha'(0) \in T_pM$; and those which are "outward pointing".

Orientations may seem a bit unfamiliar, but we will now see that they are not so complicated in easy concrete examples.

Example. Take the circle $\mathcal{S}^1$ in $\mathbb{R}^2$. A choice of orientation is then either clockwise or anticlockwise around the circle, i.e. we choose the orientation on each line $T_p\mathcal{S}^1$ to point following either the clockwise or anticlockwise direction.

Remark. For a 0-dimensional submanifold, which is just a collection of points, we have a separate convention: we say that an orientation is just a choice of either $+1$ or -1 at each point.

Most submanifolds we will encounter are orientable but some are not, including some famous ones.

Example. The Möbius band M in $\mathbb{R}^3$ is not orientable: this makes sense since if we try to choose an orientation on the line T_pM then if we go around the band and come back to p "on the other side" then we will have to reverse the orientation on T_pM, so no orientation on M can exist.

Just like for oriented vector spaces we have distinguished maps between oriented submanifolds.

Definition 5.9. A diffeomorphism $f\colon M \to N$ between oriented k-dimensional submanifolds will have that $\mathrm{d}f_p : T_pM \to T_{f(p)}N$ is an isomorphism between oriented vector spaces for all $p \in M$, and so the notion of $\operatorname{sign}\det \mathrm{d}f_p$ is well-defined. We say that f is *orientation preserving* if

$$\operatorname{sign}\det \mathrm{d}f_p = +1$$

for all $p \in M$ and *orientation reversing* if

$$\operatorname{sign}\det \mathrm{d}f_p = -1$$

for all $p \in M$.

Remark. If $f : M \to N$ is a map between *connected* oriented k-dimensional submanifolds so that $\mathrm{d}f_p : T_pM \to T_{f(p)}N$ is an isomorphism for all $p \in M$, then $\operatorname{sign}\det \mathrm{d}f_p$ is independent of $p \in M$.

Example. Consider $-\operatorname{id} : \mathbb{R}^n \to \mathbb{R}^n$. This is a diffeomorphism and for all $p \in \mathbb{R}^n$

$$\det \mathrm{d}(-\operatorname{id})_p = \det(-\operatorname{id}) = (-1)^n.$$

Hence $-\operatorname{id}$ is orientation preserving if n is even and orientation reversing if n is odd.

One small issue that we have to deal with in the case where M has boundary is what the orientation on ∂M should be.

Definition 5.10. Let $M \subseteq \mathbb{R}^n$ be an oriented k-dimensional submanifold with boundary. Let $p \in \partial M$ and let $U \ni p$ be open in $\mathbb{R}^n$, let $V \subseteq H^k$ be open and let $\varphi : U \cap M \to V$ be an orientation preserving diffeomorphism (as given by the orientation on M). Choose an ordered basis $\{v_1, \dots, v_k\}$ for T_pM defining the orientation such that $\mathrm{d}\varphi_p(v_j) \in H^k \cap (\mathbb{R}^{k-1} \times \{0\})$ for $j \geq 2$ and $\mathrm{d}\varphi_p(v_1)$ is *outward pointing*, i.e. its inner product with $-e_k$ is positive (so it points out of the upper half-space $\overline{H}^k$). Then $\{v_2, \dots, v_n\}$ defines an orientation on $T_p\partial M$.

This procedure defines an orientation on ∂M called the *induced orientation*.

Example. The case of 1-dimensional manifolds with boundary is special. Let $[-1,1]$ be the straight line in $\mathbb{R}$ with the standard orientation, i.e. in the positive direction along the line. The orientation then points inward at -1, so we assign the orientation to be -1 there, and similarly outward at $+1$ so we assign $+1$ there.

More generally, if M is an oriented submanifold and $M \times [-1,1]$ is oriented in the obvious way from the orientations on M and $[-1,1]$, then the induced orientation on $M \times \{1\}$ agrees with the given one on M, but it is reversed on $M \times \{-1\}$.

Example. Using the standard orientation on the closed unit ball $\overline{B}^{n+1}$ coming from $\mathbb{R}^{n+1}$ we get an induced orientation on $\mathcal{S}^n = \partial B^{n+1}$. For $n = 1$, we see that the ordered basis we would take in the definition above at $(1,0)$ would just be $\{e_1, e_2\}$ (since e_1 points outward at $(1,0)$) so we would take $\{e_2\}$ to define the orientation at $(1,0)$ on $\mathcal{S}^1$, i.e. the induced orientation is anticlockwise.

Example. If we take $\mathcal{S}^n$ in $\mathbb{R}^{n+1}$ with the orientation as above, then $-\operatorname{id}$ preserves $\mathcal{S}^n$, is a diffeomorphism, and will have

$$\det \mathrm{d}(-\operatorname{id})_p = (-1)^{n+1}$$

for all $p \in \mathcal{S}^n$, so it is orientation preserving/reversing if n is odd/even.

Armed with all of these ideas about orientations, we can now extend our notion of degree of a map.

Definition 5.11. Let M, N be oriented k-dimensional submanifolds and let M be compact and N connected. Let $f\colon M \to N$ be smooth. The *degree* of f is

$$\deg f = \sum_{p \in f^{-1}(x)} \operatorname{sign} \det \mathrm{d}f_p$$

where $x \in N$ is a regular value of f. Notice that $\operatorname{sign} \det \mathrm{d}f_p$ makes sense since $\mathrm{d}f_p : T_pM \to T_xN$ will be an isomorphism between oriented vector spaces.

The proof that the degree is well-defined is now pretty much the same as before, except we have to work slightly harder at the first step in the argument, taking into account the orientations. This is achieved by the following result where, again, for notational purposes we write

$$\deg(f,x) = \sum_{p \in f^{-1}(x)} \operatorname{sign} \det \mathrm{d}f_p.$$

Lemma 5.12. *Let M, N be k-dimensional oriented submanifolds, let M be compact and let $f, g : M \to N$ be smoothly homotopic. If $x \in N$ is a regular value of f and g then*

$$\deg(f,x) = \deg(g,x).$$

Proof. Let $f \sim_h g$. As we saw before in the proof of Lemma 5.4, we may assume (by Sard's Theorem and by possibly perturbing x) that x is also a regular value of h. Again, by the extension (Theorem 4.5) of the Regular Value Theorem, $h^{-1}(x)$ is a compact 1-dimensional submanifold with boundary with

$$\partial h^{-1}(x) = f^{-1}(x) \times \{0\} \cup g^{-1}(x) \times \{1\}.$$

Therefore, $h^{-1}(x)$ is a finite disjoint union of arcs A_j, $j = 1, \dots, m$, such that

$$\partial A_j = \{(p_j, 0)\} \cup \{(q_j, 1)\}.$$

We can now define an orientation on A_j as follows. We have an orientation on $M \times [0,1]$ given by the orientation on M and the standard one on $[0,1]$. For $(p,t) \in A_j$ we choose an oriented basis $\{v_1(p,t), \dots, v_{k+1}(p,t)\}$ for $T_{(p,t)}(M \times [0,1])$ such that $v_1(p,t) \in T_{(p,t)}A_j$ and

$$\{\mathrm{d}h_{(p,t)}(v_2(p,t)), \dots, \mathrm{d}h_{(p,t)}(v_{k+1}(p,t))\} \subseteq T_{h(p,t)}N$$

is an oriented basis. In this way we see that $v_1(p,t)$ will point outward at one end of A_j, say at $(q_j,1)$, and inward at the other $(p_j,0)$.

Hence,

$$\operatorname{sign}\det \mathrm{d}(h|_{M\times\{0\}})_{(p_j,0)} + \operatorname{sign}\det \mathrm{d}(h|_{M\times\{1\}})_{(q_j,1)} = 0.$$

Recall that $h|_{M\times\{0\}} = f$ and $h|_{M\times\{1\}} = g$. We also recall the induced orientation on $M\times\{1\}$ will agree with the given one on M and be opposite on $M\times\{0\}$. We deduce that

$$-\operatorname{sign}\det \mathrm{d}f_{p_j} + \operatorname{sign}\det \mathrm{d}g_{q_j} = 0.$$

Taking the sum over j then gives the answer we wanted. □

Now the same proof for Theorem 5.5 as before for the mod 2 degree goes through to yield the following.

Theorem 5.13. *Let M,N be oriented k-dimensional submanifolds with M compact and N connected. Let $f\colon M\to N$ be smooth. The degree* $\deg f$ *of f is well-defined and if $f\sim g$ then* $\deg f = \deg g$.

Example. If $f\colon M\to N$ is not surjective then $\deg f = 0$, since any point in $N\setminus f(M)$ is trivially a regular value of f.

Example. The degree of $\mathrm{id}: M\to M$ is $+1$ but if $f\colon M\to M$ is an orientation reversing diffemorphism then $\deg f = -1$. Hence f is not smoothly homotopic to id. In particular, if n is even, $-\mathrm{id}$ on $\mathcal{S}^n$ is not smoothly homotopic to id.

We have a very nice application of this fact, for which we need a definition.

Definition 5.14. Let $M\subseteq\mathbb{R}^n$ be a submanifold. A *vector field* on M is a smooth map $X: M\to\mathbb{R}^n$ such that $X(p)\in T_pM$ for all $p\in M$.

Theorem 5.15. *The sphere $\mathcal{S}^n$ admits a nowhere vanishing vector field if and only if n is odd.*

Remark. This is a version of the "Hairy Ball Theorem", which says that every continuous vector field on $\mathcal{S}^{2n}$ has a zero. The idea is that no matter how you comb the hairs on the sphere there will always be a point which looks like a bald spot.

