AN INTRODUCTION TO DIFFERENTIAL GEOMETRY

STEPHEN C. PRESTON

Contents

Date: April 25, 2013.

1. Historical overview

"When nine hundred years old you reach, look as good you will not."

Geometry is just about as old as mathematics gets; only number theory could reasonably claim to be older. The first geometry results (at least four thousand years ago) involved word problems and specific numbers; students were expected to be able to substitute any other desired numbers for other problems. They got much further with plane geometry since drawing pictures made things easier, although surfaces or three-dimensional objects were always considered when results were possible. (One of the oldest known results is the volume of a frustum, for example.)

The invention of coordinates by Descartes made some problems a lot easier, and the invention of calculus opened up a lot of possibilities. Before this, there were tangent problems, and computations of areas and volumes through exhaustion, but they required some special structures since each new formula involved inventing a whole new technique. Calculus allowed mathematicians to approximate everything locally, then build back up to global structures, using formulas that could be systematized.

Two problems were first noticed two thousand years ago and could be considered to have led to the modern development of differential geometry: constructing a map of the earth, and Euclid's parallel postulate.

1.1. **Mapping the planet.**

The Greeks knew the earth was not flat, since there were stars that could be seen in the north and not the south. They assumed it was a sphere, partly because it cast a round shadow on the moon during an eclipse, but mostly because it seemed philosophically the most elegant possibility. (Using reports of shadows from the sun in various places, Eratosthenes computed the radius to a respectable accuracy.) The problem of making a map of the earth then surfaced; it's easier to carry around a map than a globe, but there's no way to draw an accurate map of a sphere. You can see this experimentally: try to lay a piece of paper on a globe, and you'll find it has to be folded to avoid gaps. A common trick is to cut out pieces of the paper to avoid folds, but if you're careful you'll see that this doesn't work either. No matter how what you cut or where, you never quite seem to make it fit.

Ptolemy was the first to worry about this. The earliest solutions involved projecting onto a disc tangent at a pole: via orthographic projection (drawing a line from a point on the sphere perpendicular to the disc, i.e., sending (x, y, z) to $(x, y, 1)$); or stereographic projection (drawing a line from a point on the sphere to the farther pole and marking where it crosses the equatorial disc, i.e., sending (x, y, z) to $(\frac{x}{1-z}, \frac{y}{1-z}, 1)$); or gnomonic projection (drawing a line from a point on the sphere through the origin and marking where it crosses a disc tangent to the nearer pole, i.e., sending (x, y, z) to $(\frac{x}{z}, \frac{y}{z}, 1)$). This is useful if you only really care about half of the globe. Stereographic projection is nice since it preserves angles; orthographic projection is nice because it reflects how you'd see a globe; gnomonic projection is nice because it maps shortest paths on the sphere to straight lines on the map. See Figures 1.1 and 1.2.

We can also wrap a cylinder around the sphere at the equator and project everything outward via some function from latitudes to heights. The Mercator projection

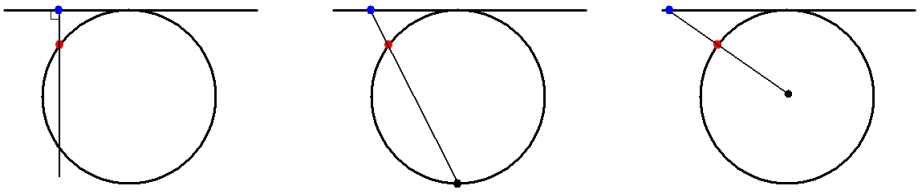

FIGURE 1.1. Orthographic, stereographic, and gnomonic projections, schematically from the side.

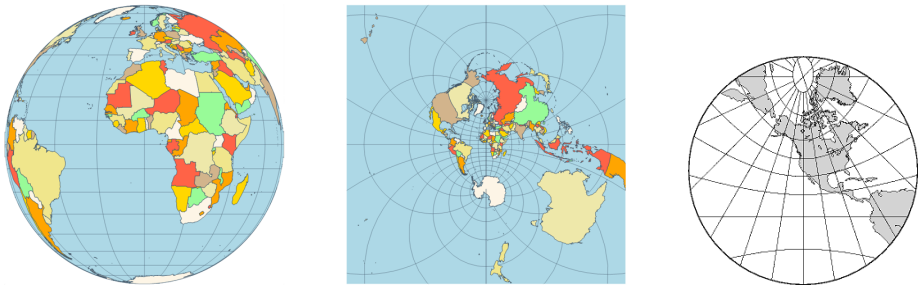

FIGURE 1.2. Orthographic, stereographic, and gnomonic maps of the world.

is what we get if we choose it to be angle-preserving, while the Gall-Peters projection is what we get if we choose it to be area-preserving. See Figure 1.3. (The area distortion in the Mercator projection was considered a political issue for years; the colonial powers all happened to be far from the equator and hence looked as large as the colonies near the equator. Cartographers generally advocate teaching geography using maps that compromise between the area-distortion of Mercator and the shape-distortion of Gall-Peters.)

More generally, the question is which surfaces are "developable," or isometric to a portion of a flat plane. You can generate accurate maps which preserve all lengths, shapes, and distances if you live on a developable surface, and you can wrap a developable surface efficiently if you have to give one as a gift. The discussion above shows that spheres are not developable, while there are several common surfaces that are developable despite looking curved in three-dimensional space. See Figure 1.4. Gauss, who worked as a surveyor as well as a mathematician, found an invariant of surfaces which could distinguish between those surfaces developable onto a flat plane and those which could not. In so doing he invented Gaussian curvature, as well as clarifying the notions of "intrinsic" and "extrinsic" geometry: that is, the distinction between geometric features that can be observed by people living on the surface, and those which depend on how the surface is bent in space. His insight served as the basis for Riemann's generalization of curvature as well as the idea that a manifold could be understood independently of its position in space.

1.2. **The parallel postulate.** The other ancient problem was the independence of Euclid's parallel (fifth) postulate from the other axioms. Playfair's version of the parallel postulate states that given any line L and any point P not on L, there is a

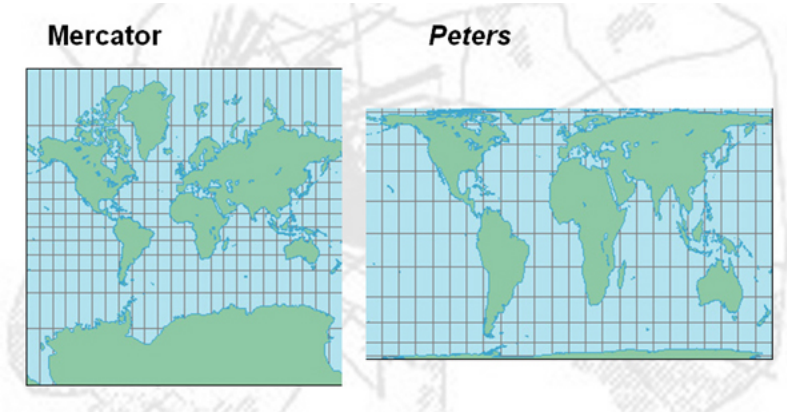

FIGURE 1.3. The Mercator projection (which preserves angles and distorts areas) vs. the Gall-Peters projection (which preserves areas and distorts angles).

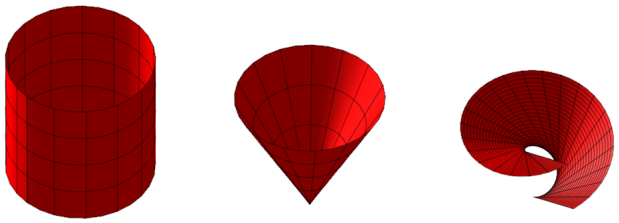

FIGURE 1.4. Three developable surfaces: although they appear curved in three-dimensional space, they can be unwrapped onto a plane, and geometers living on the surface would perceive each one as Euclidean.

unique line L' through P which never intersects L. For 1500 years, mathematicians tried to prove this uniqueness without success. (Spherical geometry with lines given by great circles seems like an obvious counterexample, but it wasn't considered valid since it violates the uniqueness of a line joining two points and the infinite extensibility of any line.) Eventually Bolyai and Lobachevsky effectively replaced the fifth postulate with the statement that there is at least one line L and point P such that there are two lines L' and L'' through P and not intersecting L. Following this through a long chain of arguments to its logical conclusion, one ends up with hyperbolic geometry. The development of this axiomatically is fascinating and fun, although it doesn't really fit in this course since it uses completely different techniques.

To prove there are no self-contradictions in the theory, one must eventually construct a model for it in some explicit way; this was first done by Beltrami and more rigorously by Klein, combining Bolyai's work with Gauss' and Riemann's. This was all in the mid-to-late 1800s. The simplest model is the upper half-plane (i.e., the set of (x, y) such that $y > 0$), where "lines" are semicircles centered on the x-axis. Playing around with this model, one can see that it's easy to draw infinitely

many lines through any given point that don't intersect a given one. See Figure 1.5.

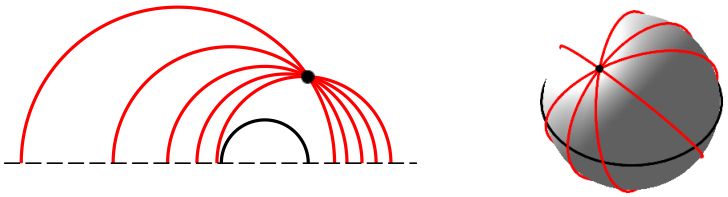

FIGURE 1.5. The parallel postulate on the hyperbolic plane (where "lines" are semicircles centered on the x-axis) and on the sphere (where "lines" are great circles). Given the black line and the black point, there are infinitely many nonintersecting red lines in hyperbolic space and no nonintersecting red lines on the sphere.

1.3. **Coordinates and manifolds.** In the mid-1800s, most geometric results were either on surfaces or on spaces with a lot of symmetry, partly since all the notation of vector calculus was still being invented, partly since it wasn't clear how to best parametrize objects, and partly since curvature was defined in a geometrically intuitive but computationally intimidating way. The theory of elasticity helped motivate progress, since it was essential to understand things like divergence and gradients in different coordinate systems. Complex analysis and the equations of electromagnetism also developed around the same time and motivated new coordinate systems and higher-dimensional computations. The work of Lamé, Ricci, and Levi-Civita led to a much better understanding of coordinate invariance, and through this eventually came the idea that coordinates were just a way of describing some underlying object, which ultimately led to the understanding of manifolds.

A major impetus to progress in geometry came from the idea that perhaps the space we live in is curved. Gauss actually tried to measure this, but it turned out space was too flat to actually notice curvature through measurements at earth-scales. Riemann casually suggested the idea that the sun might actually curve space itself, and that planets were not being pulled away from straight lines by a force but rather were following geodesics in a curved space. (For a model of this, imagine a sheet of rubber stretched out horizontally, with a heavy weight pulling it down in the center. If you slide a ball bearing across it at the right speed, it will spiral around the center for a while. So what if that's how the planets move, and we just can't see the rubber? See Figure 1.6.) It wasn't until relativity was discovered that Einstein was able to get this to work out correctly. See Misner, Thorne, and Wheeler's *Gravitation* for a nice discussion of these issues.

Since then, manifolds have come to be viewed as the central objects. Historically one worked in coordinates, which tended to obscure the global structure of things, and so manifolds were just a complication. Eventually in the 1900s it was realized that global structures really were important, and tools of algebraic topology such as index theorems were devised to study these systematically. The basic issue can already be seen in the problem of whether every curl-free field is a gradient. More precisely, if $\nabla \times X = 0$ for a vector field X on an open set $\Omega \subset \mathbb{R}^2$, is there a function

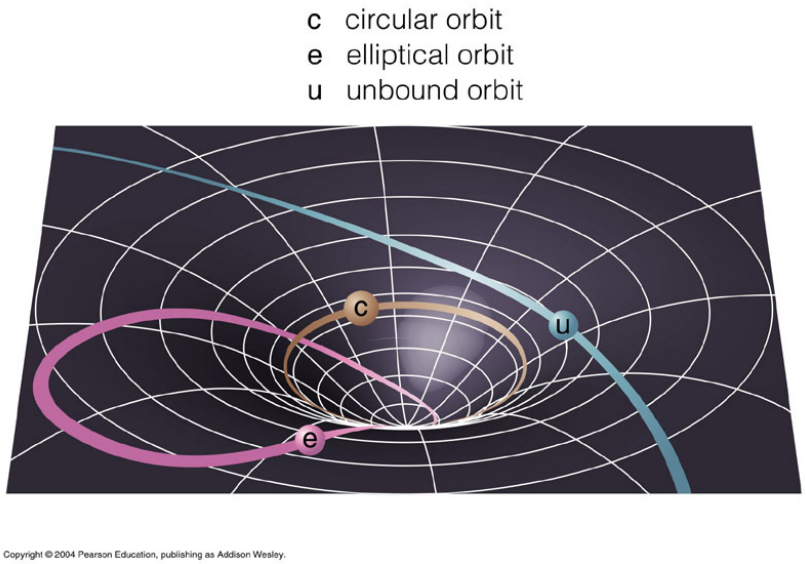

FIGURE 1.6. A common picture of gravitation as curved space: we imagine the sun in the center pulls the rubber sheet and causes celestial body trajectories to appear curved.

$f\colon \Omega \to \mathbb{R}$ such that $X = \nabla f$? This is almost true, but there are exceptions such as

$$X = \frac{-y}{x^2+y^2}\mathbf{i} + \frac{x}{x^2+y^2}\mathbf{j}$$

if Ω is the plane minus the origin. Here X is the gradient of $\theta(x,y) = \arctan(y/x)$, the usual angle function in polar coordinates, except θ is not an actual function on the plane with the origin deleted. Poincaré was the first to notice that these exceptions were actually the important part of the theory.

Mechanics has also motivated the study of manifolds. The study of constraints on a mechanical system is almost as old as mechanics itself, and they are frequently convenient to express as a manifold. For example, you could think of a pendulum as a system in $\mathbb{R}^2$ which happens to have a force generated along the rod which is just exactly enough to keep it from stretching, but it's actually easier to think of it as a system on the circle S^1. More generally one can reduce multiparticle systems with constraints to other types of manifolds; you end up with fewer coordinates and simpler systems. This works quite well for objects like rigid bodies (with six degrees of freedom) and even for things like fluids (with infinitely many degrees of freedom, but still constrained to be incompressible).

Starting with mechanics, but now with other motivations as well, mathematicians have begun to study manifolds with other structures aside from the metric that Riemann was originally interested in. Hence for example symplectic geometry and contact geometry have split off from Riemannian geometry, and both have been subsumed into "differential geometry." (Thus we teach you about manifolds first, in spite of the fact that they were basically the last things to be understood.)

Differential geometry has since found many other applications. Manifolds are useful in electrical engineering: for example, Stiefel manifolds arise in figuring out how to place cell phone towers. They also appear when one wants to define distances between objects in some way, such as when trying to get a computer to recognize a shape. Statisticians and economists have also found uses for them. Finally, many models in theoretical physics are quite naturally thought of as differential geometric structures on manifolds.

2. Introduction

"Fear is the path to the dark side. Fear leads to anger; anger leads to hate; hate leads to suffering. I sense much fear in you."

Differential geometry can be a difficult subject for students, for a couple of reasons. First, the notation can appear very cumbersome, and even seemingly familiar operations can look quite foreign. Second, the notions of manifold and curved geometry are new to students, and the idea that a manifold's geometry is best studied intrinsically (rather than, say, as something curved inside something else) is a huge conceptual leap to make. However, these two difficulties have nothing to do with each other, and it's my belief that they are best separated out. Once one has a handle on the notation, a lot of the genuinely new objects will seem very natural.

In these notes, my goal is to present the ideas of standard multivariable calculus in a way that can be generalized easily to manifolds. The vast majority of the issues that arise already appear in $\mathbb{R}^n$, the simplest manifold, when one tries to make notions such as vector fields, gradients, curls, divergences, Stokes' Theorem, etc., coordinate-independent. Thus we will obtain coordinate-independent versions of all these things on $\mathbb{R}^n$; when we generalize to do Riemannian geometry, the only things that will change will be the global manifold structure (which will introduce topological issues such as cohomology) and the metric (which will lead to curvature). Much of Riemannian geometry will then focus on the interplay between these two concepts.

Notation already becomes an issue when trying to do vector calculus on $\mathbb{R}^n$ in a coordinate-independent way: the standard notation does not generalize easily, since so much of it depends on the special structure of Euclidean space. (The Cartesian coordinates on Euclidean space have many special properties, most of which do not generalize to other coordinate systems such as polar coordinates; figuring out which properties we don't need will not only help us generalize, it will also give us greater insight into why vector calculus works the way it does.)

In doing this, we will need to give up certain things. First of all, we will no longer be able to say vectors with different base points are equivalent. (This property relies on the translation invariance of Euclidean space; since most other manifolds do not have anything similar, we want to avoid using it.) This has serious consequences; for example, differentiating vector fields no longer makes sense: in the definition of the derivative, we need to subtract vectors with different base points. Similarly, vector fields cannot be integrated. Eventually we *will* find a way to differentiate vector fields, which will look rather different from what we are used to. We will also find a way to integrate vectors in certain cases (the ones needed for Stokes' theorem and the divergence theorem), and this will help explain why operators like curl and divergence are special.

Most of the material here should be very familiar; only the way of thinking about it will be new. Until you learn this new way of thinking, little of differential geometry will make sense. And when we get to the material that is genuinely new, you won't have to make such large conceptual leaps. And then you'll be happy. And then you can make the world happier.

Part of the difficulty with this subject is that it is historically rather convoluted. Many false starts were made, with special results that were interesting but didn't seem to lead anywhere. Euler and Gauss were doing differential geometry, but their work is almost unreadable because they were inventing the notation as they went. Riemann must have done differential geometry, but his work is almost unreadable because he hardly wrote anything down. The tensor calculus (which is what most physicists still use when they do differential geometry) was not invented until the late 1800s (by Ricci and other Italian mathematicians). The notion of a manifold was not really clear until the 1930s. Even the true meaning of coordinate-invariance was not fully understood until the 1950s, which is around the same time that mathematicians started thinking of differential geometry in very abstract terms that still scare some physicists. The fact that vector calculus itself went through a number of competing developments (e.g., quaternions, which are no longer fashionable) hardly helps. Because of all this, differential geometry is a strange subject, with a fairly recent history filled with drastic upheavals in our understanding of what, exactly, it is.

By now mathematicians have pretty much settled on the notation that's most useful, but if we just skipped right to that, it wouldn't make sense to you, and you probably wouldn't be able to do much with it. The elegance of the modern theory obscures its intuitive basis, and without intuition, why do geometry? Thus the focus here will be on the history of the field and how successive generalizations led to the modern theory. We will skip the dead ends (things which are interesting by themselves but can't be generalized), and in so doing it will hopefully take us less than a hundred years to understand the subject. Good luck, and don't be afraid!

P.S. I will try to be quite detailed, but my interest is more in getting you to understand and use these tools; as such I will be somewhat less focused on giving the most general theorems or exploring technical results that aren't immediately applicable. My original motivation for writing this text was that the standard differential geometry books required prerequisites like tensor analysis, topology, calculus on manifolds, and curves and surfaces to really understand, and there was no course offered where the essential tools were fully covered. This text was thus originally designed to be used for the first month or so of a semester, but has since expanded to cover an entire course in differential geometry with fairly minimal prerequisites. Thus it's rather unique.

All students are assumed to have had at least undergraduate one-variable real analysis, vector calculus, and abstract linear algebra (beyond just matrices). Optional but helpful courses include multivariable real analysis, topology, curves and surfaces, and differential equations; we will briefly review the required results from these courses as needed. For alternative perspectives and greater depth on this material, I recommend the following books. Spivak's "Calculus on Manifolds" and Munkres' "Analysis on Manifolds" cover the earlier parts of the text, while Spivak's "Comprehensive Introduction to Differential Geometry" and Lee's "Introduction to Smooth Manifolds" cover the later parts. Other books you may want to consult are Munkres' "Topology" and Oprea's "Differential Geometry of Curves and Surfaces."

3. Linear algebra: Bases and linear transformations

"You must unlearn what you have learned."

3.1. **The role of a basis.** The vast majority of things we'll do in differential geometry will involve vectors, and thus you will need a good understanding of linear algebra to continue. We'll review the things we need and emphasize in particular the things that are most important for this subject.

First, we want to think of vector spaces as abstract things. A finite-dimensional vector space always has many possible choices of basis, and we want to think of any one of them as being equally valid for describing vectors. Thus we will rarely use the "standard basis" of $\mathbb{R}^n$, partly since that tends to bias us, and partly because it tends to obscure the more fundamental properties of linear transformations. A basic analogy that I'll repeat through the book is Plato's idea of us as only able to perceive the shadows of objects on a cave wall rather than the objects themselves: if you let go now of the idea that a vector is "really" a column of numbers in the standard basis and instead embrace the idea that an abstract vector is something not directly perceivable but rather expressible in many forms depending on where in the cave you're standing, you'll be well on your way to understanding differential geometry the right way.

So let's suppose we have a particular basis $\{e_1, e_2, \dots, e_n\}$ of an n-dimensional vector space V. Then any vector $v \in V$ may be written as $v = \sum_{k=1}^n a^k e_k$ for some unique real numbers $\{a^1, \dots, a^n\}$. (The reason for using both subscripts and superscripts will become clearer later. In general it helps us remember which objects are supposed to be invariant and which are not.) If we have some other basis $\{f_1, f_2, \dots, f_n\}$, then we can write the f-vectors as linear combinations of the e-vectors, via

$$f_i = \sum_{j=1}^n p_i^j e_j, \text{ for every } i, \tag{3.1.1}$$

for some real numbers p_i^j. In a linear algebra course you'd write this as a matrix

$$\begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix} = \begin{pmatrix} p_1^1 & p_1^2 & \cdots & p_1^n \\ p_2^1 & p_2^2 & \cdots & p_2^n \\ \vdots & \vdots & & \vdots \\ p_n^1 & p_n^2 & \cdots & p_n^n \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

(Again, please get used to the idea of writing both subscripts and superscripts to represent vectors, components, and matrices. It really is useful this way. Also get used to the idea that the summation (3.1.1) is going to be more useful than the matrix form.)

On the other hand, we also have a matrix of numbers q_j^k which give the vectors $\{e_j\}$ in terms of the vectors $\{f_k\}$, by

$$e_j = \sum_{k=1}^n q_j^k f_k.$$

Combining both formulas, we get

$$f_i = \sum_{j=1}^{n} p_i^j e_j = \sum_{j=1}^{n}\sum_{k=1}^{n} p_i^j q_j^k f_k = \sum_{k=1}^{n}\left(\sum_{j=1}^{n} p_i^j q_j^k\right) f_k,$$

and therefore since the $\{f_i\}$ are linearly independent, we must have

$$\sum_{j=1}^{n} p_i^j q_j^k = \begin{cases} 1 & \text{if } i = k \\ 0 & \text{otherwise} \end{cases}.$$

This just means that as matrices, $P = (p_i^j)$ and $Q = (q_i^j)$ are inverses: $PQ = I_n$.

Now if the same vector v is written as $v = \sum_{j=1}^{n} a^j e_j = \sum_{i=1}^{n} b^i f_i$, then the numbers a^i and b^j must be related by

$$\sum_{j=1}^{n} a^j e_j = \sum_{j=1}^{n}\sum_{k=1}^{n} a^j q_j^k f_k = \sum_{i=1}^{n} b^i f_i = \sum_{k=1}^{n} b^k f_k.$$

(In the last equality, we just changed the index from i to k; the index is a dummy variable, and so it doesn't matter which letter we use.) We therefore have

$$\sum_{k=1}^{n}\left(\sum_{j=1}^{n} a^j q_j^k\right) f_k = \sum_{k=1}^{n} b^k f_k,$$

and this implies that

$$b^k = \sum_{j=1}^{n} a^j q_j^k. \tag{3.1.2}$$

Notice what happens here: to get the *coefficients* in the f-basis from the coefficients in the e-basis, we use Q as in (3.1.2); to get the *basis vectors* in the f-basis from the basis vectors in the e-basis, we use P as in (3.1.1). This must happen since when we transform both components and basis vectors at the same time, we have to end up with the same vector, and this will only happen if the two transformations cancel each other out.

Example 3.1.1. Suppose $v = \left(\begin{smallmatrix} 3 \\ -4 \end{smallmatrix}\right)$. In the standard basis $e_1 = \left(\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right)$ and $e_2 = \left(\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right)$, we have $v = a^1 e_1 + a^2 e_2$ where $a^1 = 3$ and $a^2 = -4$.

Let $f_1 = \left(\begin{smallmatrix} 1 \\ -2 \end{smallmatrix}\right)$ and $f_2 = \left(\begin{smallmatrix} 3 \\ 2 \end{smallmatrix}\right)$. Then we have $f_1 = p_1^1 e_1 + p_1^2 e_2$ where $p_1^1 = 1$ and $p_1^2 = -2$. Similarly $p_2^1 = 3$ and $p_2^2 = 2$. To express the vector v in the basis $\{f_1, f_2\}$ we would write $v = b^1 f_1 + b^2 f_2$ for some numbers b^1 and b^2 and solve. More explicitly, in components we get $3 = b^1 + 3b^2$ and $-4 = -2b^1 + 2b^2$ with solution $b^1 = \frac{9}{4}$ and $b^2 = \frac{1}{4}$.

Alternatively we could express the vectors e_1 and e_2 in terms of f_1 and f_2, obtaining

$$e_1 = q_1^1 f_1 + q_1^2 f_2, \qquad e_2 = q_2^1 f_1 + q_2^2 f_2,$$

where $q_1^1 = \frac{1}{4}$, $q_1^2 = \frac{1}{4}$, $q_2^1 = -\frac{3}{8}$, and $q_2^2 = \frac{1}{8}$. We can then check the formulas

$$b^1 = q_1^1 a^1 + q_2^1 a^2, \qquad b^2 = q_1^2 a^1 + q_2^2 a^2$$

derived above. ☺

Observe that in all of these formulas, summing over the same index, appearing once on top and once on the bottom, has the effect of "canceling them out." We already see this in the formula $v = \sum_{j=1}^n a^j e_j$, where the summation cancels out the basis-dependent parts to give an invariant vector. Part of the usefulness of using both subscripts and superscripts is that it makes these transformation formulas somewhat easier to remember: we will find that any formula which has an index summed over without appearing once on top and once on the bottom is not correct in general.

The computations above will appear again and again whenever we work with indices, so it is convenient to have a general definition.

Definition 3.1.2. The notation δ_i^j or δ_{ij} is called the *Kronecker delta* for indices $1 \leq i, j \leq n$. It simply means

$$\delta_i^j = \begin{cases} 1 & i = j, \\ 0 & i \neq j. \end{cases}$$

The formula for δ_{ij} is the same.

You should think of the Kronecker delta as pretty much the same thing as the $n \times n$ identity matrix in linear algebra. Here are some basic properties of the Kronecker delta, which we will use repeatedly (and have already used once above).

Proposition 3.1.3. *Suppose $(a^i) = (a^1, \ldots, a^n)$ and $(b_j) = (b_1, \ldots, b_n)$ are any lists of numbers, vectors, or other quantities, for $1 \leq i \leq n$. Then for every j we have*

$$\sum_{i=1}^n \delta_i^j a^i = a^j, \tag{3.1.3}$$

and for every i we have

$$\sum_{j=1}^n \delta_i^j b_j = b_i. \tag{3.1.4}$$

Conversely, suppose ϕ_i^j is a list of numbers for $1 \leq i, j \leq n$ such that, for every list of numbers (a^i), we have

$$\sum_{i=1}^n \phi_i^j a^i = a^j \tag{3.1.5}$$

for every i. Then $\phi_i^j = \delta_i^j$.

As another converse, suppose ϕ_i^j is a list of numbers for $1 \leq i, j \leq n$ such that, for some basis $(e_1, \ldots, e_n)$ of an n-dimensional vector space, we have

$$\sum_{j=1}^n \phi_i^j e_j = e_i \tag{3.1.6}$$

for every i. Then $\phi_i^j = \delta_i^j$.

Proof. To prove (3.1.3), we note that no matter what a^i is, the quantity $\delta_i^j a^i$ is zero if $i \neq j$, and is equal to a^j if $i = j$. So we have

$$\sum_{i=1}^n \delta_i^j a^i = \sum_{i \neq j} \delta_i^j a^i + \sum_{i=j} \delta_i^j a^i = \sum_{i \neq j} 0 \cdot a^i + 1 \cdot a^j = a^j.$$

The proof of (3.1.4) is similar.

To prove (3.1.5), just use the fact that it's true for *any* list of numbers, and pick a simple list such as $a^1 = 1$ and $a^2 = \cdots = a^n = 0$. With this list we get

$$\sum_{i=1}^{n} \phi_i^j a^i = \phi_1^j = a^j,$$

which implies that $\phi_1^j = 1$ if $j = 1$ and $\phi_1^j = 0$ otherwise. Similarly we get $\phi_i^j = 1$ if $j = i$ and 0 otherwise.