Proof. Suppose that $n=2k$ is even and there exists a nowhere vanishing vector field X on $\mathcal{S}^n$. Define $h:\mathcal{S}^n\times[0,1]\to\mathcal{S}^n$ by

$$h(p,t) = p\cos(\pi t) + \frac{X(p)}{\|X(p)\|}\sin(\pi t).$$

Since X is nowhere vanishing $X(p)\in\langle p\rangle^\perp$ for all p, h is well-defined and maps into $\mathcal{S}^n$. As X is smooth, h is smooth and for all $p\in\mathcal{S}^n$ we have

$$h(p,0) = p \quad\text{and}\quad h(p,1) = -p$$

Hence h is a smooth homotopy from id to $-\mathrm{id}$ on $\mathcal{S}^n$, which is a contradiction as $n=2k$ is even.

If $n=2k+1$ is odd, then we can define

$$X(x_1,\dots,x_{2k+2}) = (x_2,-x_1,\dots,x_{2k+2},-x_{2k+1})$$

which is a smooth map $X:\mathcal{S}^{2k+1}\to\mathbb{R}^{2k+2}$ such that $X(p)\in T_p\mathcal{S}^{2k+1}$ and has $\|X(p)\|=1$ for all p. □

Example. Let $f\colon\mathbb{C}\to\mathbb{C}$ be given by $f(z)=z^k$ for $k\in\mathbb{N}$, $k>0$. Then

$$\mathrm{d}f_z = kz^{k-1}$$

and so 1 is a regular value of f. As we saw right at the start of the course, $\det \mathrm{d}f_z > 0$ for all $z \neq 0$ as f is holomorphic, so since $f^{-1}(1)$ consists of k points, the kth roots of unity, we have that

$$\deg f = k.$$

This example leads us to prove a nice fact in our next example.

Example. Let us take a monomial of degree $k > 0$ on $\mathbb{C}$:

$$z^k + a_{k-1}z^{k-1} + \ldots + a_0.$$

This extends to a smooth map $f\colon \mathcal{S}^2 \to \mathcal{S}^2$ by identifying $\mathbb{C} \cup \{\infty\}$ with $\mathcal{S}^2$ and setting $f(\infty) = \infty$.

We can define a smooth map $h : \mathcal{S}^2 \times [0,1] \to \mathcal{S}^2$ by

$$h(z,t) = z^k + t(a_{k-1}z^{k-1} + \ldots + a_0) \quad \text{and} \quad h(\infty, t) = \infty.$$

Then

$$h(z,0) = z^k \quad \text{and} \quad h(z,1) = f(z).$$

By the degree theorem (Theorem 5.13) and the previous example we deduce that

$$\deg f = \deg(z \mapsto z^k) = k.$$

In particular this means that f is surjective (since $\deg f \neq 0$) and so $f(z) = w$ has a solution in $\mathbb{C}$ for every $w \in \mathbb{C}$: this is the *Fundamental Theorem of Algebra*.

We now have two more interesting interpretations of the degree.

Definition 5.16. Let $\gamma\colon \mathcal{S}^1 \to \mathbb{R}^2$ be a smooth closed curve and let $w \in \mathbb{R}^2$ such that $\gamma^{-1}(w) = \emptyset$. Consider the smooth map $f\colon \mathcal{S}^1 \to \mathcal{S}^1$ given by

$$f(z) = \frac{\gamma(z) - w}{\|\gamma(z) - w\|}.$$

(This is well-defined and smooth as $\gamma(z) \neq w$.) Then $\deg f$ is the *winding number* $\mathrm{wind}(\gamma, w)$ of γ around w. If we have a smooth homotopy h between smooth closed curves $\gamma_1, \gamma_2 : \mathcal{S}^1 \to \mathbb{R}^2$ such that $h(p,t) \neq w$ for all $p \in \mathcal{S}^1$ and $t \in [0,1]$, then we see that $\mathrm{wind}(\gamma_1, w) = \mathrm{wind}(\gamma_2, w)$.

This is related to complex analysis as one can see from the example below.

Example. View $\mathbb{R}^2 = \mathbb{C}$ and let $\gamma_k : \mathcal{S}^1 \to \mathbb{C}$ be given by $\gamma_k(z) = z^k$ for $k \in \mathbb{N}$, $k > 0$. We see that $\mathrm{wind}(\gamma_k, 0) = \deg f$ where $f(z) = z^k$. By our previous calculations we see that $\mathrm{wind}(\gamma_k, 0) = k$ for $k > 0$, which makes intuitive sense. In fact, one can see that $\mathrm{wind}(\gamma_k, 0) = k$ for any $k \in \mathbb{Z}$. This reminds us of the answer from *Cauchy's Residue Theorem* and indeed there is a relation, which one will see in Part C Differentiable Manifolds.

We now want instead to think about two closed curves in $\mathbb{R}^3$ and try to build a number like we had the winding number above.

Definition 5.17. Let $\gamma_1, \gamma_2 : \mathcal{S}^1 \to \mathbb{R}^3$ be smooth closed curves such that $\gamma_1(\mathcal{S}^1) \cap \gamma_2(\mathcal{S}^1) = \emptyset$. Consider the smooth map $f\colon \mathcal{S}^1 \times \mathcal{S}^1 \to \mathcal{S}^2$

$$f(p,q) = \frac{\gamma_1(p) - \gamma_2(q)}{\|\gamma_1(p) - \gamma_2(q)\|}.$$

(Again, this is well-defined and smooth as γ_1, γ_2 do not intersect.) Then $\deg f$ is the *linking number* $\mathrm{link}(\gamma_1, \gamma_2)$ of the curves γ_1, γ_2. Again, if we perform smooth homotopies on the curves γ_1, γ_2 such that the curves never intersect along the homotopies, then the linking number stays the same.

Example. If we take $\gamma_1, \gamma_2 : \mathcal{S}^1 \to \mathbb{R}^3$ given by

$$\gamma_1(x_1, x_2) = (x_1 + 2, x_2, 0) \quad \text{and} \quad \gamma_2(x_1, x_2) = (x_1 - 2, x_2, 0),$$

so we have unit circles in the plane $x_3 = 0$ with centres at $(2, 0, 0)$ and $(-2, 0, 0)$. Then f can never equal $(0, 0, 1)$ for example, and thus f is not surjective. Hence $\deg f = 0$ and the linking number of γ_1 and γ_2 is 0. This makes sense as we can just pull the two circles apart.

Example. A non-trivial example is given by $\gamma_1, \gamma_2 : \mathcal{S}^1 \to \mathbb{R}^3$ with

$$\gamma_1(x_1, x_2) = (x_1 - 1, x_2, 0) \quad \text{and} \quad \gamma_2(x_1, x_2) = (x_1, 0, x_2).$$

Then

$$f(x_1, x_2, y_1, y_2) = \frac{(x_1 - y_1 - 1, x_2, -y_2)}{\sqrt{(x_1 - y_1 - 1)^2 + x_2^2 + y_2^2}}.$$

We see that $f^{-1}(0, 1, 0) = \{(0, 1, -1, 0)\}$ and from this one can deduce that $|\deg f| = 1$ and thus $|\mathrm{link}(\gamma_1, \gamma_2)| = 1$. We can see visually that these two circles are linked.

Remark. (Not examinable.) We can construct more complicated versions of winding and linking numbers by thinking more generally about appropriate non-intersecting pairs of submanifolds: the winding number is for a non-intersecting pair of compact 1-dimensional submanifold (a curve) and a connected 0-dimensional submanifold (a point) in $\mathbb{R}^2$, whereas the linking number is for a pair of non-intersecting compact 1-dimensional submanifolds in $\mathbb{R}^3$. This is interesting, but we shall not pursue it.

Remark. Given two compact submanifolds M_1, M_2 in $\mathbb{R}^n$ with dimensions k, l such that $k + l = n$, we can (after a small perturbation so that $M_1 \pitchfork M_2$) count the number of intersection points $M_1 \cap M_2$ mod 2, and if M_1, M_2 are oriented we can count the intersection points with appropriate signs, measuring the relative orientations of M_1, M_2 at the intersection points. These numbers again turn out to be well-defined and smooth homotopy invariants. However, we can always translate M_1 (say) sufficiently far in $\mathbb{R}^n$ such that $M_1 \cap M_2 = \emptyset$ (since M_2 is bounded). Therefore, these numbers must always be zero. This gives an explanation as to why the number of transverse intersection points is always even.

6 Manifolds

We have so far only thought about smooth objects (i.e. submanifolds) in $\mathbb{R}^n$ but, as we said in the introduction, it is useful to have an abstract definition of what a smooth object is, independent of any ambient space. This leads us to the notion of *manifolds*.

Definition 6.1. A set M is an *n-dimensional manifold* if there exists a family $\mathcal{A} = \{(U_i, \varphi_i) : i \in I\}$ (for some indexing set I) where:

(a) $U_i \subseteq M$ for each i and $\bigcup_{i\in I} U_i = M$;

(b) $\varphi_i : U_i \to \mathbb{R}^n$ is a bijection onto an open set $\varphi_i(U_i)$;

(c) for all $i, j \in I$, $\varphi_i(U_i \cap U_j)$ is open in $\mathbb{R}^n$;

(d) whenever $U_i \cap U_j \neq \emptyset$ the *transition map*

$$\varphi_j \circ \varphi_i^{-1} : \varphi_i(U_i \cap U_j) \to \varphi_j(U_i \cap U_j)$$

is a diffeomorphism.

Each pair (U_i, φ_i) is called a *chart* and the set of all charts $\mathcal{A}$ is called an *atlas*.

Remarks.

(a) The idea is that we are describing the space M by pasting together little flat maps (given by the charts in the atlas), just like the regular atlas describes the Earth. In this way we can describe objects which can be complicated globally in a way which is easy locally. The transition maps ensure that as we change chart (like when we turn the page of the atlas) things fit together nicely.

(b) **(Not examinable.)** One could ask what happens if we have two different atlases $\mathcal{A}_1, \mathcal{A}_2$ on the same set M. Well, we say that they are *equivalent* if $\mathcal{A}_1 \cup \mathcal{A}_2$ is still an atlas. An equivalence class of atlases is then called a *smooth structure*. So, really, a manifold is a set with a smooth structure. However, we can just take one atlas to find out what the smooth structure is, as in the definition.

(c) **(Not examinable.)** Given an atlas we can define a topology on M by saying that V is open in M if and only if $\varphi_i(V \cap U_i)$ is open in $\mathbb{R}^n$ for all $i \in I$. Technically, we would then demand that this topology is Hausdorff (i.e. given any two distinct points p, q there are two disjoint open sets $U \ni p$ and $V \ni q$) and second countable (i.e. there is a countable collection of open sets $\{V_k : k \in \mathbb{N}\}$ such every open set is a union of some of these V_k). However, we will not be concerned with such topological issues in this course.

(d) Condition (d) in Definition 6.1 is exactly where the smoothness comes in. **(Not examinable.)** We can replace this condition by asking that the transition maps and their inverses are continuous, C^k, real analytic, or holomorphic: this will then define manifolds which are topological, C^k, real analytic or complex respectively.

Example. The simplest example of an n-dimensional manifold is $\mathbb{R}^n$ where we can take $\mathcal{A} = \{(\mathbb{R}^n, \mathrm{id})\}$. This also shows that any open set in $\mathbb{R}^n$ is an n-dimensional manifold.

Example. Since $\mathrm{GL}(n, \mathbb{R})$ is open in $\mathbb{R}^{n^2}$ it is an n^2-dimensional manifold.

We now show one of the simplest, but non-trivial, examples is an n-dimensional manifold: $\mathcal{S}^n$.

Example. Consider $\mathcal{S}^n$.

(a) Let $N = (0, \ldots, 0, 1)$ and $S = (0, \ldots, 0, -1)$ be the "North" and "South" poles. Let $U_N = \mathcal{S}^n \setminus \{N\}$ and $U_S = \mathcal{S}^n \setminus \{S\}$. These satisfy $U_N \cup U_S = \mathcal{S}^n$.

(b) Let $\varphi_N : U_N \to \mathbb{R}^n$ be given by

$$\varphi_N(x) = \frac{(x_1, \dots, x_n)}{1 - x_{n+1}}$$

and $\varphi_S : U_S \to \mathbb{R}^n$ be given by

$$\varphi_S(x) = \frac{(x_1, \dots, x_n)}{1 + x_{n+1}}.$$

(These are the stereographic projections.) We have explicit inverses:

$$\varphi_N^{-1}(y) = \left(\frac{2y_1}{1 + \|y\|^2}, \dots, \frac{2y_n}{1 + \|y\|^2}, \frac{\|y\|^2 - 1}{1 + \|y\|^2} \right)$$

and

$$\varphi_S^{-1}(y) = \left(\frac{2y_1}{1 + \|y\|^2}, \dots, \frac{2y_n}{1 + \|y\|^2}, \frac{1 - \|y\|^2}{1 + \|y\|^2} \right),$$

so φ_N, φ_S are clearly bijections onto the open set $\mathbb{R}^n$.

(c) We see that $U_N \cap U_S = \mathcal{S}^n \setminus \{N, S\}$ so $\varphi_N(U_N \cap U_S) = \mathbb{R}^n \setminus \{0\} = \varphi_S(U_N \cap U_S)$, which is open.

(d) We now calculate that $\varphi_S \circ \varphi_N^{-1} : \mathbb{R}^n \setminus \{0\} \to \mathbb{R}^n \setminus \{0\}$ is

$$\varphi_S \circ \varphi_N^{-1}(y) = \frac{y}{\|y\|^2},$$

which is a diffeomorphism because it is smooth, as $y \neq 0$, and it is its own inverse. (Essentially the transition map is the "inversion" map.)

We deduce that $\mathcal{S}^n$ is an n-dimensional manifold.

We know that $\mathcal{S}^n$ is an n-dimensional submanifold or $\mathbb{R}^{n+1}$, so it would be nice if all submanifolds were manifolds. This is precisely what we now show.

Proposition 6.2. *A k-dimensional submanifold M in $\mathbb{R}^n$ is a k-dimensional manifold.*

Proof. Let $p \in M$. By the generalisation of the Implicit Function Theorem (Theorem 3.3) there is an open set $U_p' \ni p$ in $\mathbb{R}^n$, an open set $V_p \ni 0$ in $\mathbb{R}^k$, an open set $W_p \subseteq \mathbb{R}^{n-k}$ and a diffeomorphism $\psi_p : V_p \times W_p \to U_p'$ with $\psi_p(0) = p$ such that

$$M \cap U_p' = \psi_p(V_p \times \{0\}).$$

We therefore let

$$U_p = M \cap U_p'$$

and $\varphi_p : U_p \to V_p \subseteq \mathbb{R}^k$ be given by

$$\varphi_p(\psi_p(x, 0)) = x$$

for all $x \in V_p$. Then $\cup_{p \in M} U_p = M$, $\varphi_p(U_p) = V_p$ is open and $\varphi_p : U_p \to V_p$ is a bijection. This defines $\mathcal{A} = \{(U_p, \varphi_p) : p \in M\}$.

We see that $U_p \cap U_q$ will be given by $U' \cap M$ for some (possibly empty) open set $U' \subseteq U_p'$ in $\mathbb{R}^n$ and therefore, since ψ is a diffeomorphism, $\varphi_p(U_p \cap U_q)$ will also be open in $\mathbb{R}^k$.

Finally, if p, q are such that $U_p \cap U_q \neq \emptyset$ then

$$\varphi_q \circ \varphi_p^{-1}(x) = \varphi_q(\psi_p(x, 0)) = \psi_q^{-1} \circ \psi_p(x, 0).$$

Since ψ_p, ψ_q are diffeomorphisms, the transition map is a diffeomorphism. □

Remark. (Not examinable). A deep result, known as the *(strong) Whitney Embedding Theorem*, states that every n-dimensional manifold can be realised as an embedded submanifold in $\mathbb{R}^{2n}$. This further motivates why at the start of the course we restricted to submanifolds as this is really every manifold. It is interesting to note that the proof of the Whitney Embedding Theorem relies on the transversality results we gave earlier.

We now want to say what it means for a map between manifolds to be smooth.

Definition 6.3. Let M be an m-dimensional manifold and N be an n-dimensional manifold. Then $f\colon M \to N$ is *smooth at* $p \in M$ if there exist coordinate charts (U, φ) on M with $p \in U$ and (V, ψ) on N with $f(p) \in V$ such that the map

$$\psi \circ f \circ \varphi^{-1} : \varphi(U) \subseteq \mathbb{R}^m \to \psi(V) \subseteq \mathbb{R}^n$$

is smooth where it is defined. We say f is *smooth* if it is smooth at all $p \in M$.

Remark. This definition makes sense precisely because the transition maps are diffeomorphisms. If we had two charts $(U_1, \varphi_1), (U_2, \varphi_2)$ at $p \in M$ and $(V_1, \psi_1), (V_2, \psi_2)$ at $f(p) \in N$, then

$$\psi_2 \circ f \circ \varphi_2^{-1} = (\psi_2 \circ \psi_1^{-1}) \circ (\psi_1 \circ f \circ \varphi_1^{-1}) \circ (\varphi_1 \circ \varphi_2^{-1}).$$

Hence $\psi_2 \circ f \circ \varphi_2^{-1}$ is smooth if and only if $\psi_1 \circ f \circ \varphi_1^{-1}$ is smooth.

Many natural maps are smooth.

Example. The identity map $\mathrm{id} : M \to M$ is smooth because given any chart (U, φ) on M we have that $\varphi \circ \mathrm{id} \circ \varphi^{-1} = \mathrm{id}$ on $\varphi(U)$, and the identity map on $\mathbb{R}^n$ is smooth.

Example. If $M \subseteq \mathbb{R}^m$ is a submanifold, then the restriction of any smooth map $f\colon \mathbb{R}^m \to \mathbb{R}^n$ to M is a smooth map in the sense above. If $N \subseteq \mathbb{R}^n$ is also a submanifold and the map $f\colon \mathbb{R}^m \to \mathbb{R}^n$ is smooth such that $f(M) \subseteq N$ then the restriction $f\colon M \to N$ is smooth.

Example. For any of the groups G of matrices we have shown are manifolds, the multiplication map $m : G \times G \to G$ given by $m(A, B) = AB$ and the inversion map $i : G \to G$ given by $i(A) = A^{-1}$ are smooth. This is what makes them *Lie groups*.

The left and right multiplication maps $L_A\colon G \to G$ and $R_A\colon G \to G$ given by $L_A(B) = AB$ and $R_A(B) = BA$ are smooth. Moreover, the determinant $\det : G \to \mathbb{R}$ and trace $\mathrm{tr} : G \to \mathbb{R}$ are smooth.

We are particularly interested in special types of smooth maps, as for submanifolds.

Definition 6.4. Let M, N be manifolds. A map $f\colon M \to N$ is a *diffeomorphism* if it is a smooth bijection with a smooth inverse. The manifolds M and N are said to be *diffeomorphic* if there exists a diffeomorphism between them.

A diffeomorphism is the natural notion of equivalence between manifolds, so diffeomorphic manifolds are "the same".

Example. The identity map $\mathrm{id} : M \to M$ is a diffeomorphism. If f, g are diffeomorphisms then so is $f \circ g$ and so is f^{-1}. Hence, the diffeomorphisms form a group which we write $\mathrm{Diff}(M)$.