The difference between (3.1.5) and (3.1.6) is that (3.1.5) is assumed true for any list of numbers while (3.1.6) is assumed true for one particular set of basis vectors. Notice that assuming (3.1.6) means that

$$\sum_{j=1}^{n} \phi_i^j e_j = \sum_{j=1}^{n} \delta_i^j e_j$$

using (3.1.4). By linearity of summations, we can write

$$\sum_{j=1}^{n} (\phi_i^j - \delta_i^j) e_j = 0.$$

Let i be any integer in $\{1, \ldots, n\}$. Then the equation says there is some linear combination of the $\{e_1, \ldots, e_n\}$ which is zero, and linear independence means that $\phi_i^j - \delta_i^j = 0$ for every j. But this is also true for any i since i was arbitrary. □

3.2. **Linear transformations.** Now suppose we have a linear transformation T from an m-dimensional vector space V to another n-dimensional vector space W. (Even if $m = n$, it is useful to treat V and W as being two different spaces in general.) If we choose a particular basis $\{e_1, \ldots, e_m\}$ of V and $\{h_1, \ldots, h_n\}$ of W, then T will have some coefficients T_i^j defined by the formula $T(e_i) = \sum_{j=1}^{n} T_i^j h_j$. Obviously if we change either the $\{e_i\}$ or the $\{h_j\}$, the coefficients T_i^j will change. In fact the ability to do this is what allows us to perform the standard reduction operations on matrices.

Example 3.2.1. Suppose for example $V = \mathbb{R}^3$ and $W = \mathbb{R}^2$. In the basis $\{e_1, e_2, e_3\}$ of V and $\{h_1, h_2\}$ of W, suppose we can write T as the matrix

$$T \sim \begin{pmatrix} 1 & 1 & 2 \\ 2 & 5 & -5 \end{pmatrix}.$$

Explicitly, this means $T(e_1) = h_1 + 2h_2$, $T(e_2) = h_1 + 5h_2$, and $T(e_3) = 2h_1 - 5h_2$. The obvious row-reduction—replacing row two (R_2) with ($R_2 - 2R_1$)—corresponds to changing the basis of the range space to the new basis $\{h_1', h_2'\}$ given by $h_1' = h_1 + 2h_2$ and $h_2' = h_2$, since then the matrix is

$$T \sim \begin{pmatrix} 1 & 1 & 2 \\ 0 & 3 & -9 \end{pmatrix},$$

and the corresponding vector equations are $T(e_1) = h_1'$, $T(e_2) = h_1' + 3h_2'$, and $T(e_3) = 2h_1' - 9h_2'$.

Dividing the second row through by 3 corresponds to another change of basis of the range, to $h_1'' = h_1'$ and $h_2'' = 3h_2'$. Then we get $T(e_1) = h_1''$, $T(e_2) = h_1'' + h_2''$,

and $T(e_3) = 2h_1'' - 3h_2''$, corresponding to matrix

$$T \sim \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & -3 \end{pmatrix}.$$

The last row reduction is subtracting row two from row one, which gives the basis $h_1''' = h_1''$ and $h_2''' = h_1'' + h_2''$ with matrix

$$T \sim \begin{pmatrix} 1 & 0 & 5 \\ 0 & 1 & -3 \end{pmatrix}.$$

So $T(e_1) = h_1'''$, $T(e_2) = h_2'''$, and $T(e_3) = 5h_1''' - 3h_2'''$.

This is reduced row echelon form, and as much as we can do with row operations. With column operations we can clear out more, by replacing column three (C_3) with $(C_3 - 5C_1 + 3C_2)$. This corresponds to a change of basis on V, given by $e_1' = e_1$, $e_2' = e_2$, and $e_3' = e_3 - 5e_1 + 3e_2$, which gives

$$T(e_1') = h_1''', \qquad T(e_2') = h_2''', \qquad T(e_3') = 0,$$

and the final simplest possible form of the matrix

$$T \sim \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

In summary, every row operation corresponds to a change of basis of the range, and every column operation corresponds to a change of basis in the domain.

☺

It is in this sense that elementary row operations do not change the linear transformation. There is nothing special about row operations either; you could do everything in terms of column operations and get the same result. The only reason row operations were historically preferred and are still primarily taught is that, when your matrix comes from a system of linear equations, the row operations do not change the unknowns while column operations do.

We now come to a very important set of concepts whose relationship forms the foundation of linear algebra. They get used most importantly in the Implicit Function Theorem.

Definition 3.2.2. Suppose $T\colon V \to W$ is a linear transformation from an m-dimensional vector space V to an n-dimensional vector space W. The *kernel* of T is

$$\ker T = \{v \in V \,|\, T(v) = 0\},$$

the *image* of T is

$$\operatorname{im} T = \{w \in W \,|\, \exists\, v \in V \text{ s.t. } T(v) = w\},$$

and the *rank* of T is the dimension of the image. The maximal rank of a linear transformation is $\min\{m, n\}$. The Fundamental Theorem of Linear Algebra says that

$$\operatorname{rank} T + \dim \ker T = \dim W. \tag{3.2.1}$$

Following the technique of Example 3.2.1, every matrix can be written in reduced row echelon form, which means that there is a basis $\{e_1', e_2', \ldots, e_m'\}$ of V and a basis $\{h_1', h_2', \ldots, h_n'\}$ of W for which $T(e_i') = h_i'$ for $i \leq r$ and $T(e_i') = 0$ for $i \geq r$, where r is the rank. Since r is the dimension of the image of T, it is obviously not dependent

on any particular basis, and hence any method of reducing the matrix must yield the same final reduced form.

All this depends on being able to change the bases for V and W separately. If T is a transformation from V *to itself*, then we are much more limited in the changes we can make to the matrix: for any basis $\{e_i\}$ of V, we write $T(e_i) = \sum_{j=1}^n T_i^j e_j$, and since both domain and range are expressed in the *same* basis, any change of the domain basis must be accompanied by the same change of the range basis. So if as before we have a new basis $\{f_j\}$, then the linear operator T can be written as $T(f_i) = \sum_{j=1}^n \tilde{T}_i^j f_j$ for some coefficients $\tilde{T}_i^j$. Let's compute $\tilde{T}_i^j$ in terms of T_i^j.

Proposition 3.2.3. *Suppose V is a vector space with bases $\{f_i\}$ and $\{e_i\}$ related by the formulas $f_j = \sum_{i=1}^n p_j^i e_i$ and $e_i = \sum_{j=1}^n q_i^j f_j$. Let $T\colon V \to V$ be a linear transformation expressed as*

$$T(e_i) = \sum_{j=1}^n T_i^j e_j \qquad \text{and} \qquad T(f_i) = \sum_{j=1}^n \tilde{T}_i^j f_j.$$

Then for every indices i and ℓ,

$$\tilde{T}_i^\ell = \sum_{j=1}^n \sum_{k=1}^n p_i^j q_k^\ell T_j^k. \tag{3.2.2}$$

Proof. We just write

$$\begin{aligned} T(f_i) &= \sum_{j=1}^n p_i^j T(e_j) \\ &= \sum_{j=1}^n \sum_{k=1}^n p_i^j T_j^k e_k \\ &= \sum_{j=1}^n \sum_{k=1}^n \sum_{\ell=1}^n p_i^j T_j^k q_k^\ell f_\ell. \end{aligned}$$

Since this right-hand side must be equal to $\sum_{\ell=1}^n \tilde{T}_i^\ell f_\ell$ for every i, we must have

$$\sum_{\ell=1}^n \left(\sum_{j=1}^n \sum_{k=1}^n p_i^j T_j^k q_k^\ell - \tilde{T}_i^\ell \right) f_\ell = 0.$$

Now linear independence of the $\{f_\ell\}$ implies (3.2.2). □

In matrix algebra formula (3.2.2) might be written as $T_f = PT_eQ = PT_eP^{-1}$. Because of this formula, if $T\colon V \to V$, then we cannot transform T into the identity in a new basis unless T was already the identity in the old basis. Thus for example we have eigenvalues which are well-defined independently of basis. We will discuss this more in the next Section.

3.3. **Transformation invariants.** The coefficients of a linear transformation in a particular basis have no real meaning. For a transformation $T\colon V \to W$, the only invariant is the rank, although for transformations $T\colon V \to V$ there are more. Although all the important ones reduce to the determinant, we will prove that the trace is invariant independently since the technique illustrates the general concept better and is frequently used on its own. We first discuss transformations from a

space V to itself, then at the end discuss how these invariants can be used for more general transformations.

The trace is used frequently for "averaging" a linear transformation, or more generally for averaging a tensor in certain directions.

Proposition 3.3.1. *If $T\colon V \to V$, then the* trace *of T is defined to be*

$$\operatorname{Tr}(T) = \sum_{i=1}^{n} T_i^i,$$

in any basis, and does not depend on choice of basis.

Proof. We just have to check that $\operatorname{Tr}(T)$ does not change when we change the basis. Let $\{e_i\}$ and $\{f_i\}$ be two bases as in Proposition 3.2.3. Then in the f-basis, the trace is

$$\begin{aligned}\operatorname{Tr}(\tilde{T}) &= \sum_{i=1}^{n} \tilde{T}_i^i = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} p_i^j q_k^i T_j^k = \sum_{j=1}^{n}\sum_{k=1}^{n} T_j^k \sum_{i=1}^{n} p_i^j q_k^i \\ &= \sum_{j=1}^{n}\sum_{k=1}^{n} T_j^k \delta_k^j = \sum_{j=1}^{n} T_j^j = \operatorname{Tr}(T),\end{aligned}$$

which is the trace in the e-basis. □

Similarly the determinant of T is basis-invariant. This will be very important when we get to Stokes' Theorem in Chapter 17. Before proceeding with the awkward but systematic definition, we recall some group theory: the permutations of the set $\{1, \cdots, n\}$ form a group S_n, with $n!$ elements. Furthermore, each permutation σ has a sign $\operatorname{sgn}(\sigma) = (-1)^k$, where k is the number of transpositions, i.e., the number of distinct pairs $\{i, j\}$ with $i < j$ and $\sigma(i) > \sigma(j)$. Finally, the sign is a homomorphism: $\operatorname{sgn}(\sigma \circ \tau) = \operatorname{sgn}(\sigma) \cdot \operatorname{sgn}(\tau)$. The definition of determinant is then as follows.

Definition 3.3.2. Suppose $T\colon V \to V$ is a linear transformation expressed in some basis $\{e_i\}$ of V. Then the *determinant* of T is defined by

$$\det T = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma)\, T^1_{\sigma(1)} T^2_{\sigma(2)} \cdots T^n_{\sigma(n)} \tag{3.3.1}$$

This is not a real definition until we know that $\det T$ does not depend on the choice of basis, which we will check in a moment. First let's work out an example.

Example 3.3.3. Suppose V is two-dimensional. Any linear transformation $T\colon V \to V$ is determined by the four numbers T_1^1, T_1^2, T_2^1, and T_2^2, which we can write in the matrix form as

$$\begin{pmatrix} T_1^1 & T_2^1 \\ T_1^2 & T_2^2 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

There are only two permutations of the two-element set $\{1, 2\}$: the identity ι and the transposition τ that exchanges them. (In other words, $\iota(1) = 1$, $\iota(2) = 2$, $\tau(1) = 2$, and $\tau(2) = 1$.) The identity has positive sign and the switcher has negative sign.

The formula (3.3.1) then gives

$$\det T = \operatorname{sgn}(\iota) T^1_{\iota(1)} T^2_{\iota(2)} + \operatorname{sgn}(\tau) T^1_{\tau(1)} T^2_{\tau(2)} = T_1^1 T_2^2 - T_2^1 T_1^2 = ad - bc,$$

which is the usual formula.

As a simple computation, let's show that the determinant of a product is the product of the determinants for 2×2 matrices. (We will do this in full generality in the next proof.) If $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and $B = \begin{pmatrix} e & f \\ g & h \end{pmatrix}$, then

$$AB = \begin{pmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{pmatrix},$$

and the determinant is

$$\begin{aligned} \det(AB) &= (ae+bg)(cf+dh) - (af+bh)(ce+dg) \\ &= aecf + bgcf + aedh + bgdh - afce - bhce - afdg - bhdg \\ &= bgcf + aedh - bhce - afdg \\ &= (ad-bc)(eh-fg) \\ &= (\det A)(\det B) \end{aligned}$$

The important thing here is the cancellation of the terms $aecf$ and $bhdg$: we can then factor the rest. In the next Proposition we will see why this cancellation happens and how it leads to the basis-invariance of the Definition 3.3.2. ☺

Proposition 3.3.4. *The determinant defined by Definition 3.3.2 does not depend on choice of basis.*

Proof. The first step is to prove that if A and B are two matrices, then $\det(AB) = \det A \det B$. This is often done by breaking up a matrix into elementary components, but we'll just plow through it using the definition (3.3.1). So suppose the components of A are (a_i^j) and the components of B are (b_i^j) in some basis.

Then $C = AB$ has components $(c_i^j) = (\sum_{k=1}^n a_k^j b_i^k)$. Thus the determinant of the product AB is

$$\begin{aligned} \det(AB) &= \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma)\, c^1_{\sigma(1)} \cdots c^n_{\sigma(n)} \\ &= \sum_{\sigma \in S_n} \sum_{k_1=1}^{n} \cdots \sum_{k_n=1}^{n} \operatorname{sgn}(\sigma)\, a^1_{k_1} b^{k_1}_{\sigma(1)} \cdots a^n_{k_n} b^{k_n}_{\sigma(n)}. \end{aligned} \tag{3.3.2}$$

On the other hand, we know that the product of the determinants is

$$\det A \det B = \left(\sum_{\kappa \in S_n} \operatorname{sgn}(\kappa)\, a^1_{\kappa(1)} \cdots a^n_{\kappa(n)} \right) \left(\sum_{\tau \in S_n} \operatorname{sgn}(\tau)\, b^1_{\tau(1)} \cdots b^n_{\tau(n)} \right). \tag{3.3.3}$$

So to get (3.3.2) to simplify to (3.3.3), the first thing we have to do is show that of the n^n terms in the sum over k_1 through k_n, all but $n!$ of them drop out. So observe that

$$\det(AB) = \sum_{k_1, \cdots, k_n = 1}^{n} a^1_{k_1} \cdots a^n_{k_n} \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma)\, b^{k_1}_{\sigma(1)} \cdots b^{k_n}_{\sigma(n)}. \tag{3.3.4}$$

Now just consider the inner sum on the last line. If any two of the k's are equal, then the sum must vanish. Here's the reasoning: suppose for example that $k_1 = k_2$. Then for every permutation σ_1, we will have a permutation σ_2 such that $\sigma_2(1) = \sigma_1(2)$,

$\sigma_2(2) = \sigma_1(1)$, and $\sigma_2(m) = \sigma_1(m)$ for $3 \leq m \leq n$. Then $\operatorname{sgn}(\sigma_2) = -\operatorname{sgn}(\sigma_1)$, and

$$\begin{aligned}\operatorname{sgn}(\sigma_1)\, b^{k_1}_{\sigma_1(1)} b^{k_2}_{\sigma_1(2)} \cdots b^{k_n}_{\sigma_1(n)} + \operatorname{sgn}(\sigma_2)\, b^{k_1}_{\sigma_2(1)} b^{k_2}_{\sigma_2(2)} \cdots b^{k_n}_{\sigma_2(n)} \\ = \operatorname{sgn}(\sigma_1)\, b^{k_1}_{\sigma_1(1)} b^{k_1}_{\sigma_1(2)} \cdots b^{k_n}_{\sigma_1(n)} - \operatorname{sgn}(\sigma_1)\, b^{k_1}_{\sigma_1(2)} b^{k_1}_{\sigma_1(1)} \cdots b^{k_n}_{\sigma_1(n)} = 0.\end{aligned}$$

In this way, half the permutations in S_n cancel out the other half, so we get zero for the inner sum in (3.3.4) when $k_1 = k_2$. Similarly we get zero whenever any two k_i's are equal, and the only time we can get something nonzero is when all k_i are distinct.

Now when all k's are distinct, then $(k_1, \cdots, k_n)$ must be a permutation of $(1, \cdots, n)$, so that $k_1 = \kappa(1), \cdots, k_n = \kappa(n)$ for some permutation κ. As a result, we can write

$$\det(AB) = \sum_{\kappa \in S_n} \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma)\, a^1_{\kappa(1)} \cdots a^n_{\kappa(n)} b^{\kappa(1)}_{\sigma(1)} \cdots b^{\kappa(n)}_{\sigma(n)}. \tag{3.3.5}$$

We're getting closer. The next step is to notice that for each fixed $\kappa \in S_n$, we can write any $\sigma \in S_n$ as $\sigma = \tau \circ \kappa$ for some τ (since S_n is a group). Then we obtain

$$\begin{aligned}\det(AB) &= \sum_{\kappa \in S_n} \sum_{\tau \in S_n} \big(\operatorname{sgn}(\tau \circ \kappa)\big)\, a^1_{\kappa(1)} \cdots a^n_{\kappa(n)} b^{\kappa(1)}_{\tau(\kappa(1))} \cdots b^{\kappa(n)}_{\tau(\kappa(n))} \\ &= \sum_{\kappa \in S_n} \operatorname{sgn}(\kappa)\, a^1_{\kappa(1)} \cdots a^n_{\kappa(n)} \sum_{\tau \in S_n} \operatorname{sgn}(\tau)\, b^{\kappa(1)}_{\tau(\kappa(1))} \cdots b^{\kappa(n)}_{\tau(\kappa(n))}.\end{aligned}$$

This now looks very close to (3.3.3). In order to see that it's exactly the same, just notice that for any particular permutation κ, we have

$$b^{\kappa(1)}_{\tau(\kappa(1))} \cdots b^{\kappa(n)}_{\tau(\kappa(n))} = b^1_{\tau(1)} \cdots b^n_{\tau(n)},$$

since each term shows up exactly once, and we're just multiplying them all together. So we have derived (3.3.3) from (3.3.2), and thus we know $\det(AB) = \det A \det B$.

As a consequence, the determinant of the identity matrix is $\det I_n = 1$. From there, since $PQ = I_n$, we have $\det P \det Q = 1$. And finally, under a change of basis with $T_f = P T_e Q$, we have

$$\det T_f = \det P \det T_e \det Q = \det T_e,$$

which establishes that the determinant is indeed independent of basis. □

Again, it needs to be emphasized that these quantities are only invariants when T operates as a linear map from a space *to itself*. If T is a linear nonsingular map from V to W, then we can always find a basis $\{e_1, \cdots, e_n\}$ of V and a basis $\{f_1 = T(e_1), \cdots, f_n = T(e_n)\}$ of W so that $T^i_j = \delta^i_j$, which makes $\det T = 1$. Thus, if we're allowed to change basis of both domain and range separately, then the determinant can be anything; if the domain and range are the same space, and all changes of basis have to be done simultaneously to both, then the determinant is a genuine invariant.

The most important invariant is the *characteristic polynomial*, given by $p(\lambda) = \det(\lambda I - T)$. If V is n-dimensional, this takes the form

$$p(\lambda) = \lambda^n - (\operatorname{Tr} T)\lambda^{n-1} + \cdots + (-1)^n (\det T).$$

Its roots are the eigenvalues of T, so those are also invariant. The other coefficients are also invariants, but they are used less frequently than the trace or determinant.

Finally we discuss the determinant for more general transformations $T\colon V \to W$. Obviously the determinant only makes sense if the matrix is square (i.e., if V and W have the same dimension). The numerical value certainly is not invariant, but whether it's zero or not *is* invariant.

Proposition 3.3.5. *Suppose V and W have the same dimension, and that $T\colon V \to W$ is a linear operator. Compute the determinant of T as a matrix from some basis of V to some basis of W. Then either the determinant is zero for every choice of basis, or the determinant is nonzero for every choice of basis.*

Proof. Suppose T is invertible. Then it has maximal rank, so there is a choice of basis $\{e_1, \dots, e_n\}$ of V and basis $\{h_1, \dots, h_n\}$ of W such that $T(e_i) = h_i$ for every $1 \leq i \leq n$. We now pretend that $h_i = e_i$ so that the matrix of T is the identity and the determinant of T is one. (More precisely, we are defining an isomorphism $\pi\colon W \to V$ by the formula $\pi(h_i) = e_i$ and computing the determinant of $\pi \circ T$. Of course the isomorphism π depends on our choice of bases.)

Now suppose we have a new basis $\{f_1, \dots, f_n\}$ of the domain V. Then we can write $f_i = \sum_{j=1}^n p_i^j e_j$ for some coefficients p_i^j, where the (p_i^j) form an invertible matrix. And we then have

$$T(f_i) = \sum_{j=1}^{n} p_i^j h_j,$$

so that pretending again that $h_i = f_i$, the matrix of T is the matrix of $P = (p_i^j)$. Since there is a matrix $Q = (q_i^j)$ such that PQ is the identity, we know $\det P \neq 0$. So the determinant of T (computed in this basis) is still nonzero. The same thing works if we change the basis of the range.

If T is not invertible, then its rank is less than n, and its reduced echelon form has zeros on the diagonal, which means that its determinant will be zero. Changing the basis of the domain or range again corresponds to multiplying by nonsingular matrices, and so the determinant of the new matrix will still be zero. □

In this course, the previous proposition is often applied practically to determine whether we can use the Inverse Function Theorem 5.2.4. The next proposition is often applied practically to determine whether we can use the Implicit Function Theorem 5.2.2. For small matrices, it is often easier to compute determinants than to compute a full row reduction in order to compute a rank.

Proposition 3.3.6. *Let $T\colon V \to W$ be a linear transformation with $m = \dim V$ and $n = \dim W$, and suppose $m < n$. Define a "subdeterminant" to be the determinant of an $m \times m$ submatrix formed by deleting $n - m$ columns from the matrix. Then T has maximal rank m if and only if there is at least one nonzero subdeterminant.*

The same result works if $m > n$ and we delete $m-n$ rows to get subdeterminants.

Proof. As in Section 3.1, the rank r is determined independently of a basis by performing row and column operations. Now suppose we can perform all the standard row operations without ever switching columns and obtain a reduced row echelon matrix that looks like the identity in its first m columns. In this case the determinant of the matrix formed by the first m columns must have *always* had nonzero determinant, since no terms from the later columns were ever used to determine how to perform the reduction.

More generally, if we have an $m \times n$ matrix, row reduction may lead to a column full of zeros even if the linear transformation has full rank. In this case switching columns will give a nonsingular matrix in the first m columns. The determinant of the corresponding submatrix in the original matrix is the same as the determinant of the first block of the column-switched matrix, and hence it's nonzero if the matrix has full rank.

On the other hand, if the rank is less than m, then no amount of column switching will ever lead to an $m \times m$ identity matrix after row reduction, and so every submatrix must also be degenerate, so all of their determinants are zero. □

One very important theme that has already come up is the connection between a change of basis on a single vector space and a linear transformation from one vector space to another. The basic idea, which will appear again in Chapter 15, is that anything we can do on a vector space that does not depend on the basis will also be invariant under linear transformations from one vector space to another, if we can make sense of it. This has appeared in the way we were able to treat basis changes as matrices in the proof of Proposition 3.3.4 (which means we were essentially treating a change of basis as a linear transformation by pretending our two different bases were actually elements of two different vector spaces, or conversely in Proposition 3.3.5 pretending that two different bases of different vector spaces were actually the same). Strictly speaking this confusion is incorrect, but it's very common, and it helps form an intuition for why results like Propositions 3.3.1 and 3.3.4 are so important. This will become clearer when you work with coordinate charts and objects defined in them.

4. Multilinear algebra: Duals, tensors, and forms

"An elegant weapon for a more civilized age."

Everything in Chapter 3 is standard in an undergraduate linear algebra course. We now study some aspects of linear algebra that are a little more obscure, but very useful for differential geometry.

4.1. **The dual space.**

Definition 4.1.1. If V is a vector space, then the *dual* V^* is defined as the set of linear functions on V. Explicitly,

$$V^* = \{\omega \colon V \to \mathbb{R} \,|\, \omega(av + bw) = a\omega(v) + b\omega(w) \text{ for all } v, w \in V\}.$$

Elements of V^* are called *covectors* or, in certain contexts, 1-*forms*.

V^* has its own vector space structure defined by

$$(c\omega)(v) = c\omega(v), \qquad (\omega + \beta)(v) = \omega(v) + \beta(v), \qquad \text{for all } v \in V.$$

Let's work out an example in two dimensions.

Example 4.1.2. Suppose V is a 2-dimensional vector space with a basis given by $e_1 = \left(\begin{smallmatrix}1\\1\end{smallmatrix}\right)$ and $e_2 = \left(\begin{smallmatrix}1\\2\end{smallmatrix}\right)$. Any element α of V^* is completely determined by what it does to e_1 and e_2, since a general vector v can be written as $v = c^1e_1 + c^2e_2$ and linearity of α implies that

$$\alpha(c^1e_1 + c^2e_2) = c^1\alpha(e_1) + c^2\alpha(e_2).$$

Suppose $\alpha(e_1) = 1$ and $\alpha(e_2) = 0$. Let's find a concrete way to represent α. Recall that you can multiply a row vector by a column vector to get a number: in matrix algebra you'd say that a $1 \times n$ matrix multiplied by an $n \times 1$ matrix gives a 1×1 matrix, or in other words a number. So let's suppose $\alpha = \begin{pmatrix} a & b \end{pmatrix}$ for some numbers c and d. Then

$$\alpha(e_1) = \begin{pmatrix} a & b \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = a + b = 1,$$

$$\alpha(e_2) = \begin{pmatrix} a & b \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = a + 2b = 0.$$

Solving these two equations gives $\alpha = \begin{pmatrix} 2 & -1 \end{pmatrix}$.

Similarly if $\beta \in V^*$ is the covector defined by $\beta(e_1) = 0$ and $\beta(e_2) = 1$, then we can check that β must be $\beta = \begin{pmatrix} -1 & 1 \end{pmatrix}$.

Finally, suppose ω is any other covector in V^*, and let's show that you can express ω in terms of α and β. Let $p = \omega(e_1)$ and $q = \omega(e_2)$, and let $\phi = p\alpha + q\beta$. Then ϕ is also an element of V^*, and

$$\phi(e_1) = p\alpha(e_1) + q\beta(e_1) = p \cdot 1 + q \cdot 0 = p$$
$$\phi(e_2) = p\alpha(e_2) + q\beta(e_2) = p \cdot 0 + q \cdot 1 = q.$$

Since ω and ϕ agree on the basis vectors e_1 and e_2, they must agree on *any* vector, and hence they must be the same covector. We have thus shown that α and β span the covector space V^*, and it is easy to see that they form a basis as well. ☺

Now let's generalize this to an n-dimensional vector space.

Proposition 4.1.3. *If V is finite-dimensional, then so is V^*, and the dimensions are the same.*

Proof. To prove this, we construct an explicit basis for V^*. So start with a basis $\{e_1, e_2, \ldots, e_n\}$ of V. Define linear functions $\alpha^1, \alpha^2, \ldots, \alpha^n$ by the formulas $\alpha^j(e_k) = \delta^j_k$ for all j and k. This is equivalent to

$$\alpha^j(v) = \alpha^j\left(\sum_{k=1}^n a^k e_k\right) = a^j \tag{4.1.1}$$

for each j, on any vector $v = \sum_{k=1}^n a^k e_k$.

To prove these n functions on V actually are a basis, we just have to express every other linear function in terms of them. So let ω be any linear function on V. Then for any vector $v = \sum_{k=1}^n a^k e_k$, we have

$$\omega(v) = \omega\left(\sum_{k=1}^n a^k e_k\right) = \sum_{k=1}^n a^k \omega(e_k) = \sum_{k=1}^n \alpha^k(v)\omega(e_k),$$

where $\omega(e_k)$ is just some real number for each k. Since this is true for every $v \in V$, we have

$$\omega = \sum_{k=1}^n \omega(e_k)\alpha^k.$$

This shows that the functions $\{\alpha^1, \ldots, \alpha^n\}$ span V^*.

Now we just need to show that they are linearly independent. So suppose we have some numbers $c_1, \ldots, c_n$ for which $c_1\alpha^1 + \cdots + c_n\alpha^n = 0$. We want to show that all c's are zero. Since the function $c_1\alpha^1 + \cdots + c_n\alpha^n$ is the zero function, its value on any vector is the number zero. Thus we have for any i that

$$0 = \sum_{j=1}^n c_j \alpha^j(e_i) = \sum_{j=1}^n c_j \delta^j_i = c_i.$$

Thus $c_i = 0$ for every i, and so the α's are a basis. Hence V^* is an n-dimensional vector space. □

The first thing that dual spaces can be used for is to properly generalize the transpose of a matrix.

Definition 4.1.4. Every linear transformation $T\colon V \to W$ has a corresponding *dual transformation* $T^*\colon W^* \to V^*$, defined by

$$\big(T^*(\beta)\big)(v) = \beta\big(T(v)\big)$$

for every $\beta \in W^*$.

This is a natural definition; if we want to get a linear function from V to $\mathbb{R}$ from one on W, we first go from V to W linearly by T, then go from W to $\mathbb{R}$ by the linear function on W. See Figure 4.1.

Suppose V has a basis $\{e_i\}$ with dual basis $\{\alpha^i\}$ of V^*, and W has a basis $\{f_j\}$ with dual basis $\{\beta^j\}$ of W^*. If T is given by $T(e_i) = \sum_{j=1}^n T^j_i f_j$, then we can compute the coefficients of T^* in the dual basis:

$$\big(T^*(\beta^i)\big)(e_j) = \beta^i\big(T(e_j)\big) = \beta^i\left(\sum_{k=1}^n T^k_j f_k\right) = \sum_{k=1}^n T^k_j \beta^i(f_k) = \sum_{k=1}^n T^k_j \delta^i_k = T^i_j.$$

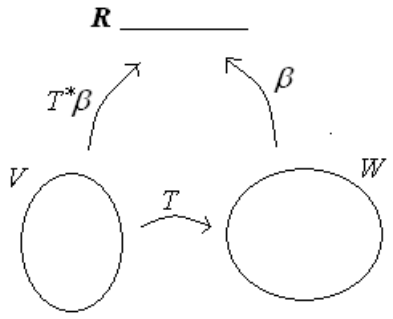

FIGURE 4.1. The pull-back $T^*\beta \in V^*$ of a $\beta \in W^*$, as a commuting diagram.