Example. On the matrix groups we have seen are manifolds, the left and right multiplication maps L_A, R_A are diffeomorphisms (since their inverses are $L_{A^{-1}}, R_{A^{-1}}$).

7 Projective space

In this section we focus on very important examples of manifolds which play a key role in geometry, particularly algebraic geometry: *projective spaces*.

Definition 7.1. Let V be a finite-dimensional vector space over a field $\mathbb{F}$, which in practice we will take to be either $\mathbb{R}$ or $\mathbb{C}$. The *projective space* $\mathbb{P}(V)$ associated to V is the set of 1-dimensional subspaces of V. Equivalently, $\mathbb{P}(V)$ is the quotient $V \setminus \{0\}/\sim$ where $\sim$ is the equivalence relation

$$v \sim w \quad \Leftrightarrow \quad v = \lambda w \text{ for some } \lambda \in \mathbb{F}^* = \mathbb{F} \setminus \{0\}.$$

Another equivalent definition is that $\mathbb{P}(V)$ is the quotient of $V \setminus \{0\}$ under the action of $\mathbb{F}^*$ given by scalar multiplication.

Points in $\mathbb{P}(V)$ are denoted $[v]$ for $v \in V \setminus \{0\}$. We define the *dimension* of $\mathbb{P}(V)$ by

$$\dim \mathbb{P}(V) = \dim V - 1,$$

where we use the convention that the empty set has dimension -1. Therefore, if $\dim V = 2$ we say that $\mathbb{P}(V)$ is the projective line and if $\dim V = 3$ then $\mathbb{P}(V)$ is the projective plane.

There is an important special case that will be our main concern.

Definition 7.2. If $V = \mathbb{F}^n$ we write $\mathbb{P}(\mathbb{F}^{n+1}) = \mathbb{FP}^n$. Points on $\mathbb{FP}^n$ are represented by equivalence classes of points $(x_0, \ldots, x_n) \in \mathbb{F}^{n+1}$ where the coordinates are not all zero. We then use the notation

$$[x_0 : \ldots : x_n]$$

for these points and see that

$$[x_0 : \ldots : x_n] = [\lambda x_0 : \ldots : \lambda x_n]$$

for all $\lambda \in \mathbb{F}^*$. These are called *homogeneous coordinates* on $\mathbb{FP}^n$.

Example. Consider $\mathbb{FP}^1$ where points are given in homogeneous coordinates as $[x_0 : x_1]$. We see that if $x_0 = 0$ this gives one point $[0 : x_1] = [0 : 1]$ as $x_1 \neq 0$. If $x_0 \neq 0$ we can rescale it to be 1 and get the point $[x_0 : x_1] = [1 : t]$ for $t = x_1/x_0$ which can be any element of $\mathbb{F}$. In this way we have decomposed

$$\mathbb{FP}^1 = \mathbb{F} \sqcup \{[0 : 1]\}.$$

(Here $\sqcup$ means "disjoint union".) Moreover, $[1 : t] = [t^{-1} : 1]$ for $t \neq 0$, so we see that $[0 : 1]$ is the point we get by letting $t \to \infty$.

In particular, we see that $\mathbb{RP}^1 = \mathcal{S}^1$ (since we take $\mathbb{R}$ and add one point at infinity) and $\mathbb{CP}^1 = \mathcal{S}^2$ (since we take $\mathbb{C}$ and add a point at infinity).

We saw that $\mathbb{RP}^1$ and $\mathbb{CP}^1$ are manifolds of dimension 1 and 2: this is part of a general trend.

Lemma 7.3. *The real projective n-space $\mathbb{RP}^n$ is an n-dimensional manifold.*

Proof. For $i = 0, \ldots, n$ we define

$$U_i = \{[x_0 : \ldots : x_n] \in \mathbb{RP}^n \,:\, x_i \neq 0\}.$$

We then define $\varphi_i : U_i \to \mathbb{R}^n$ by

$$\varphi_i([x_0 : \ldots : x_n]) = \left(\frac{x_0}{x_i}, \ldots, \frac{x_{i-1}}{x_i}, \frac{x_{i+1}}{x_i}, \ldots, \frac{x_n}{x_i} \right).$$

It then follows that $\mathcal{A} = \{(U_i, \varphi_i) : i = 0, \ldots, n\}$ is an atlas on $\mathbb{RP}^n$. The details are left as an exercise. □

Lemma 7.4. *The complex projective n-space $\mathbb{CP}^n$ is a $2n$-dimensional manifold.*

Proof. The definition is the same as above except we replace $\mathbb{C}^n$ by $\mathbb{R}^n$, i.e. for $i = 0, \dots, n$ we let

$$U_i = \{[z_0 : \dots : z_n] \in \mathbb{CP}^n \,:\, z_i \neq 0\}$$

and $\varphi_i : U_i \to \mathbb{C}^n$ be

$$\varphi_i[z_0 : \dots : z_n]) = \left(\frac{z_0}{z_i}, \dots, \frac{z_{i-1}}{z_i}, \frac{z_{i+1}}{z_i}, \dots, \frac{z_n}{z_i}\right).$$

Again, this gives an atlas on $\mathbb{CP}^n$: the details are left as an exercise. □

Remark. (Not examinable). Actually, one sees that $\mathbb{CP}^n$ is an n-dimensional *complex manifold*, since the transition maps are holomorphic.

The constructions above suggest a general discussion.

Definition 7.5. On $\mathbb{FP}^n$ the subsets

$$U_i = \{[x_0 : \dots : x_n] \,:\, x_i \neq 0\}$$

for $i = 0, \dots, n$ are called *affine patches*. These are copies of $\mathbb{F}^n$ and we see that

$$\mathbb{FP}^n \setminus U_i = \{[x_0 : \dots : x_n] \,:\, x_i = 0\} \cong \mathbb{FP}^{n-1}.$$

Hence we can decompose

$$\mathbb{FP}^n = \mathbb{F}^n \sqcup \mathbb{FP}^{n-1}$$

in several different ways. In each case, we are adding in points at infinity to $\mathbb{F}^n$ to obtain $\mathbb{FP}^n$.

Projective spaces have distinguished subobjects which are also projective spaces.

Definition 7.6. If V is a finite-dimensional vector space over a field $\mathbb{F}$ and U is a subspace of V, then $\mathbb{P}(U) \subseteq \mathbb{P}(V)$ is called a *projective linear subspace*. If $\dim U = 2$ then $\mathbb{P}(U)$ is a *projective line* in $\mathbb{P}(V)$ and if $\dim U = \dim V - 1$ then $\mathbb{P}(U)$ is called a *hyperplane*.

We now want to think about intersections of projective linear subspaces in $\mathbb{P}(V)$. The first is familiar just from usual Euclidean geometry.

Lemma 7.7. *Given any two distinct points in $\mathbb{P}(V)$ there is a unique projective line in $\mathbb{P}(V)$ passing through the points.*

Proof. If the two distinct points are $[v]$, $[w]$ for v, w non-zero vectors in V, then v, w must be linearly independent. Therefore the projective line is $\mathbb{P}\langle v, w\rangle$. □

However, the next result is definitely false in the Euclidean case.

Lemma 7.8. *In the projective plane any two distinct projective lines meet in a unique point.*

Proof. The projective plane is $\mathbb{P}(V)$ for a 3-dimensional vector space V. The lines are given by $\mathbb{P}(U_1)$, $\mathbb{P}(U_2)$ for 2-dimensional subspaces U_1, U_2 of V. The dimension formula says that

$$\dim(U_1 + U_2) + \dim(U_1 \cap U_2) = \dim U_1 + \dim U_2.$$

Since U_1, U_2 define distinct lines $U_1 + U_2$ contains both U_1, U_2 and has strictly larger dimension, so must equal V. The dimension formula then implies that

$$\dim(U_1 \cap U_2) = 1.$$

We then see that $\mathbb{P}(U_1 \cap U_2)$ is the unique intersection point. □

We can generalise this as follows, with the proof left as an exercise.

Theorem 7.9. *Let $L_1 = \mathbb{P}(U_1), L_2 = \mathbb{P}(U_2)$ be projective linear subspaces of $\mathbb{P}(V)$. We define the projective span of L_1, L_2 by*

$$\langle L_1, L_2 \rangle = \mathbb{P}(U_1 + U_2).$$

Then

$$\dim(L_1 \cap L_2) = \dim(L_1) + \dim(L_2) - \dim\langle L_1, L_2 \rangle.$$

We can also relate projective subspaces, and other subobjects in projective space, to *homogeneous polynomials*.

Definition 7.10. Let f be a homogeneous polynomial on $\mathbb{F}^{n+1}$, i.e.

$$f(\lambda x_0, \ldots, \lambda x_n) = \lambda^d f(x_0, \ldots, x_n)$$

for all $\lambda \in \mathbb{F}^*$ and all $(x_0, \ldots, x_n) \in \mathbb{F}^{n+1}$. Then the equation

$$f[x_0 : \ldots : x_n] = 0$$

is well-defined on $\mathbb{FP}^n$.

A subset $M \subseteq \mathbb{FP}^n$ is a *projective algebraic variety* if there exist homogeneous polynomials $f_1, \ldots, f_k$ on $\mathbb{F}^{n+1}$ such that

$$M = \{[x] \in \mathbb{FP}^n \,:\, f_1(x) = \ldots = f_k(x) = 0\}.$$

Example. If $f \neq 0$ is a homogeneous polynomial of degree 1 on $\mathbb{F}^{n+1}$ then $f(x) = 0$ defines a subspace U of $\mathbb{F}^{n+1}$ of dimension n. Hence the projective algebraic variety defined by f is just $\mathbb{P}(U)$, which is a hyperplane.

It is perhaps natural to ask when a projective algebraic variety over $\mathbb{R}$ or $\mathbb{C}$ is a manifold. This can be achieved by the following result.

Proposition 7.11. *Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$. Let $M \subseteq \mathbb{FP}^n$ be a projective algebraic variety defined by homogeneous polynomials $f_1, \ldots, f_k$ on $\mathbb{F}^{n+1}$. If the Jacobian matrix*

$$\left(\frac{\partial f_i}{\partial x_j}\right)$$

has rank k at every $x \in \mathbb{F}^{n+1} \setminus \{0\}$ such that $[x] \in M$, then M is an $(n-k)$-dimensional manifold if $\mathbb{F} = \mathbb{R}$ or a $2(n-k)$-dimensional manifold if $\mathbb{F} = \mathbb{C}$.

Proof. **(Not examinable.)** Define $f : \mathbb{FP}^n \to \mathbb{F}^k$ by $f = (f_1, \ldots, f_k)$. Then f is smooth (since it is defined by polynomials) and $M = f^{-1}(0)$. The assumption on the Jacobian matrix is exactly that 0 is a regular value of f. A version of the Regular Value Theorem for maps between manifolds then shows that M is a manifold. □

Example. Suppose that $M \subseteq \mathbb{FP}^n$ is a projective algebraic variety defined by a single homogeneous polynomial f of degree $d \geq 1$ on $\mathbb{F}^{n+1}$, i.e.

$$M = \{[x] \in \mathbb{F}^{n+1} \,:\, f(x) = 0\}.$$

In this case, M is called a *divisor* If $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$, the condition for M to be a manifold of (real or complex) dimension $n-1$ is that the polynomials

$$f, \frac{\partial f}{\partial x_0}, \ldots, \frac{\partial f}{\partial x_n}$$

do *not* have a common zero in $\mathbb{F}^{n+1} \setminus \{0\}$. For example, if we take

$$f(z_0, \ldots, z_n) = z_0^d + \ldots + z_n^d$$

we see that

$$\frac{\partial f}{\partial z_j} = dz_j^{d-1} = 0$$

if and only if $z_j = 0$, so the only common zero of f and $\frac{\partial f}{\partial z_j}$ for all j is the origin. We deduce that M is a manifold, often known as a *Fermat hypersurface*.

Remark. (Not examinable.) The special case of divisors in $\mathbb{CP}^2$ are examples of *algebraic curves*, which are the subject of a Part B course. A fundamental result in that course will concern intersections of pairs of algebraic curves. In general, the study of intersections of projective algebraic varieties is an important part of algebraic geometry.

We now want to talk about transformations of projective space. Clearly if $T : V \to W$ is a linear map then

$$T(\lambda v) = \lambda T v$$

for $v \in V$ and $\lambda \in \mathbb{F}^*$, which means that

$$[Tv] = [T(\lambda v)] \in \mathbb{P}(W)$$

provided $Tv \neq 0$. Hence we need $T : V \to W$ to be injective.

Definition 7.12. Let V, W be finite-dimensional vector spaces over $\mathbb{F}$ and let $T : V \to W$ be an injective linear map. The associated *projective linear transformation* $\tau : \mathbb{P}(V) \to \mathbb{P}(W)$ is defined by

$$\tau[v] = [Tv]$$

for $v \in V \setminus \{0\}$.

A particularly interesting case is when $V = W$. In this case $T : V \to V$ lies in $\mathrm{GL}(V)$, the invertible linear maps from V to V, and the map $T \mapsto \tau$ is a homomorphism ϕ from $\mathrm{GL}(V)$ to the group of projection linear transformations of V. It is clear that

$$\ker \phi = \{\lambda I \,:\, \lambda \in \mathbb{F}^*\}.$$

By the First Isomorphism Theorem (for groups) we have the following definition.

Definition 7.13. The group $\mathrm{PGL}(V)$ of projective linear transformations of a finite-dimensional vector space V over $\mathbb{F}$ is given by the quotient group:

$$\mathrm{PGL}(V) = \mathrm{GL}(V)/\{\lambda I \,:\, \lambda \in \mathbb{F}^*\}.$$

We write $\mathrm{PGL}(n+1, \mathbb{F}) = \mathrm{PGL}(\mathbb{F}^{n+1})$, the *projective linear group*.

To close this section we relate the projective linear group in the case $n = 2$ to maps which we already know very well.

Example. Let $T : \mathbb{F}^2 \to \mathbb{F}^2$ be an invertible linear map. Then we can write it as

$$T(x_0, x_1) = (ax_0 + bx_1, cx_0 + dx_1)$$

where $ad - bc \neq 0$. The associated projective linear transformation on $\mathbb{FP}^1$ is then

$$\tau : [x_0 : x_1] \to [ax_0 + bx_1 : cx_0 + dx_1].$$

If we consider the affine patch where $x_1 \neq 0$ then

$$\tau : \left[\frac{x_0}{x_1} : 1\right] \mapsto \left[\frac{ax_0 + bx_1}{cx_0 + dy_1} : 1\right].$$

In terms of $z = \frac{x_0}{x_1}$ on this affine patch τ is just

$$\tau : z \mapsto \frac{az+b}{cz+d},$$

which is a *Möbius transformation* when $\mathbb{F} = \mathbb{C}$.

In particular, what we have shown is that $\mathrm{PGL}(2,\mathbb{C})$ is nothing other than the Möbius group. We should also notice that we can always rescale the representative of a class in $\mathrm{PGL}(2,\mathbb{C})$ to have determinant 1, i.e. lie in $\mathrm{SL}(2,\mathbb{C})$, and the remaining freedom in the equivalence class is to multiply by elements in $\{\lambda I : \lambda \in \mathbb{C}, \det(\lambda I) = 1\} = \{\pm I\}$. Hence

$$\mathrm{PGL}(2,\mathbb{C}) = \mathrm{PSL}(2,\mathbb{C}) = \mathrm{SL}(2,\mathbb{C})/\{\pm I\};$$

that is, matrices in $M_2(\mathbb{C})$ with determinant 1, up to sign. Note also that to compose Möbius transformations you can just multiply matrices in $\mathrm{SL}(2,\mathbb{C})$ representing the classes in $\mathrm{PSL}(2,\mathbb{C})$.

We conclude this section by discussing the notion of *projective duality*. For this we first recall some ideas from Part A Linear Algebra. If V is a finite-dimensional vector space over $\mathbb{F}$, then its *dual space*

$$V^* = \mathrm{Hom}(V,\mathbb{F}) = \{f : V \to \mathbb{F} \,:\, \mathbb{F} \text{ is linear}\}.$$

Then $\dim V = \dim V^*$ and V, V^* are isomorphic *provided* we choose a basis for V. However, if we take the dual of the dual space V^{**} (often called the double dual), we have a *canonical isomorphism* between V and V^{**}:

$$\phi\colon V \to V^{**}, \qquad \phi(v)(f) = f(v) \quad \text{for } f \in V^*.$$

In this way, we can identify V^{**} with V and drop the notation ϕ.

Another important construction is the *annihilator* $U^\circ \leqslant V^*$ of a subspace $U \leqslant V$:

$$U^\circ = \{f \in V^* \,:\, f(u) = 0 \text{ for all } u \in U.$$

Recall that

$$\dim U^\circ = \dim V - \dim U$$

and

$$U^{\circ\circ} = U.$$

Hence, the map $U \mapsto U^\circ$ defines a bijection between subspaces of V and subspaces of V^*. Moreover, if $U_1, U_2 \leqslant V$,

$$U_1 \leqslant U_2 \quad \Rightarrow \quad U_2^\circ \leqslant U_1^\circ$$

so the bijection is *inclusion-reversing*. This leads us to the following definition.

Definition 7.14. The *projective duality correspondence* Φ between linear subspaces of $\mathbb{P}(V)$ and linear subspaces of $\mathbb{P}(V^*)$ is the inclusion-reversing bijection given by

$$\Phi(\mathbb{P}(U)) = \mathbb{P}(U^\circ)$$

for $U \leqslant V$. Note that $\dim \mathbb{P}(V) = n$ and $\dim \mathbb{P}(U) = m$ implies that

$$\dim \mathbb{P}(U^\circ) = \dim U^\circ - 1 = \dim V - \dim U - 1 = (n+1) - (m+1) - 1 = n - m - 1.$$

Example. Let $\dim V = n+1$. Then if $[f] \in \mathbb{P}(V)^*$ we have that $\mathbb{P}(U) = \Phi^{-1}[f] \in \mathbb{P}(V)$ has dimension $n - 0 - 1 = n - 1$, i.e. $\mathbb{P}(U)$ is a hyperplane. Using the isomorphism ϕ and the fact that $[f] = \mathbb{P}(\langle f \rangle)$, we see that $U = \ker f$:

$$U = \{x \in V \,:\, f(x) = 0\}.$$

Conversely hyperplanes in $\mathbb{P}(V)$ correspond to points in $\mathbb{P}(V^*)$ under Φ. Concretely, if $V = \mathbb{F}^n$ and $f = [a_0 : \ldots : a_n]$ in homogeneous coordinates we see that

$$U = \{[x_0 : \ldots : x_n] \in \mathbb{FP}^n \,:\, a_0x_0 + \ldots + a_nx_n = 0\}$$

in the correspondence between $[f]$ and U.