As a result we must have

$$T^*(\beta^i) = \sum_{j=1}^{n} T^i_j \alpha^j. \tag{4.1.2}$$

Notice how the components of T^* form a matrix which is the transpose of the matrix of T. The transpose of a matrix is not a basis-independent object, because $(PTP^{-1})^\dagger = (P^{-1})^\dagger T^\dagger P^\dagger \neq T^\dagger$; but the dual transformation *is* a basis-independent object, because we defined it without reference to any basis.

Remark 4.1.5. You have probably noticed that we have been consistently avoiding matrix notation, and you may feel like this is essentially "Linear Algebra Made Difficult." The reason for our use of index notation is that it becomes awkward to put linear operators into matrix form consistently, since for example we may want to apply a matrix of coefficients (p^j_i) either to vectors or covectors or coefficients (which would involve summing over either i or j). Personally I always find trying to do this very confusing. There are certainly ways to deal with most things without using indices each time, which we will aim for eventually, but one should be very careful not to forget that the index computations really are fundamental.

Example 4.1.6. Let us continue working in the situation from Example 4.1.2, where V is a two-dimensional vector space. Let $W = \mathbb{R}^3$ with the standard basis

$$f_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad f_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \qquad f_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

so that of course W^* has the standard dual basis

$$\beta^1 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}, \quad \beta^2 = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}, \quad \beta^3 = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}.$$

Consider a linear operator $T\colon V \to W$ defined by the vector equations $T(e_1) = f_1 - f_2 + 2f_3$ and $T(e_2) = -f_1 + 3f_2$. By Definition 4.1.4, the dual transformation T^* maps W^* to V^*, and since it is linear, we will know what it is as soon as we know what it does to the basis $\{\beta^1, \beta^2, \beta^3\}$ of W^*. Now $T^*(\beta^1)$ is an element of V^*, and we will know which one it is if we know what it does to e_1 and e_2. So we

compute:

$$T^*(\beta^1)(e_1) = \beta^1\big(T(e_1)\big) = \beta^1(f_1 - f_2 + 2f_3) = 1,$$
$$T^*(\beta^1)(e_2) = \beta^1\big(T(e_2)\big) = \beta^1(-f_1 + 3f_2) = -1,$$

and we conclude that

$$T^*(\beta^1) = T^*(\beta^1)(e_1)\alpha^1 + T^*(\beta^1)(e_2)\alpha^2 = \alpha^1 - \alpha^2.$$

Similarly we conclude that $T^*(\beta^2) = -\alpha^1 + 3\alpha^2$ and $T^*(\beta^3) = 2\alpha^1$.

Thus the matrix of T in basis $\{e_1, e_2\} \to \{f_1, f_2, f_3\}$ is

$$T \sim \begin{pmatrix} 1 & -1 \\ -1 & 3 \\ 2 & 0 \end{pmatrix},$$

while the matrix of T^* in basis $\{\beta^1, \beta^2, \beta^3\} \to \{\alpha^1, \alpha^2\}$ is

$$T^* \sim \begin{pmatrix} 1 & -1 & 2 \\ -1 & 3 & 0 \end{pmatrix}.$$

As a result, any reference to the transpose of a matrix really means the corresponding dual transformation. ☺

The formula (4.1.2) has an important and frequently useful consequence. In matrix algebra, the following statement would be $\det T^\dagger = \det T$, or the determinant of the transpose is the determinant of the original matrix.

Corollary 4.1.7. *If $T\colon V \to V$ is a linear transformation, then the dual linear transformation $T^*\colon V^* \to V^*$ (defined by Definition 4.1.4) has the same determinant.*

Proof. Recall the formula (3.3.1) for $\det T$ is

$$\det T = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma)\, T^1_{\sigma(1)} T^2_{\sigma(2)} \cdots T^n_{\sigma(n)}.$$

Since the coefficients of T^* are the reverse of those of T by (4.1.2), we have

$$\det T^* = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma)\, T_1^{\sigma(1)} \cdots T_n^{\sigma(n)}.$$

But recalling that $\operatorname{sgn}(\sigma) = \operatorname{sgn}(\sigma^{-1})$, we see that

$$\begin{aligned}
\det T^* &= \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma)\, T^{\sigma(1)}_{\sigma(\sigma^{-1}(1))} \cdots T^{\sigma(n)}_{\sigma(\sigma^{-1}(n))} \\
&= \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma^{-1})\, T^1_{\sigma^{-1}(1)} \cdots T^n_{\sigma^{-1}(n)} \\
&= \sum_{\rho \in S_n} \operatorname{sgn}(\rho)\, T^1_{\rho(1)} \cdots T^n_{\rho(n)} \\
&= \det T.
\end{aligned}$$

The equality on the second line follows as in Proposition 3.3.4: the product of all n terms is the same regardless of which permutation we take, since every term appears once. In the third line we renamed $\rho = \sigma^{-1}$ to match the definition of the determinant. □

Remark 4.1.8. **A vector space is not naturally isomorphic to its dual space.** One difficulty with the usual notation is that if one happens to have a transformation $T\colon \mathbb{R}^n \to \mathbb{R}^n$, then the transpose represents a transformation $T^*\colon (\mathbb{R}^n)^* \to (\mathbb{R}^n)^*$, and it is easy to forget (if we only write the matrix) that T is supposed to be operating on column vectors while T^* is supposed to be operating on row vectors. One gets away with this by implicitly treating $\mathbb{R}^n$ and $(\mathbb{R}^n)^*$ as "the same," since they have the same dimension. This is wrong, and it will be important in this field to avoid it.

Once we get used to taking the transpose of a matrix, it becomes tempting to just take the transpose of a column vector to get a row vector. What does such a thing mean? Well, we can think of any vector in $V = \mathbb{R}^2$ for example as being $v = v^1 e_1 + v^2 e_2$, and if we wanted to we could just define an isomorphism from $\xi\colon V \to V^*$ by

$$(v^1 e_1 + v^2 e_2) \mapsto (v^1 \alpha^1 + v^2 \alpha^2).$$

Here $\{e_1, e_2\}$ and $\{\alpha^1, \alpha^2\}$ are the bases from Example 4.1.2. This certainly is an isomorphism from V to V^*, but it depends on the basis we chose. Explicitly the vector $v = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$ has coefficients 2 and 4 in the standard basis but -1 and 3 in the basis $\{e_1, e_2\}$. Thus the transpose of it is $v^* = \begin{pmatrix} 2 & 4 \end{pmatrix}$ using the standard basis but $v^* = -1\alpha^1 + 3\alpha^2 = \begin{pmatrix} -5 & 4 \end{pmatrix}$ using the basis $\{e_1, e_2\}$. It would be something else entirely using another basis.

So there is no basis-independent notion of v^*. More generally there is no basis-independent isomorphism from V to V^*, although there are many possible isomorphisms since they have the same dimension as vector spaces. We must be careful about which operations actually are basis-independent, and hence it is important to distinguish a vector space from its dual.

Although we have seen that V and V^* are not "naturally" isomorphic, it turns out that $(V^*)^*$ and V *are* naturally isomorphic. That is, there is an isomorphism that can be defined without reference to a basis. The following proposition explains the reason for the term "dual": if we perform it twice, we end up back where we started.

Proposition 4.1.9. *If V is any vector space, then the dual of the dual of V is naturally isomorphic to V: that is, there is an isomorphism $\iota\colon V \to (V^*)^*$ defined independently of any basis.*

In addition, for any linear transformation $T\colon V \to W$, the dual of the dual is isomorphic to T: that is, if $\iota_V\colon V \to (V^*)^*$ and $\iota_W\colon W \to (W^*)^*$ are the isomorphisms, then $(T^*)^* \circ \iota_V = \iota_W \circ T$.

Proof. Every element of V^* is a function $\omega\colon V \to \mathbb{R}$ which is linear and completely determined by its action on every element $v \in V$. Thus we are essentially considering ω fixed and v varying, and looking at the number $\omega(v)$. But we could also consider a fixed v and let ω vary, and in this way obtain a function from V^* to $\mathbb{R}$. This will be an element of $(V^*)^*$, since it's obviously linear (by definition of addition and scalar multiplication in V^*). So for every $v \in V$, we have a linear function $\tilde{v}\colon V^* \to \mathbb{R}$ defined by $\tilde{v}(\omega) = \omega(v)$ for every $\omega \in V^*$. We define $\iota\colon V \to (V^*)^*$ by $\iota(v) = \tilde{v}$. It is clear that ι is a linear map and that ι is one-to-one, and since V and $(V^*)^*$ have the same dimension, ι must be an isomorphism by the Fundamental Theorem of Linear Algebra, (3.2.1).

To prove $(T^*)^*$ is equivalent to T, we take any $v \in V$; then for any $\omega \in V^*$, we have

$$(T^*)^*(\tilde{v})(\omega) = \tilde{v}\big(T^*(\omega)\big) = T^*(\omega)(v) = \omega\big(T(v)\big) = \widetilde{T(v)}(\omega).$$

Thus $(T^*)^*(\tilde{v}) = \widetilde{T(v)}$ for every $v \in V$, which is equivalent to saying $(T^*)^* \circ \iota_V(v) = \iota_W \circ T(v)$ for every v. Thus $(T^*)^* \circ \iota_V = \iota_W \circ T$. □

In finite dimensions, the dual space is always isomorphic to the original vector space (since all finite-dimensional vector spaces of the same dimension are isomorphic). In infinite dimensions, this is no longer true: as a well-known example from a graduate real analysis course, the dual of $L^1(\mathbb{R}) = \left\{ f\colon \mathbb{R} \to \mathbb{R} \,\middle|\, \int_{-\infty}^{\infty} |f(x)|\,dx < \infty \right\}$ is $L^\infty(\mathbb{R}) = \left\{ f\colon \mathbb{R} \to \mathbb{R} \,\middle|\, \exists M \text{ s.t. } |f(x)| \le M \text{ for almost every } x \in \mathbb{R} \right\}$. The fact that V^* may not be equal to V in infinite dimensions is what leads to the importance of distinguishing a vector space from its dual, even in finite-dimensions.

To get an intuition for this, notice that we do have an isomorphism from V to V^*: just take any basis $\{e_1, \dots, e_m\}$ of V and the dual basis $\{\alpha^1, \dots, \alpha^m\}$ of V^* satisfying $\alpha^i(e_j) = \delta^i_j$. An isomorphism is given by

$$v = a^1 e_1 + \cdots + a^m e_m \mapsto a^1 \alpha^1 + \cdots + a^m \alpha^m.$$

However if we chose a different basis of V, we would get a different dual basis of V^*, and a different isomorphism from this procedure. We saw this already in Remark 4.1.8 in two dimensions. (The position of the indices is a hint that the object is not basis-invariant, which I remind you is why we use subscripts and superscripts separately.)

On the other hand, the isomorphism between V and $(V^*)^*$ does not depend on any particular choice of basis, and is therefore more natural. However, the fact that ι is an isomorphism depends on finite-dimensionality: we used the fact that both V and $(V^*)^*$ have the same finite dimension. In infinite dimensions this too can break down; for example L^1 is not isomorphic to $(L^1)^{**}$, because although the map ι is one-to-one, there's nothing to force it to be onto.

4.2. **Tensors on vector spaces.** Once we have both vectors in V and covectors in V^*, we can talk about tensors in general.

Definition 4.2.1. If V is a finite-dimensional vector space, a *tensor of order* (p, q) *on* V is a multilinear map $T\colon V^p \times (V^*)^q \to \mathbb{R}$; in other words, T takes p vectors and q covectors and gives a real number. Multilinearity means that if all but one slots in T are held fixed, then T is a linear operator on the remaining slot.

The most important example is a bilinear form $g\colon V \times V \to \mathbb{R}$, which is a tensor of order $(2, 0)$. Such a form is determined by its coefficients in any basis; for example, if the basis of V is $\{e_1, \dots, e_n\}$, then for any vectors u and v in V, we have

$$u = \sum_{j=1}^n a^j e_j \text{ and } v = \sum_{k=1}^n b^k e_k \quad \text{for some numbers } a^j, b^k, \tag{4.2.1}$$

so that

$$g(u, v) = g\left(\sum_{j=1}^n a^j e_j, \sum_{k=1}^n b^k e_k \right) = \sum_{j=1}^n a^j g\left(e_j, \sum_{k=1}^n b^k e_k \right) = \sum_{j=1}^n \sum_{k=1}^n a^j b^k g(e_j, e_k),$$

so that g is completely determined by the n^2 numbers $g_{jk} \equiv g(e_j, e_k)$. If we define the tensor product $\otimes$ on two covectors ω and ξ in V^* as a tensor $\omega \otimes \xi \colon V \times V \to \mathbb{R}$ by

$$(\omega \otimes \xi)(u, v) = \omega(u) \cdot \xi(v),$$

then we can express g as

$$g = \sum_{j=1}^{n} \sum_{k=1}^{n} g_{jk} \alpha^j \otimes \alpha^k,$$

since both sides have the same operations on any pair of vectors. For example, on basis vectors e_ℓ and e_m the left side is $g_{\ell m}$ by definition while the right side is

$$\sum_{j=1}^{n} \sum_{k=1}^{n} g_{jk} \alpha^j(e_\ell) \cdot \alpha^k(e_m) = \sum_{j=1}^{n} \sum_{k=1}^{n} g_{jk} \delta_\ell^j \delta_m^k = g_{\ell m}$$

using the Kronecker delta identities in Proposition 3.1.3.

More generally, any tensor B of order (p, q) can be expressed in terms of a basis $\{e_i\}$ of V and dual basis $\{\alpha^i\}$ of V^* using

$$B = \sum_{i_1=1}^{n} \cdots \sum_{i_p=1}^{n} \sum_{j_1=1}^{n} \cdots \sum_{j_q=1}^{n} B_{i_1 \cdots i_p}^{j_1 \cdots j_q} \alpha^{i_i} \otimes \cdots \otimes \alpha^{i_p} \otimes e_{j_1} \otimes \cdots \otimes e_{j_q}.$$

Here we are using the identification of V with $(V^*)^*$ to view each e_k as a function on V^*. The coefficients $B_{i_1 \cdots i_p}^{j_1 \cdots j_q}$ clearly depend on the basis; changing the basis will require us to multiply these coefficients by $(p+q)$ transformation matrices involving the matrices P and Q. As a result, general tensor algebra can be an awful ugly mess. The main thing to remember is that since vectors are defined abstractly, so are all tensors; the components are just a convenient way to calculate specific numbers.