Remark. It can be useful to use the following identities from Part A Linear Algebra for subspaces U_1, U_2 in V:

$$(U_1 + U_2)^\circ = U_1^\circ \cap U_2^\circ \quad \text{and} \quad (U_1 \cap U_2)^\circ = U_1^\circ + U_2^\circ.$$

For example, if $v, w \in V \setminus \{0\}$ then

$$\langle v, w\rangle^\circ = \langle v\rangle^\circ \cap \langle w\rangle^\circ.$$

Example. In the projective plane $\mathbb{P}(V)$ (for $\dim V = 3$) projective duality exchanges points and projective lines. Let $[v], [w]$ be two distinct points on the projective line $\mathbb{P}\langle v, w, \rangle$. We see that $\dim\langle v\rangle^\circ = 2$ so $\mathbb{P}(\langle v\rangle^\circ)$ is a projective line and, by the remark above,

$$\dim\langle v, w\rangle^\circ = \dim\langle v\rangle^\circ \cap \langle w\rangle^\circ = \dim\langle v\rangle^\circ + \dim\langle w\rangle^\circ - 3 = 1,$$

so $\mathbb{P}\langle v, w\rangle^\circ$ is a point. Hence, the projective lines $\mathbb{P}(\langle v\rangle^\circ), \mathbb{P}(\langle w\rangle^\circ)$ meet at the point $\mathbb{P}(\langle v, w\rangle^\circ)$. By extending this, we see that collinear points in $\mathbb{FP}^2$ (i.e. points lying along a line) corresponds under duality to a set of concurrent lines.

A consequence of this discussion is that $\mathbb{P}\langle v\rangle^\circ \subseteq \mathbb{P}(V^*)$ can be viewed as the set of points parametrising lines through $[v]$ in $\mathbb{P}(V)$.

Remark. The example above shows that Lemma 7.7 and Lemma 7.8 are duals of each other. This is a general phenomenon: for every result in projective geometry there will be a dual version, and we need only prove one to deduce the other.

Remark. Projective geometry can be used to tackle classical problems in plane geometry: we will see such applications on a problem sheet.

8 Hyperbolic space

In this final section of the course, we now want to introduce a new kind of geometry: *hyperbolic geometry*. Its study represented one of the most important developments in geometry and continues to play a key role in many aspects of current mathematics. It can be defined in all dimensions, but in this course we will restrict ourselves to 2 dimensions.

We will see a number of different ways of thinking of hyperbolic geometry, but we will start with an analogue of the construction of the sphere. Remember that on $\mathbb{R}^3$ we have the inner product

$$\langle (u_1, u_2, u_3), (v_1, v_2, v_3) \rangle = u_1 v_1 + u_2 v_2 + u_3 v_3.$$

The unit 2-sphere is then given by the points where $\langle x, x \rangle = 1$.

We can think of $\mathbb{R}^3 = \mathbb{R} \times \mathbb{R}^2$ and give coordinates (x_0, x_1, x_2) where we think of x_0 as "time": this is relevant for relativity in physics. We can then define a new "inner product" as follows.

Definition 8.1. We define the Lorentz inner product on $\mathbb{R} \times \mathbb{R}^2$ by

$$\langle (u_0, u_1, u_2), (v_0, v_1, v_2) \rangle_L = u_0 v_0 - u_1 v_1 - u_2 v_2.$$

We often denote $\mathbb{R} \times \mathbb{R}^2$ with the Lorentz inner product as $\mathbb{R}^{1,2}$. Let $\mathcal{H}^2 \subseteq \mathbb{R}^{1,2}$ be given by

$$\mathcal{H}^2 = \{x = (x_0, x_1, x_2) \in \mathbb{R}^{1,2} \,:\, \langle x, x \rangle_L = x_0^2 - x_1^2 - x_2^2 = 1,\, x_0 > 0\}.$$

The reason for $x_0 > 0$ is because the hyperboloid defined by $\langle x, x \rangle_L = 1$ has two components.

We have already seen that $\mathcal{H}^2$ is a 2-dimensional manifold diffeomorphic to $\mathbb{R}^2$. Note that any $x \in \mathcal{H}^2$ can be written as $x = (\cosh t, \sinh t \cos\theta, \sinh t \sin\theta)$.

Remark. It is clear how to define $\mathcal{H}^n$ in all dimensions, by adding more "space directions" $x_1, x_2, \ldots, x_n$ to give $\mathcal{H}^n \subseteq \mathbb{R}^{1,n}$.

We now want to define the hyperbolic geometry on $\mathcal{H}^2$. The easiest way to think about it is to think about what the shortest paths between two points in $\mathcal{H}^2$ look like: these are called *geodesics*.

Definition 8.2. A *Lorentz plane* through the origin in $\mathbb{R}^{1,2}$ is one containing a vector x such that $\langle x, x \rangle_L > 0$. Lorentz planes are those which can meet $\mathcal{H}^2$.

A *geodesic hyperbola* in $\mathcal{H}^2$ is the intersection of a Lorentz plane through the origin in $\mathbb{R}^{1,2}$ with $\mathcal{H}^2$. The arcs of geodesic hyperbolae are called *geodesics*.

Example. The plane $\Pi = \{(x_0, x_1, x_2) \,:\, x_2 = 0\}$ contains the vector $x = (1, 0, 0)$, which satisfies $\langle x, x \rangle_L = 1 > 0$. Hence Π is a Lorentz plane.

This gives a geodesic hyperbola as

$$\Pi \cap \mathcal{H}^2 = \{(x_0, x_1, 0) \,:\, x_0^2 - x_1^2 = 1,\, x_0 > 0\} = \{(\cosh t, \sinh t, 0) \,:\, t \in \mathbb{R}\}.$$

In other words, we can define it as $x_2 = 0$ on $\mathcal{H}^2$.

We have the following easy fact which reminds us of both Euclidean and projection geometry.

Lemma 8.3. *Given any two distinct points in $\mathcal{H}^2$ there is a unique geodesic between them.*

Proof. Given any two points in $\mathcal{H}^2$, there is a unique Lorentz plane through the origin containing the two points. □

Example. Suppose we have two geodesic hyperbolae defined by distinct Lorentz planes Π_1, Π_2. Then, by the dimension formula for vector spaces, we know that $\Pi_1 \cap \Pi_2 = \mathrm{Span}\{x\}$ and there are three cases.

(a) $\langle x,x\rangle_L < 0$ (x is 'space-like'): the hyperbolae are disjoint and diverge.

If we let $\Pi_1 = \text{Span}\{(1,0,0),(0,1,0)\}$ and $\Pi_2 = \text{Span}\{(2,0,1),(0,1,0)\}$ then $\Pi_1 \cap \Pi_2$ is spanned by $x = (0,1,0)$, so $\langle x,x\rangle_L = -1$. Then the hyperbolae are $\{(\cosh t, \sinh t, 0)\}$, corresponding to $x_2 = 0$, and the one given by $x_0 = 2x_2$:

$$\{(2x_0,x_1,x_0) : 4x_0^2 - x_1^2 - x_0^2 = 3x_0^2 - x_1^2 = 1,\ x_0 > 0\} = \{(\frac{2}{\sqrt{3}}\cosh t, \sinh t, \frac{1}{\sqrt{3}}\cosh t)\},$$

so they diverge (the x_2 component of the second goes to infinity whilst the other remains 0).

(b) $\langle x,x\rangle_L > 0$ (x is 'time-like'): the hyperbolae intersect at one point.

If we let $\Pi_1 = \text{Span}\{(1,0,0),(0,1,0)\}$ and $\Pi_2 = \{(1,0,0),(0,0,1)\}$ then $\Pi_1 \cap \Pi_2$ is spanned by $x = (1,0,0)$ so $\langle x,x\rangle_L = 1$ and the hyperbolae are $x_2 = 0$: $\{(\cosh t, \sinh t, 0) : t \in \mathbb{R}\}$ and $x_1 = 0$: $\{(\cosh t, 0, \sinh t) : t \in \mathbb{R}\}$ which only meet at $(1,0,0)$.

(c) $\langle x,x\rangle_L = 0$ (x is 'null'): the hyperbolae are disjoint but approach each other at infinity, and are called ultraparallel.

If we let $\Pi_1 = \text{Span}\{(2,0,1),(1,1,0)\}$ and $\Pi_2 = \text{Span}\{(2,0,-1),(1,1,0)\}$ then $\Pi_1 \cap \Pi_2$ is spanned by $x = (1,1,0)$ so $\langle x,x\rangle_L = 0$ and the hyperbolae are given by $x_0 = x_1 \pm 2x_2$:

$$\{(2x_0 + x_1, x_1, \pm x_0) : (2x_0+x_1)^2 - x_1^2 - x_0^2 = x_0(3x_0 + 4x_1) = 1,\ 2x_0 + x_1 > 0\}$$

which are disjoint because we cannot have $x_0 = 0$ and we see that as $x_1 \to \infty$ we must have $x_0 \to 0$, so they approach each other.

The Lorentz inner product looks bad because it can be positive, negative and even zero. However, on $\mathcal{H}^2$ it is well-behaved as we now see.

Lemma 8.4. *Given any two points x, y in $\mathcal{H}^2$ we can choose coordinates, preserving the Lorentz inner product, on $\mathbb{R}^{1,2}$ so that $x = (1,0,0)$ and $y = (\cosh t, \sinh t, 0)$.*

Hence $\langle x,y\rangle_L \geq 1$ with equality if and only if $x = y$.

Proof. Let $x, y \in \mathcal{H}^2$. Since $\langle x,x\rangle_L = 1$ we can choose a Lorentz orthonormal basis $\{e_0, e_1, e_2\}$ for $\mathbb{R}^{1,2}$ with $e_0 = x$. We can then rotate the plane $\langle e_1, e_2\rangle$ so that the coefficient of y in the e_2 direction is zero (and this preserves $\mathcal{H}^2$). In these coordinates $x = (1,0,0)$ and $y = (y_0, y_1, 0)$. Since $y \in \mathcal{H}^2$, $y = (\cosh t, \sinh t, 0)$.

Hence $\langle x,y\rangle_L = \cosh t \geq 1$ and $\langle x,y\rangle_L = 1$ if and only if $t = 0$, which is equivalent to $x = y$. □

Remark. The fact that one can choose good coordinates as in Lemma 8.4 is a useful tool.