Another example of a tensor is the evaluation tensor, $E \colon V \times V^* \to \mathbb{R}$, defined by the easy formula $E(v, \omega) = \omega(v)$. Of course this is actually a tensor, since it's linear in both v and ω separately. Its components in a basis $\{e_i\}$ and $\{\alpha^i\}$ are found from

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a^i q_j E_i^j = E\left(\sum_{i=1}^{n} a^i e_i, \sum_{j=1}^{n} q_j \alpha^j \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} a^i q_j \alpha^j(e_i) = \sum_{i=1}^{n} \sum_{j=1}^{n} a^i q_j \delta_i^j.$$

Thus $E_i^j = \delta_i^j$, and we can write

$$E = \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i^j \alpha^i \otimes e_j,$$

since both sides give the same answer when applied to any pair (v, ω).

If you understand how both bilinear forms and the evaluation tensor work, you've pretty much got most of tensor analysis.

4.3. **k-forms on vector spaces.** A very important special case of tensors is k-forms, which are tensors of type $(k, 0)$ which are totally antisymmetric.

Definition 4.3.1. Let $k \in \mathbb{N}$. A tensor ω of type $(k, 0)$ on V is called a *k-form* if it is totally antisymmetric, i.e., if

$$\omega(v_1, \ldots, v_i, \ldots, v_j, \ldots, v_k) = -\omega(v_1, \ldots, v_j, \ldots, v_i, \ldots, v_k)$$

for every two indices i and j, and for any k vectors $v_1, \cdots, v_k \in V$. The vector space of all k-forms on V is denoted by $\Omega^k(V)$.

We define a 0-form to be a real number, so that $\Omega^0(V) = \mathbb{R}$.

Observe that *any* tensor of type $(1,0)$ is a 1-form under this definition, since the antisymmetry imposes no restriction. A 2-form ω will satisfy the single equation $\omega(v,w) = -\omega(w,v)$ for all $v, w \in V$. A 3-form ω will satisfy the three equations $\omega(u,v,w) = -\omega(v,u,w)$, $\omega(u,v,w) = -\omega(w,v,u)$, and $\omega(u,v,w) = -\omega(u,w,v)$ for all $u, v, w \in V$.

Example 4.3.2. Let's see what $\Omega^k(V)$ looks like on an n-dimensional vector space V for small values of k and n. Of course for $k = 1$ we always have $\Omega^1(V) = V^*$, so we will only worry about $k \geq 2$.

- $k = 2$ and $n = 1$. There is only one basis vector e_1, and the condition $\omega(e_1, e_1) = -\omega(e_1, e_1)$ requires that $\omega(e_1, e_1) = 0$. Hence by linearity, $\omega(u, v) = 0$ for every pair of vectors, and thus $\Omega^2(V) = 0$.
- $k = 2$ and $n = 2$. We have four possible choices of $\omega_{ij} = \omega(e_i, e_j)$. As above we must have $\omega(e_1, e_1) = 0$ and $\omega(e_2, e_2) = 0$, and we also have $\omega(e_1, e_2) = -\omega(e_2, e_1)$. So $\omega_{11} = \omega_{22} = 0$ and $\omega_{21} = -\omega_{12}$. Thus $\Omega^2(V)$ is one-dimensional, and every 2-form on a two-dimensional vector space can be written
$$\omega = \omega_{12}\alpha^1 \otimes \alpha^2 + \omega_{21}\alpha^2 \otimes \alpha^1 = \omega_{12}(\alpha^1 \otimes \alpha^2 - \alpha^2 \otimes \alpha^1).$$
- $k = 3$ and $n = 2$. We can check in the same way that the independent components are ω_{12}, ω_{23}, and ω_{31}, so that $\Omega^2(V)$ is three-dimensional. The most general 2-form can be written as
$$\begin{aligned}\omega = \omega_{12}(\alpha^1 \otimes \alpha^2 - \alpha^2 \otimes \alpha^1) + \omega_{23}(\alpha^2 \otimes \alpha^3 - \alpha^3 \otimes \alpha^2) \\ + \omega_{31}(\alpha^3 \otimes \alpha^1 - \alpha^1 \otimes \alpha^3).\end{aligned}$$
- $k = 3$ and $n = 2$. We need to determine $\omega_{ijk} = \omega(e_i, e_j, e_k)$ to get the basis components, and since there are only two basis vectors, some pair of the indices $\{i, j, k\}$ must be equal. By antisymmetry in each pair, this forces ω_{ijk} to always be zero. So $\Omega^3(V) = 0$.
- $k = 3$ and $n = 3$. We need to determine all possible values of $\omega_{ijk} = \omega(e_i, e_j, e_k)$. The only way we get anything nonzero is if the three indices are distinct, which means they must be a permutation of $\{1, 2, 3\}$. So everything is determined by ω_{123}, and $\Omega^3(V)$ is one-dimensional. Antisymmetry in each pair then implies all the other components are $\omega_{132} = -\omega_{123}$, $\omega_{213} = -\omega_{123}$, $\omega_{231} = \omega_{123}$, $\omega_{312} = \omega_{123}$, and $\omega_{321} = -\omega_{123}$. So we can write the most general 3-form on a three-dimensional space as
$$\begin{aligned}\omega = \omega_{123}\big(&\alpha^1 \otimes \alpha^2 \otimes \alpha^3 + \alpha^2 \otimes \alpha^3 \otimes \alpha^1 + \alpha^3 \otimes \alpha^1 \otimes \alpha^2 \\ &- \alpha^1 \otimes \alpha^3 \otimes \alpha^2 - \alpha^2 \otimes \alpha^1 \otimes \alpha^3 - \alpha^3 \otimes \alpha^2 \otimes \alpha^1\big).\end{aligned}$$

For obvious reasons, we will want a more concise notation to express a k-form in a basis of covectors. This is what the wedge product is for. ☺

To understand what $\Omega^k(V)$ looks like in general, we first need to define the wedge product of forms. Our goal is to get k-forms in terms of 1-forms, which are just elements of V^*.

Definition 4.3.3. Suppose β is a j-form and γ is a k-form. Then we define the *wedge product* of β and γ to be a $(j+k)$-form satisfying, for all vectors $v_1, v_2, \ldots, v_{j+k} \in V$, the formula

$$\begin{aligned}(4.3.1)\quad &\beta \wedge \gamma(v_1, \ldots, v_{j+k}) \\ &\qquad = \frac{1}{j!k!} \sum_{\sigma \in S_{j+k}} \operatorname{sgn}(\sigma) \beta\big(v_{\sigma(1)}, \cdots, v_{\sigma(j)}\big) \gamma\big(v_{\sigma(j+1)}, \cdots, v_{\sigma(j+k)}\big),\end{aligned}$$

where, as before, S_m denotes the set of $m!$ permutations of the numbers $\{1, \cdots, m\}$, while sgn denotes the sign of each permutation.

The term $\frac{1}{j!k!}$ is to compensate for the fact that all the terms will appear more than once in the sum. For example, suppose β is a 2-form and γ is a 1-form. Then the 3-form $\beta \wedge \gamma$ will satisfy, for any three vectors $u, v, w \in V$,

$$\begin{aligned}\beta \wedge \gamma(u, v, w) = \frac{1}{2}\Big(&1 \cdot \beta(u, v)\gamma(w) + 1 \cdot \beta(v, w)\gamma(u) + 1 \cdot \beta(w, u)\gamma(v) \\ &+ (-1) \cdot \beta(v, u)\gamma(w) + (-1) \cdot \beta(w, v)\gamma(u) + (-1) \cdot \beta(u, w)\gamma(v)\Big) \\ = \beta(u, v)\gamma(w) &+ \beta(v, w)\gamma(u) + \beta(w, u)\gamma(v).\end{aligned}$$

Without the factor of $\frac{1}{2}$, we would have a redundancy, since for example we know $\beta(u, v)\gamma(w) = -\beta(v, u)\gamma(w)$ by the antisymmetry of the 2-form. The factorials also ensure that the wedge product is associative, as in the next proposition.

Since the proof is rather involved, let's do an example first to see how it works. If you're better at combinatorics and finite group theory than I am, you can probably skip this example entirely.

Example 4.3.4. Suppose α is a 1-form, β is a 2-form, and γ is a 1-form on some vector space. Let's prove that $(\alpha \wedge \beta) \wedge \gamma = \alpha \wedge (\beta \wedge \gamma)$. First we take four vectors $v_1, \ldots, v_4$ and plug them into both sides.

For the left side, we have

$$\big((\alpha \wedge \beta) \wedge \gamma\big)(v_1, v_2, v_3, v_4) = \frac{1}{3! \cdot 1!} \sum_{\sigma \in S_4} \operatorname{sgn}(\sigma)(\alpha \wedge \beta)(v_{\sigma(1)}, v_{\sigma(2)}, v_{\sigma(3)})\gamma(v_{\sigma(4)}).$$

To compute $(\alpha \wedge \beta)\big(v_{\sigma(1)}, v_{\sigma(2)}, v_{\sigma(3)}\big)$, write $w_1 = v_{\sigma(1)}$, $w_2 = v_{\sigma(2)}$, and $w_3 = v_{\sigma(3)}$. Then

$$(\alpha \wedge \beta)(w_1, w_2, w_3) = \frac{1}{1! \cdot 2!} \sum_{\tau \in S_3} \alpha(w_{\tau(1)}) \cdot \beta(w_{\tau(2)}, w_{\tau(3)}).$$

To write $w_{\tau(1)}$ in terms of v, notice that if for example $\tau(1) = 3$ then $w_{\tau(1)} = w_3 = v_{\sigma(3)}$, and so $w_{\tau(1)} = v_{\sigma(\tau(1))}$. The same reasoning shows that $w_{\tau(i)} = v_{\sigma(\tau(i))}$ in general. Putting the left side all together, we get

$$\begin{aligned}(4.3.2)\quad &\big((\alpha \wedge \beta) \wedge \gamma\big)(v_1, v_2, v_3, v_4) \\ &\qquad = \frac{1}{12} \sum_{\sigma \in S_4} \operatorname{sgn}(\sigma) \sum_{\tau \in S_3} \operatorname{sgn}(\tau) \cdot \alpha(v_{\sigma(\tau(1))}) \cdot \beta(v_{\sigma(\tau(2))}, v_{\sigma(\tau(3))}) \cdot \gamma(v_{\sigma(4)}).\end{aligned}$$

Now each $\tau \in S_3$ can actually be thought of as a permutation $\tau' \in S_4$ for which $\tau'(4) = 4$; let's refer to this subgroup of S_4 as $S_{3,0}$. It's easy to see that $\operatorname{sgn}(\tau') =$

$\operatorname{sgn}(\tau)$, and the formula (4.3.2) rearranges, using $\operatorname{sgn}(\sigma)\cdot\operatorname{sgn}(\tau') = \operatorname{sgn}(\sigma\circ\tau')$ to

$$\begin{aligned}&\big((\alpha\wedge\beta)\wedge\gamma\big)(v_1,v_2,v_3,v_4)\\ &\quad= \frac{1}{12}\sum_{\tau'\in S_{3,0}}\sum_{\sigma\in S_4}\operatorname{sgn}(\sigma\circ\tau')\cdot\alpha(v_{\sigma\circ\tau'(1)})\cdot\beta(v_{\sigma\circ\tau'(2)},v_{\sigma\circ\tau'(3)})\cdot\gamma(v_{\sigma\circ\tau'(4)}).\end{aligned}$$

Fix a $\tau'\in S_{3,0}$ and consider the inside sum. Everything is in terms of $\sigma\circ\tau'$, so let $\rho=\sigma\circ\tau'$; the map $\sigma\mapsto\rho$ is a bijection from S_4 to itself and we obtain

(4.3.3)

$$\begin{aligned}\big((\alpha\wedge\beta)\wedge\gamma\big)(v_1,v_2,v_3,v_4) &= \frac{1}{12}\sum_{\tau'\in S_{3,0}}\sum_{\rho\in S_4}\operatorname{sgn}(\rho)\cdot\alpha(v_{\rho(1)})\cdot\beta(v_{\rho(2)},v_{\rho(3)})\cdot\gamma(v_{\rho(4)})\\ &= \frac{1}{2}\sum_{\rho\in S_4}\operatorname{sgn}(\rho)\cdot\alpha(v_{\rho(1)})\cdot\beta(v_{\rho(2)},v_{\rho(3)})\cdot\gamma(v_{\rho(4)}),\end{aligned}$$

where we notice in the last line that nothing depends on τ' anymore, and thus the sum over $S_{3,0}$ involves six identical terms.

The fact that the right side of (4.3.3) treats α and γ symmetrically means that the right side $\big(\alpha\wedge(\beta\wedge\gamma)\big)(v_1,v_2,v_3,v_4)$ must simplify to the same thing, and this gives associativity in this special case. ☺

The general case is more involved but uses the same basic concept of considering subgroups of the full permutation group that fix the first few or last few elements.

Proposition 4.3.5. *The wedge product is associative: if α is a j-form, β is a k-form, and γ is an l-form, then we have*

$$(\alpha\wedge\beta)\wedge\gamma = \alpha\wedge(\beta\wedge\gamma).$$

In addition, the wedge product is either commutative or anticommutative, depending on the size of the forms. If α is a j-form and β is a k-form, then

$$\alpha\wedge\beta = (-1)^{jk}\beta\wedge\alpha. \tag{4.3.4}$$

Proof. Associativity is important, but it's a bit tricky to prove. We expect it should somehow follow from the fact that the tensor product $\otimes$ is associative, so the first thing is to get the wedge product in terms of the tensor product.

The tool for doing this is the "alternation" operator $\mathcal{A}$, which takes an ordinary tensor B of type $(k,0)$ and gives a k-form $\mathcal{A}(B)$ by antisymmetrizing. We define its operation on k vectors $v_1,\cdots,v_k$ by

$$\mathcal{A}(B)(v_1,\cdots,v_k) = \sum_{\sigma\in S_k}\operatorname{sgn}(\sigma)B(v_{\sigma(1)},\cdots,v_{\sigma(k)}). \tag{4.3.5}$$

It is easy to see that if B is *already* antisymmetric, then all terms in (4.3.5) are the same, up to sign, so that $\mathcal{A}(B)=k!B$. It is also easy to see that the wedge product of a j-form α and a k-form β is

$$\alpha\wedge\beta = \frac{1}{j!k!}\mathcal{A}(\alpha\otimes\beta) \tag{4.3.6}$$

by formula (4.3.1).