We now want to introduce the notion of distance that defines the hyperbolic geometry on $\mathcal{H}^2$.

Definition 8.5. The *hyperbolic distance* $d(x,y)$ between $x, y \in \mathcal{H}^2$ is given by

$$d(x,y) = \cosh^{-1}(\langle x,y\rangle_L),$$

where $\cosh^{-1}$ takes values in $[0,\infty)$. Lemma 8.4 shows that this is well-defined and we will see on a problem sheet that $(\mathcal{H}^2, d)$ is a metric space which we call the *hyperbolic 2-space*. (It is clear how to generalise the definition to define the hyperbolic n-space $(\mathcal{H}^n, d)$.)

Example. Like $\mathbb{R}^2$, but unlike $\mathcal{S}^2$, distances in $(\mathcal{H}^2, d)$ can be arbitrarily large. For example, if $x = (1,0,0)$ and $y = (\cosh t, \sinh t, 0)$ then $d(x,y) = t$.

We now discuss very briefly isometries of the hyperboloid model as we will be more focussed on isometries for the other models. Let us briefly recall the definition of isometry.

Definition 8.6. An *isometry* between metric spaces (M_1, d_1) and (M_2, d_2) is a map $T : (M_1, d_1) \to (M_2, d_2)$ such that $d(T(x), T(y)) = d(x, y)$ for all x, y. (Note this forces T to be injective.) The set of bijective isometries from a metric space (M, d) to itself forms a group $\mathrm{Isom}(M)$ called the *isometry group*.

The group we are interested for hyperbolic geometry is the following.

Definition 8.7. Let

$$g = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

We define $\mathrm{O}^+(1,2)$ to be the set of $A = (a_{ij}) \in M_3(\mathbb{R})$ (where $i, j \in \{0, 1, 2\}$) such that

$$A^{\mathrm{T}} g A = g$$

and $a_{00} > 0$. This forms a group of *orthochronous Lorentz transformations* of $\mathbb{R}^{1,2}$. (Orthochronous is because $a_{00} > 0$, which means that "time" points in the same direction after the Lorentz transformation.)

We may also define

$$\mathrm{SO}^+(1,2) = \{A \in \mathrm{O}^+(1,2) \,:\, \det A = 1\},$$

the group of *proper orthochronous Lorentz transformations* of $\mathbb{R}^{1,2}$.

We see that the Lorentz group contains a familiar subgroup.

Example. We see that $\mathrm{O}(2)$ is a subgroup of $\mathrm{O}^+(1,2)$, since if $B \in \mathrm{O}(2)$ then

$$A = \begin{pmatrix} 1 & 0 \\ 0 & B \end{pmatrix}$$

lies in $\mathrm{O}^+(1,2)$. Hence rotations in the plane $x_0 = 0$, and reflections in planes of the form $\mathbb{R} \times \ell$ where ℓ is a line through 0 which lies in the plane $x_0 = 0$, both lie in $\mathrm{O}^+(1,2)$.

The following theorem is proved just like in the case of $\mathbb{R}^2$ so we will omit it.

Theorem 8.8. *Any isometry of $\mathcal{H}^2$ is given by $T(x) = Ax$ where $A \in \mathrm{O}^+(1,2)$ and sends geodesics to geodesics. Hence $\mathrm{Isom}(\mathcal{H}^2) = \mathrm{O}^+(1,2)$ and the group of orientation-preserving isometries $\mathrm{Isom}^+(\mathcal{H}^2) = \mathrm{SO}^+(1,2)$.*

Example. We see that

$$\begin{pmatrix} \cosh s & \sinh s & 0 \\ \sinh s & \cosh s & 0 \\ 0 & 0 & \pm 1 \end{pmatrix} \in \mathrm{O}^+(1,2).$$

These are examples of what are sometimes known as *Lorentz translations/glides* (for $+1$ or -1)

Remark. (Not examinable). We can generate all of $\mathrm{Isom}(\mathcal{H}^2)$ using elements in $\mathrm{O}(2)$ together with Lorentz translations and glides.

I want to continue this section by giving two more ways of thinking about hyperbolic 2-space, which are useful for different purposes.

(Not examinable). Consider the map $f \colon \mathcal{H}^2 \to \mathbb{C}$ given by

$$f(x_0, x_1, x_2) = \frac{-x_2 + i}{x_0 - x_1}.$$

Since $x_0^2 - x_1^2 - x_2^2 = 1$ and $x_0 > 0$ we see that

$$(x_0 - x_1)(x_0 + x_1) = 1 + x_2^2 \geq 1$$

and hence either $x_0 \pm x_1$ are both positive or both negative. Since $(1,0,0) \in \mathcal{H}^2$ which is connected and $x_0 \pm x_1$ is continuous, we must have $x_0 \pm x_1 > 0$ and thus $x_0 - x_1 > 0$. This means that f maps $\mathcal{H}^2$ into the open upper half-plane H^2. We also see that f is invertible:

$$f^{-1}(x+iy) = \frac{(x^2+y^2+1, x^2+y^2-1, -2x)}{2y},$$

so $f\colon \mathcal{H}^2 \to H^2$ is a bijection. If $z_1, z_2 \in H$ then one may easily calculate that

$$\langle f^{-1}(z_1), f^{-1}(z_2)\rangle_L = 1 + \frac{|z_1 - z_2|^2}{2\mathrm{Im}z_1 \mathrm{Im}z_2}.$$

Therefore, if we declare f to be an isometry, we can define a new hyperbolic distance on the upper half-space.

This gives our second model of hyperbolic 2-space.

Definition 8.9. The *upper half-space model* of hyperbolic 2-space is the (open) upper half-space

$$H^2 = \{x + iy \in \mathbb{C} : y > 0\}$$

with distance

$$d(z_1, z_2) = \cosh^{-1}\left(1 + \frac{|z_1 - z_2|^2}{2\mathrm{Im}z_1 \mathrm{Im}z_2}\right).$$

(If we write $d(x_1 + iy_1, x_2 + iy_2)$ in real coordinates it is then clear how to extend the definition to give the hyperbolic distance on H^n.)

Let us look at how this distance behaves on H^2.

Example. Let $z_1 = iy_1$ and $z_2 = iy_2$ where $0 < y_1 \leq y_2$. Then

$$\cosh d(iy_1, iy_2) = 1 + \frac{(y_1 - y_2)^2}{2y_1 y_2} = \frac{1}{2}\left(\frac{y_2}{y_1} + \frac{y_1}{y_2}\right) = \cosh\left(\log \frac{y_2}{y_1}\right).$$

Hence $d(iy_1, iy_2) = \log \frac{y_2}{y_1} \geq 0$. We deduce that $d(ir, i) = \log \frac{1}{r} \to \infty$ as $r \to 0$, and $d(i, ir) = \log r \to \infty$ as $r \to \infty$.

The examples shows that d is quite different from the Euclidean metric on H^2, since distances are now "becoming stretched" as we approach the real axis, which is the boundary $\partial\overline{H}^2$ of H^2.

The formula for the distance is not that important, but what we do care about is what the isometries and geodesics in (H^2, d) looks like. We already know, in some sense, what the isometries of (H^2, d) are, but we want to give another interpretation in terms of complex analysis.

Theorem 8.10. *The group* $\mathrm{Isom}^+(H^2)$ *of orientation-preserving isometries of the hyperbolic upper half-place* (H^2, d) *is the Möbius group of* H^2, $\mathrm{Möb}(H^2) \cong \mathrm{PSL}(2,\mathbb{R}) = \mathrm{SL}(2,\mathbb{R})/\{\pm I\}$. *Moreover,* $\mathrm{Isom}(H^2)$ *is generated by* $\mathrm{Möb}(H^2)$ *together with the map* $z \mapsto -\bar{z}$.

Remark. We know that a Möbius transformation preserves H^2 if and only if $T(z) = \frac{az+b}{cz+d}$ with $a, b, c, d \in \mathbb{R}$ and $ad - bc = 1$, which is how we see that $\mathrm{Möb}(H^2) = \mathrm{PSL}(2,\mathbb{R})$.

Proof. **(Not examinable).** Using the formula $T(z) = \frac{az+b}{cz+d}$ for $a, b, c, d \in \mathbb{R}$ with $ad - bc = 1$, we may calculate that

$$\mathrm{Im}\, T(z) = \frac{\mathrm{Im}z}{|cz+d|^2}$$

and that

$$|T(z_1) - T(z_2)|^2 = \frac{|(az_1+b)(cz_2+d) - (az_2+b)(cz_1+d)|^2}{|cz_1+d|^2|cz_2+d|^2} = \frac{|z_1 - z_2|^2}{|cz_1+d|^2|cz_2+d|^2}.$$

Therefore $d(T(z_1), T(z_2)) = d(z_1, z_2)$ and T is an orientation-preserving (as it is holomorphic) isometry.

It takes more work to show that these are the only orientation-preserving isometries, so I will not prove this.