So now to prove associativity, we want to show

$$(\alpha \wedge \beta) \wedge \gamma = \frac{1}{j!k!(j+k)!l!}\mathcal{A}\big(\mathcal{A}(\alpha \otimes \beta) \otimes \gamma\big) \\ = \frac{1}{j!k!(k+l)!l!}\mathcal{A}\big(\alpha \otimes \mathcal{A}(\beta \otimes \gamma)\big) = \alpha \wedge (\beta \wedge \gamma),$$

or essentially that

$$\frac{1}{(j+k)!}\mathcal{A}\big(\mathcal{A}(\alpha \otimes \beta) \otimes \gamma\big) = \frac{1}{(k+l)!}\mathcal{A}\big(\alpha \otimes \mathcal{A}(\beta \otimes \gamma)\big). \tag{4.3.7}$$

The basic idea is that we want to get rid of the $\mathcal{A}$ operators on the inside, then use associativity of $\otimes$ to finish it off. What allows us to do this is the fact that the outer antisymmetrization essentially takes care of antisymmetrizing the inner parts as well. The formal statement of this is the following lemma.

Lemma 4.3.6. *If B and C are tensors of type $(j,0)$ and $(k,0)$ respectively, and if $\mathcal{A}(C) = 0$, then $\mathcal{A}(B \otimes C) = 0$. Similarly if $\mathcal{A}(B) = 0$, then $\mathcal{A}(B \otimes C) = 0$.*

Proof. Knowing some finite group theory will help a lot in this proof. From the formula, we have

$$\mathcal{A}(B \otimes C)(v_1, \cdots, v_j, v_{j+1}, \cdots, v_{k+1}) \\ = \sum_{\sigma \in S_{j+k}} \operatorname{sgn}(\sigma) B(v_{\sigma(1)}, \cdots, v_{\sigma(j)}) C(v_{\sigma(j+1)}, \cdots, v_{\sigma(j+k)}).$$

The trick here is to consider the subgroup $S_{0,k}$ of S_{j+k} consisting of the permutations fixing the first j indices. It's obviously isomorphic to S_k. Furthermore, like any group, S_{j+k} can be partitioned into cosets of any subgroup like $S_{0,k}$. Each such coset is of the form $[\tau] = \tau S_{0,k} \equiv \{\tau \circ \chi \,|\, \chi \in S_{0,k}\}$, for some element $\tau \notin S_{0,k}$, and there must obviously be $|S_{j+k}|/|S_k| = (j+k)!/k!$ of these cosets. Let Q denote the set of cosets, and pick one representative $\tau_q \in S_{j+k}$ from each $q \in Q$.

We can now break up the sum over S_{j+k} into sums over the cosets, since every $\sigma \in S_{j+k}$ is $\sigma = \tau_q \circ \chi$ for some $q \in Q$ and $\chi \in S_{0,k}$. We use the fact that sgn is a homomorphism, $\operatorname{sgn}(\tau \circ \chi) = \operatorname{sgn}(\tau)\operatorname{sgn}(\chi)$:

$$\begin{aligned}
&\mathcal{A}(B \otimes C)(v_1, \cdots, v_j, v_{j+1}, \cdots, v_{k+1}) \\
&\quad = \sum_{\sigma \in S_{j+k}} \operatorname{sgn}(\sigma) B(v_{\sigma(1)}, \cdots, v_{\sigma(j)}) C(v_{\sigma(j+1)}, \cdots, v_{\sigma(j+k)}) \\
&\quad = \sum_{q \in Q} \sum_{\chi \in S_{0,k}} \operatorname{sgn}(\tau_q)\operatorname{sgn}(\chi) B(v_{\tau_q(\chi(1))}, \cdots, v_{\tau_q(\chi(j))}) C(v_{\tau_q(\chi(j+1))}, \cdots, v_{\tau_q(\chi(j+k))}) \\
&\quad = \sum_{q \in Q} \operatorname{sgn}(\tau_q) \sum_{\chi \in S_{0,k}} \operatorname{sgn}(\chi) B(v_{\tau_q(1)}, \cdots, v_{\tau_q(j)}) C(v_{\tau_q(\chi(j+1))}, \cdots, v_{\tau_q(\chi(j+k))})
\end{aligned}$$

where we used the fact that $\chi(1) = 1, \ldots, \chi(j) = j$ because by construction every χ fixes the first j elements. We can then pull the B term out of the χ sum to obtain

$$\mathcal{A}(B \otimes C)(v_1, \cdots, v_j, v_{j+1}, \cdots, v_{k+1}) = \sum_{q \in Q} \operatorname{sgn}(\tau_q) B(v_{\tau_q(1)}, \cdots, v_{\tau_q(j)}) \\ \cdot \left(\sum_{\chi \in S_{0,k}} \operatorname{sgn}(\chi) C(v_{\tau_q(\chi(j+1))}, \cdots, v_{\tau_q(\chi(j+k))}) \right). \tag{4.3.8}$$

We want to show that the term inside parentheses here is zero for any q in the coset quotient Q. Now suppose we go in the other direction and compute $\mathcal{A}(C)(v_{\tau_q(j+1)}, \cdots, v_{\tau_q(j+k)})$, which we know to be zero since $\mathcal{A}(C) = 0$. Rename $w_{j+1} = v_{\tau_q(j+1)}, \ldots, w_{j+k} = v_{\tau_q(j+k)}$. (Recall we are holding all the indices from 1 to j fixed and only permuting the indices $j+1$ through $j+k$.) Then

$$\mathcal{A}(C)(w_{j+1}, \cdots, w_{j+k}) = \sum_{\chi \in S_{0,k}} \operatorname{sgn}(\chi) C(w_{\chi(j+1)}, \cdots, w_{\chi(j+k)}).$$

Now $w_{\chi(j+i)} = v_{\tau_q(\chi(j+i))}$ for any $1 \leq i \leq k$, exactly as we worked out in Example 4.3.2. Thus we have

$$0 = \mathcal{A}(C)(v_{\tau_q(j+1)}, \cdots, v_{\tau_q(j+k)}) = \sum_{\chi \in S_{0,k}} \operatorname{sgn}(\chi) C(v_{\tau_q(\chi(j+1))}, \cdots, v_{\tau_q(\chi(j+k))}).$$

So the term inside parentheses in equation (4.3.8) is actually zero, and we conclude $\mathcal{A}(B \otimes C) = 0$.

If we suppose $\mathcal{A}(B) = 0$ instead, then we would just use the subgroup $S_{j,0}$ instead and get the same result. □

Now let's go back to the proof of Proposition 4.3.5. We saw before that if a tensor B of type $(j+k, 0)$ is already antisymmetric, then $\mathcal{A}(B) = (j+k)!B$. Therefore $\mathcal{A}(\mathcal{A}(\alpha \otimes \beta)) = (j+k)!\mathcal{A}(\alpha \otimes \beta)$, so that (combining)

$$\mathcal{A}\Big(\mathcal{A}(\alpha \otimes \beta) - (j+k)!(\alpha \otimes \beta)\Big) = 0.$$

Now that means we can apply Lemma 4.3.6 to get, for any γ, that

$$\mathcal{A}\Big(\big[\mathcal{A}(\alpha \otimes \beta) - (j+k)!(\alpha \otimes \beta)\big] \otimes \gamma\Big) = 0,$$

which (separating again) means

$$\mathcal{A}\big(\big[\mathcal{A}(\alpha \otimes \beta)\big] \otimes \gamma\big) = (j+k)!\mathcal{A}\big((\alpha \otimes \beta) \otimes \gamma\big).$$

Doing the same thing on the other side gives us

$$\mathcal{A}\big(\alpha \otimes \big[\mathcal{A}(\beta \otimes \gamma)\big]\big) = (k+l)!\mathcal{A}\big(\alpha \otimes (\beta \otimes \gamma)\big).$$

So plugging both of these formulas into (4.3.7) means we just have to check that

$$\mathcal{A}\big((\alpha \otimes \beta) \otimes \gamma\big) = \mathcal{A}\big(\alpha \otimes (\beta \otimes \gamma)\big),$$

and this is true since $\otimes$ is obviously associative. So after all that, we've finally proved that the wedge product is associative.

As a nice little cool-down, let's prove the graded anticommutativity formula (4.3.4). Recall from the definition that

$$(\alpha \wedge \beta)(v_1, \cdots, v_{j+k}) = \sum_{\sigma \in S_{j+k}} \operatorname{sgn}(\sigma) \alpha(v_{\sigma(1)}, \cdots, v_{\sigma(j)}) \beta(v_{\sigma(j+1)}, \cdots, v_{\sigma(j+k)})$$

while

$$(\beta \wedge \alpha)(v_1, \cdots, v_{j+k}) = \sum_{\tau \in S_{j+k}} \operatorname{sgn}(\tau) \alpha(v_{\tau(k+1)}, \cdots, v_{\tau(k+j)}) \beta(v_{\tau(1)}, \cdots, v_{\tau(j)}).$$

So to compare these, we want some correspondence between permutations σ and τ in S_{j+k} so that $\sigma(i) = \tau(k+i)$ for $1 \leq i \leq j$ and $\sigma(j+i) = \tau(i)$ for $1 \leq i \leq k$. If this is the case, then how are $\operatorname{sgn}(\sigma)$ and $\operatorname{sgn}(\tau)$ related?

Well, clearly the permutation $\kappa = \tau^{-1} \circ \sigma$ in S_{j+k} is $\kappa(i) = k + i$ for $1 \leq i \leq j$ and $\kappa(j + i) = i$ for $1 \leq i \leq k$, or

$$\begin{pmatrix} 1 & \cdots & j & j+1 & \cdots & j+k \\ k+1 & \cdots & k+j & 1 & \cdots & k \end{pmatrix}.$$

How many transpositions does this involve? Think of it this way: starting from the identity, we would need to slide the term $k + 1$ from the $k + 1$ place to the 1^{st} place, which requires k adjacent-entry transpositions. We then have to do the same thing for $k + 2$, to move it into the 2^{nd} place, which requires another k transpositions. There are j such terms we have to move, and each one requires k transpositions, so there are jk in total. Hence $\operatorname{sgn}(\kappa) = (-1)^{jk}$, and thus since sgn is a homomorphism, we have

$$\operatorname{sgn}(\sigma) = \operatorname{sgn}(\tau) \operatorname{sgn}(\kappa) = (-1)^{jk} \operatorname{sgn}(\tau).$$

The result $\alpha \wedge \beta = (-1)^{jk} \beta \wedge \alpha$ follows immediately. □

Despite all the work we've done above, we will only use associativity to write a basis for k-forms as $\alpha^{i_1} \wedge \cdots \wedge \alpha^{i_k}$ in terms of the basic 1-forms α^i. And since k-forms can always be written in terms of a wedge product basis of 1-forms (as we will see in a moment), virtually the only time we will need the graded anticommutativity (4.3.4) is for two 1-forms. Here's how this sort of thing generally works.

Example 4.3.7 (Wedge products in $\mathbb{R}^4$)**.** Consider $V = \mathbb{R}^4$ with standard basis $\{e_1, e_2, e_3, e_4\}$ and dual basis $\{\alpha^1, \alpha^2, \alpha^3, \alpha^4\}$. Let β be the 1-form $\beta = \alpha^1 + 2\,\alpha^3$ and let γ be the 2-form $\gamma = \alpha^1 \wedge \alpha^4 + 3\,\alpha^3 \wedge \alpha^4 - 2\,\alpha^1 \wedge \alpha^2$. Let's compute the 3-form $\beta \wedge \gamma$, using bilinearity, associativity, and anticommutativity.

$$\begin{aligned} \beta \wedge \gamma &= (\alpha^1 + 2\,\alpha^3) \wedge (\alpha^1 \wedge \alpha^4 + 3\,\alpha^3 \wedge \alpha^4 - 2\,\alpha^1 \wedge \alpha^2) \\ &= \alpha^1 \wedge (\alpha^1 \wedge \alpha^4) + 2\alpha^3 \wedge (\alpha^1 \wedge \alpha^4) + 3\alpha^1 \wedge (\alpha^3 \wedge \alpha^4) \\ &\qquad + 6\alpha^3 \wedge (\alpha^3 \wedge \alpha^4) - 2\alpha^1 \wedge (\alpha^1 \wedge \alpha^2) - 4\alpha^3 \wedge (\alpha^1 \wedge \alpha^2) \\ &= (\alpha^1 \wedge \alpha^1) \wedge \alpha^4 + 2(\alpha^3 \wedge \alpha^1) \wedge \alpha^4 + 3\alpha^1 \wedge \alpha^3 \wedge \alpha^4 \\ &\qquad 6(\alpha^3 \wedge \alpha^3) \wedge \alpha^4 - 2(\alpha^1 \wedge \alpha^1) \wedge \alpha^2 - 4(\alpha^1 \wedge \alpha^2) \wedge \alpha^3 \\ &= (-2 + 3)\alpha^1 \wedge \alpha^3 \wedge \alpha^4 - 4\alpha^1 \wedge \alpha^2 \wedge \alpha^3. \end{aligned}$$

Observe that if β is any 1-form, then $\beta \wedge \beta = 0$ by the anticommutativity, while if β is a 2-form, we can easily have $\beta \wedge \beta \neq 0$. For example, on $\mathbb{R}^4$ with the 2-form $\beta = \alpha^1 \wedge \alpha^2 + \alpha^3 \wedge \alpha^4$, we have

$$\beta \wedge \beta = (\alpha^1 \wedge \alpha^2 + \alpha^3 \wedge \alpha^4) \wedge (\alpha^1 \wedge \alpha^2 + \alpha^3 \wedge \alpha^4) = 2\alpha^1 \wedge \alpha^2 \wedge \alpha^3 \wedge \alpha^4.$$

☺

Now once we have a well-defined wedge product, we can express every k-form in terms of the wedge products of basic 1-forms, as we did in Examples 4.3.2 and 4.3.7 . This enables us to count the dimension of the space of k-forms. To get an idea for how this works, let's see how it works for 2-forms $\omega \in \Omega^2(V)$.

Take any basis $\{e_1, \cdots, e_n\}$ of V. Then we can write any u and v in V as $u = \sum_{i=1}^n u^i e_i$ and $v = \sum_{j=1}^n v^j e_j$. As a result, we have from bilinearity that

$$\omega(u, v) = \sum_{i=1}^n \sum_{j=1}^n u^i v^j \omega(e_i, e_j).$$

Thus the action of ω is determined by the numbers $\omega(e_i, e_j)$. Now observe that if $i = j$, then $\omega(e_i, e_j) = 0$ by antisymmetry; furthermore, if $j < i$, then $\omega(e_i, e_j) = -\omega(e_j, e_i)$. As a result, ω is completely determined by the numbers $\omega_{ij} \equiv \omega(e_i, e_j)$ where $i < j$. How many such terms are there? There are as many as pairs (i, j) with $1 \leq i < j \leq n$, and clearly there are $\binom{n}{2}$ of these. (Pick any two distinct numbers from $\{1, \cdots, n\}$ and disregard the order.)

Once you understand how it works on 2-forms, it's pretty easy to see how it must work for general k-forms. The notation is just a bit messier, but the idea is the same.