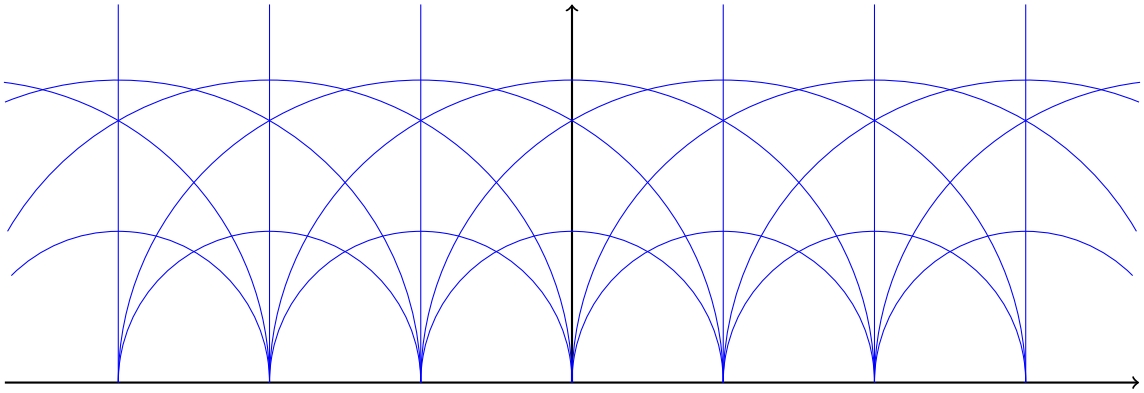
The map $T\colon z \mapsto -\bar{z}$ preserves H^2 and is an orientation-reversing isometry just by looking at the formula. Again, I will not show that it suffices to take the group generated by $\mathrm{Isom}^+(H^2)$ and T to obtain $\mathrm{Isom}(H^2)$. □

Example. We have that translations $T(z) = z + b$ lie in $\mathrm{Isom}(H^2)$, just like in $\mathbb{R}^2$.

We also have that dilations $T(z) = a^2 z$ are isometries of (H^2, d): this is very surprising, as these are definitely not isometries in $\mathbb{R}^2$.

Using the description of the isometries and the fact that the map $f : \mathcal{H}^2 \to H^2$ above takes geodesics to geodesics, we can describe the geodesics on (H^2, d).

Theorem 8.11. *The geodesics in the hyperbolic upper-half space consist of vertical half-lines given by* $\mathrm{Re}\, z = c$ *for* $c \in \mathbb{R}$ *and semi-circles with centre on the real axis.*



We finish the course with our third model of hyperbolic 2-space.

(Not examinable). Consider the map $f\colon \mathcal{H}^2 \to \mathbb{C}$ given by

$$f(x_0, x_1, x_2) = \frac{x_1 + ix_2}{1 + x_0},$$

which looks like stereographic projection. Since $x_0^2 - x_1^2 - x_2^2 = 1$ we see that

$$|f(x_0, x_1, x_2)|^2 = \frac{x_1^2 + x_2^2}{(1 + x_0)^2} = \frac{x_0^2 - 1}{x_0^2 + 2x_0 + 1} < 1.$$

Hence $f\colon \mathcal{H}^2 \to B^2$, the open unit disk in $\mathbb{C}$. We also see that f is invertible:

$$f^{-1}(x + iy) = \frac{(1 + x^2 + y^2, 2x, 2y)}{1 - x^2 - y^2},$$

so f is a bijection. If $z_1, z_2 \in B^2$ then we may calculate that

$$\langle f^{-1}(z_1), f^{-1}(z_2)\rangle_L = \frac{(1 + |z_1|^2)(1 + |z_2|^2) - 4\mathrm{Re}\,(z_1\bar{z_2})}{(1 - |z_1|^2)(1 - |z_2|^2)} = 1 + \frac{2|z_1 - z_2|^2}{(1 - |z_1|^2)(1 - |z_2|^2)}.$$

Therefore, if we declare f to be an isometry, we can define a new distance on the unit disk.

This gives us our final model of hyperbolic space.

Definition 8.12. The *Poincaré disk model* of hyperbolic 2-space is the (open) unit disk

$$D = B^2 = \{z \in \mathbb{C} : |z| < 1\}$$

with distance given by

$$d(z_1, z_2) = \cosh^{-1}\left(1 + \frac{2|z_1 - z_2|^2}{(1 - |z_1|^2)(1 - |z_2|^2)}\right).$$

(We clearly see how to extend this to all dimensions to define the hyperbolic metric on the open unit ball B^n in $\mathbb{R}^n$.)

Again, let us examine how the distance behaves.

Example. We see that if $z_1 = 0$ and $|z_2| = r$ where $r \in [0, 1]$ is real, then

$$d(0, r) = \cosh^{-1}\left(\frac{1 + r^2}{1 - r^2}\right).$$

We see that as $r \to 1$ then $d(0, z_2) \to \infty$.

This example shows that distances "become stretched" as we get closer to the boundary $\partial\overline{B}^2$ of the disk $D = B^2$, just like in the upper half-space model. We may now describe the isometries of the hyperbolic disk.

Theorem 8.13. *The group* $\mathrm{Isom}^+(D)$ *of orientation-preserving isometries of the hyperbolic disk is the Möbius group of the disk* $\mathrm{Möb}(D)$*. Moreover,* $\mathrm{Isom}(D)$ *is generated by* $\mathrm{Möb}(D)$ *together with complex conjugation* $z \mapsto \bar{z}$*.*

Remark. We know that $T \in \mathrm{Möb}(D)$ if and only if we can write $T(z) = \frac{az+b}{\bar{b}z+\bar{a}}$ with $|a|^2 - |b|^2 = 1$.

Proof. **(Not examinable).** Using the formula $T(z) = \frac{az+b}{\bar{b}z+\bar{a}}$ with $|a|^2 - |b|^2 = 1$ we may calculate that

$$|T(z_1) - T(z_2)|^2 = \frac{|(az_1 + b)(\bar{b}z_2 + \bar{a}) - (az_2 + b)(\bar{b}z_1 + \bar{a})|^2}{|\bar{b}z_1 + \bar{a}|^2|\bar{b}z_2 + \bar{a}|^2} = \frac{|z_1 - z_2|^2}{|\bar{b}z_1 + \bar{a}|^2|\bar{b}z_2 + \bar{a}|^2}$$

since $|a|^2 - |b|^2 = 1$. We can then see that

$$(1 - |T(z_1)|^2)(1 - |T(z_2)|^2) = \frac{(|\bar{b}z_1 + \bar{a}|^2 - |az_1 + b|^2)(|\bar{b}z_2 + \bar{a}|^2 - |az_2 + b|^2)}{|\bar{b}z_1 + \bar{a}|^2|\bar{b}z_2 + \bar{a}|^2} = \frac{(1 - |z_1|^2)(1 - |z_2|^2)}{|\bar{b}z_1 + \bar{a}|^2|\bar{b}z_2 + \bar{a}|^2}$$

and compute

$$|\bar{b}z + \bar{a}|^2 - |az + b|^2 = (|a|^2 - |b|^2)(1 - |z|^2) = 1 - |z|^2$$

to deduce that

$$\frac{|T(z_1) - T(z_2)|^2}{(1 - |T(z_1)|^2)(1 - |T(z_2)|^2)} = \frac{|z_1 - z_2|^2}{(1 - |z_1|^2)(1 - |z_2|^2)}.$$

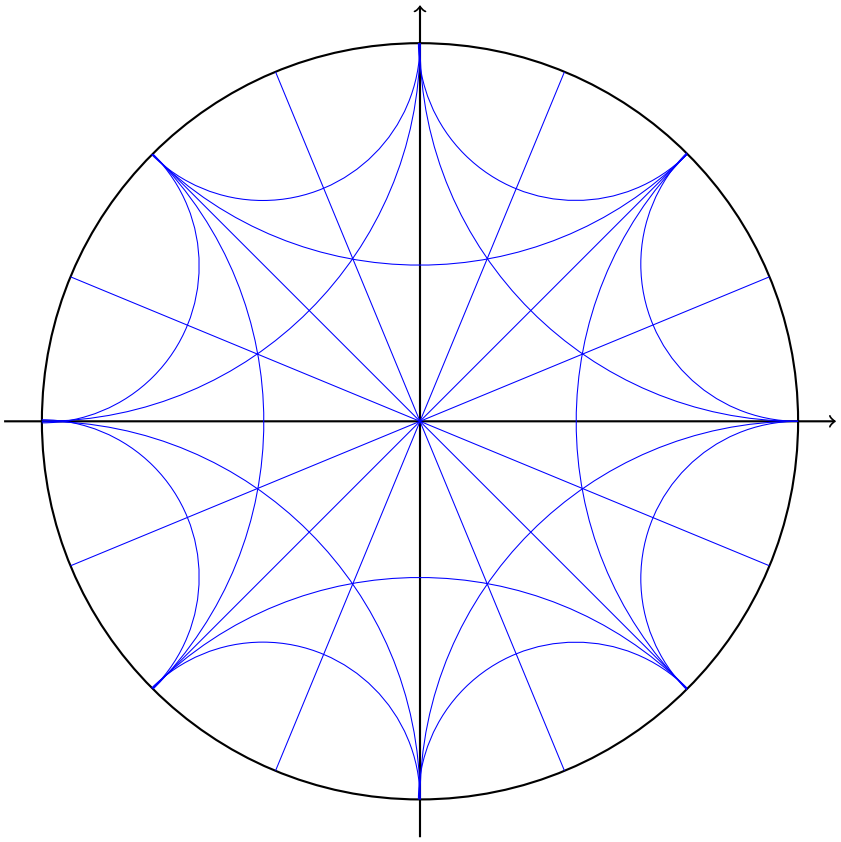
Hence $d(T(z_1), T(z_2)) = d(z_1, z_2)$ and thus T is an isometry.

Therefore, $\mathrm{Isom}^+(D)$ contains $\mathrm{Möb}(D)$. It also contains complex conjugation, by inspection, which is orientation-reversing. Again, I will not prove that these transformation generate the isometries. □

Again, using similar techniques as for the hyperbolic upper-half space, we can understand the geodesics as follows.

Theorem 8.14. *The geodesics in the hyperbolic disk consist of straight lines through* 0 *and arcs of circles which meet the unit circle orthogonally.*

In fact, we can see that the geodesics are projections of the geodesics in the hyperboloid model to the unit disc.



Remark. There is much more one can say about hyperbolic space, particularly about the geodesics and its *curvature*. We will see some of this on the problem sheet, but this will be discussed in detail in Part B Geometry of Surfaces.