Proposition 4.3.8. *Suppose V is an n-dimensional vector space. The vector space $\Omega^k(V)$ of k-forms on V has dimension $\binom{n}{k}$. If $k = 0$, then the space of 0-forms has dimension 1, and if $k > n$, then there are no nontrivial k-forms.*

Proof. We will prove this by constructing an explicit basis for the k-forms. Take any basis $\{e_1, \ldots, e_n\}$ of V, and let $\{\alpha^1, \ldots, \alpha^n\}$ be the dual basis of V^*. Then we claim that the k-forms $\alpha^{i_1} \wedge \alpha^{i_2} \wedge \cdots \wedge \alpha^{i_k}$, where the indices i_1 range from 1 to n and satisfy $i_1 < i_2 < \cdots < i_k$, form a basis of the k-forms. Once we prove this, we just have to count how many such k-forms there are. It is easy to see that there must be $\binom{n}{k}$ of them: we merely need to choose k distinct elements of the set $\{1, \cdots, n\}$ to get the numbers i_1 through i_k, then order them; there are $\binom{n}{k}$ ways to do this. Since a k-form is a multilinear operator, we can compute the operation of any k-form ω on k vectors v_1 through v_k in V in terms of their coefficients in the basis $\{e_1, \ldots, e_n\}$.

So we just need to know, for any vectors $v_1, \cdots, v_k$,

$$\begin{aligned}\omega(v_1, v_2, \ldots, v_k) &= \sum_{i_1, i_2, \cdots, i_k = 1}^{n} v_1^{i_1} v_2^{i_2} \cdots v_k^{i_k} \omega(e_{i_1}, e_{i_2}, \cdots, e_{i_k}) \\ &= \sum_{i_1, i_2, \cdots, i_k = 1}^{n} \omega(e_{i_1}, e_{i_2}, \cdots, e_{i_k}) \alpha^{i_1} \wedge \alpha^{i_2} \wedge \cdots \wedge \alpha^{i_k}(v_1, v_2, \cdots, v_k).\end{aligned}$$

Thus we must have

$$\omega = \sum_{i_1, i_2, \cdots, i_k = 1}^{n} \omega(e_{i_1}, e_{i_2}, \cdots, e_{i_k}) \alpha^{i_1} \wedge \alpha^{i_2} \wedge \cdots \wedge \alpha^{i_k}.$$

Now not all basic k-forms $\alpha^{i_1} \wedge \cdots \alpha^{i_k}$ are distinct. First of all, this basic k-form is nonzero if and only if the numbers $\{i_1, \ldots, i_k\}$ are all distinct. Furthermore, if the sets $\{i_1, \cdots, i_k\}$ and $\{j_1, \cdots, j_k\}$ contain the same k elements, then the j's must be a permutation of the i's, and thus

$$\alpha^{i_1} \wedge \cdots \wedge \alpha^{i_k} = \pm \alpha^{j_1} \wedge \cdots \wedge \alpha^{j_k},$$

depending on the sign of that permutation.

As a result, the number of distinct basic k-forms $\alpha^{i_1} \wedge \cdots \wedge \alpha^{i_k}$ is the same as the number of ways to choose k numbers out of $\{1, \cdots, n\}$, up to reorderings. This is precisely $\binom{n}{k}$.

The only other thing to prove is that if $k > n$, then there are no nontrivial k-forms, and this follows from the fact that any k-form is determined by its action on k different vectors. Now if $k > n$, then any k vectors *must* be linearly dependent, and thus the action of a k-form must be zero on them. (If any argument appears twice in a k-form, we must get zero by antisymmetry). □

As mentioned, if you understand how it works for 2-forms, then it's not at all surprising how it works for higher forms. But what's interesting is that since $\binom{n}{n} = 1$, there is only *one* nontrivial n-form on an n-dimensional vector space, in the sense that any n-form must be a multiple of any other one.

Now in general, if $T\colon V \to W$, then there is an induced transformation (the pull-back) defined as a map from $\Omega^k(W)$ to $\Omega^k(V)$, in much the same way that $T^*\colon W^* \to V^*$ is defined by Definition 4.1.4.

Definition 4.3.9. If $T\colon V \to W$ is a linear transformation, we define the pull-back $T^*\colon \Omega^k(W) \to \Omega^k(V)$ by the following operation: for any k-form ω on W, we define $T^*\omega \in \Omega^k(V)$ by

$$T^*\omega(v_1, \cdots, v_k) = \omega\big(T(v_1), \cdots, T(v_k)\big). \tag{4.3.9}$$

Observe how natural this definition is; it's the only way we could relate forms on one space to forms on another space. The operator T^* *has* to go backwards.

Now here's something that will be very useful much later. Suppose V is an n-dimensional vector space, and that $T\colon V \to V$ is a linear transformation. There is only one n-form on V (up to multiplication by scalars), so that if μ is an n-form on V, then $T^*\mu$ must be $c\mu$ for some constant c. Once you understand the effect of antisymmetry, the following proposition shouldn't be too surprising.

Proposition 4.3.10. *If $T\colon V \to V$ is a linear transformation and μ is an n-form on V, then $T^*\mu = (\det T)\mu$.*

Proof. Take any basis $\{e_1, \cdots, e_n\}$ of V and let $\{\alpha^1, \cdots, \alpha^n\}$ be the dual basis of V^*. Write $T(e_i) = \sum_{j=1}^n T_i^j e_j$. Then $T^*\mu$, as an n-form, is completely determined by its operation on the vectors e_1 through e_n in order. We have

$$\begin{aligned} T^*\mu(e_1, \cdots, e_n) &= \mu\big(T(e_1), \cdots, T(e_n)\big) \\ &= \sum_{j_1=1}^n \cdots \sum_{j_n=1}^n \mu\big(T_1^{j_1} e_{j_1}, \cdots, T_n^{j_n} e_{j_n}\big) \\ &= \sum_{j_1=1}^n \cdots \sum_{j_n=1}^n T_1^{j_1} \cdots T_n^{j_n} \mu\big(e_{j_1}, \cdots, e_{j_n}\big). \end{aligned}$$

Now by antisymmetry, the term $\mu(e_{j_1}, \cdots, e_{j_n})$ is zero if any two of the j_k's are the same. Hence we only get a nonzero term if $(j_1, \cdots, j_n)$ is a permutation of $(1, \cdots, n)$. Furthermore, if it is a permutation $j_k = \sigma(k)$, then we must have

$$\mu(e_{j_1}, \cdots, e_{j_n}) = \mu(e_{\sigma(1)}, \cdots, e_{\sigma(n)}) = \operatorname{sgn}(\sigma)\mu(e_1, \cdots, e_n).$$

As a result, the formula becomes

$$T^*\mu(e_1, \cdots, e_n) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma)\, T_1^{\sigma(1)} \cdots T_n^{\sigma(n)} \mu(e_1, \cdots, e_n) = \mu(e_1, \cdots, e_n)(\det T),$$

by the definition (3.3.1). As a result, we must have $T^*\mu = (\det T)\,\mu$. □

5. Multivariable calculus

"The dark side of the Force is a pathway to many abilities some consider to be unnatural."

Everything we will mention in this Chapter is proved in undergraduate multivariable real analysis, so we will give only the ideas of proofs, freely specializing to the 2-dimensional case whenever it's at all convenient. We will also never care about the sharpest or strongest possible results: many theorems are true if the function is only assumed continuous or once-differentiable, but in differential geometry all functions are C^∞. Whenever this simplifies the statement of the theorem or the proof, I will assume it. If anything looks unfamiliar to you, make sure you review it from another textbook. (Munkres' "Analysis on Manifolds" and Spivak's "Calculus on Manifolds" both cover these theorems and proofs in detail.) This material is what ends up being most important for differential geometry.

The subject of interest is functions $F\colon \mathbb{R}^n \to \mathbb{R}^k$, as given explicitly by some formulas

$$F(x_1, \cdots, x_n) = \big(f_1(x_1, \cdots, x_n), \cdots, f_k(x_1, \cdots, x_n)\big).$$

This will be our interest as well, although we will have a very different perspective on it. (For this Chapter, we will index every object by subscripts. Later on, we will start using both subscripts and superscripts for different purposes, as is common in differential geometry.)

5.1. **Derivatives.**

Definition 5.1.1. A function $F\colon \mathbb{R}^n \to \mathbb{R}^k$ is called *continuous everywhere* if, for every $a \in \mathbb{R}^n$, we have $\lim_{h\to 0} F(a+h) = F(a)$.

Definition 5.1.2. A function $F\colon \mathbb{R}^n \to \mathbb{R}^k$ is called *differentiable everywhere* if, for every $a \in \mathbb{R}^n$, there is a linear map $DF(a)\colon \mathbb{R}^n \to \mathbb{R}^k$ such that

$$\lim_{h\to 0} \frac{|F(a+h) - F(a) - DF(a)h|}{|h|} = 0.$$

This linear map $DF(a)$ is unique and is called the *derivative of F at a*.

Theorem 5.1.3. *If $F = (f_1, \cdots, f_k)\colon \mathbb{R}^n \to \mathbb{R}^k$ and $DF(a)$ exists at $a = (a_1, \cdots, a_n)$ in the sense of Definition 5.1.2, then the partial derivatives*

$$\frac{\partial f_j}{\partial x_i}\Big|_{x=a} \equiv \lim_{h\to 0} \frac{f_j(a_1, \cdots, a_i + h, \cdots, a_n) - f_j(a_1, \cdots, a_i, \cdots, a_n)}{h}$$

exist for every $i \in \{1, \cdots, n\}$ and every $j \in \{1, \cdots, k\}$. Furthermore, we can write the linear operator $DF(a)$ as the $k \times n$ matrix

$$DF(a) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}\Big|_{x=a} & \cdots & \frac{\partial f_1}{\partial x_n}\Big|_{x=a} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_k}{\partial x_1}\Big|_{x=a} & \cdots & \frac{\partial f_k}{\partial x_n}\Big|_{x=a} \end{pmatrix}. \tag{5.1.1}$$

I will skip the proof.

Example 5.1.4. Let $F\colon \mathbb{R}^3 \to \mathbb{R}^2$ be defined by the formula

$$F(x,y,z) = (x^3 - 3zy^2, 4x^2 + y^2)$$

and let $a = (0,1,0)$. Then

$$DF(a) = \begin{pmatrix} 3x^2 & 6yz & -3y^2 \\ 8x & 2y & 0 \end{pmatrix}\Big|_{(x,y,z)=(0,1,0)} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 2 & 0 \end{pmatrix}.$$

This matrix has rank one by Proposition 3.3.6: every 2×2 submatrix has determinant zero, while there is a 1×1 submatrix that has nonzero determinant. ☺

Theorem 5.1.5. *Suppose $F\colon \mathbb{R}^n \to \mathbb{R}^k$. If for every $i \in \{1, \cdots, n\}$, the i^{th} partial derivative function*

$$\begin{aligned} \frac{\partial F}{\partial x_i}(x_1, \cdots, x_n) &\equiv D_i F(x_1, \cdots, x_n) \\ &\equiv \lim_{h \to 0} \frac{F(x_1, \cdots, x_i + h, \cdots, x_n) - F(x_1, \cdots, x_i, \cdots, x_n)}{h} \end{aligned}$$

exists and is a continuous function (in the sense of Definition 5.1.1), then F is differentiable everywhere in the sense of Definition 5.1.2.

I'll skip this proof as well.

Definition 5.1.6. A function $F\colon \mathbb{R}^n \to \mathbb{R}^k$ is called C^∞ or *smooth* if all iterated partial derivatives $\frac{\partial^m F}{\partial x_{i_1} \partial x_{i_2} \cdots \partial x_{i_m}} \equiv D_{i_1} D_{i_2} \cdots D_{i_m} F$ exist and are continuous functions from $\mathbb{R}^n$ to $\mathbb{R}^k$.

Remark 5.1.7. The partial derivative is often written as

$$\frac{\partial f}{\partial x}(x,y) = f_x(x,y),$$

or simply as f_x if it's clear where we are evaluating. Similarly in higher dimensions we often write

$$\frac{\partial F}{\partial x_i}(x_1, \cdots, x_n) = F_{x_i}(x_1, \cdots, x_n).$$

We also occasionally write formulas such as

$$\partial_{x_i} F = \frac{\partial F}{\partial x_i}$$

if the subscript notation could cause confusion. If we have more than one partial derivative, we write

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial}{\partial x}\frac{\partial f}{\partial y} = \partial_x(f_y) = f_{yx}.$$

Note that the order change looks like a cheat, but what's happening is that f is getting differentiated first with respect to y, then with respect to x, so the operator that differentiates f with respect to y should appear closer to f.

Many of the theorems of multivariable calculus depend on having a certain number of derivatives being continuous, and become false if not enough derivatives are continuous. Furthermore, there are functions which have partial derivatives existing but not continuous. There are also functions which have partial derivatives at a point but are not differentiable at that point. These pathological features are interesting in analysis, where one likes to do as many things as possible without assuming much smoothness. However in differential geometry, there is almost never

any reason to deal with any functions that are not C^∞. As such we will immediately specialize to assuming this.

In one-variable real analysis, there is essentially only one serious theorem that gives results about derivatives: the Mean Value Theorem. There is no analogous result in higher dimensions. Hence for many purposes it is both convenient and necessary to reduce questions about higher-dimensional situations to one-dimensional calculus, where everything is very well understood. This continues to be true in differential geometry: we will find that often the best way to understand a high-dimensional manifold M is to understand the functions from reals to M or functions from M to reals, and this philosophy allows us to reduce many difficult things to simple one-dimensional calculus.

Theorem 5.1.8 (Mean Value Theorem)**.** *If $f\colon [a,b] \to \mathbb{R}$ is continuous on $[a,b]$ and differentiable on (a,b), then there is a point $c \in (a,b)$ such that*

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Proof. The idea is to subtract a linear function L from f so that $g = f - L$ satisfies $g(a) = g(b) = 0$, then check that it's sufficient to prove there is a c with $g'(c) = 0$. Either g has an interior extreme value where its derivative is zero, or must be identically zero. □

The following theorem on equality of mixed partials ends up being extremely important, and we will refer to it often.

Theorem 5.1.9 (Commuting mixed partials)**.** *If F is C^∞, then for any indices i and j, the mixed partials are equal:*

$$\frac{\partial^2 F}{\partial x_j \partial x_i} = \frac{\partial^2 F}{\partial x_i \partial x_j}.$$

Proof. For simplicity assume there are only two variables, x and y; then by definition of the partials we have

$$\begin{aligned} F_{yx}(a,b) &= \lim_{h\to 0} \frac{F_y(a+h,b) - F_y(a,b)}{h} \\ &= \lim_{h\to 0} \lim_{k\to 0} \frac{F(a+h,b+k) - F(a+h,b) - F(a,b+k) + F(a,b)}{hk}, \\ F_{xy}(a,b) &= \lim_{k\to 0} \frac{F_x(a,b+k) - F_x(a,b)}{k} \\ &= \lim_{k\to 0} \lim_{h\to 0} \frac{F(a+h,b+k) - F(a,b+k) - F(a+h,b) + F(a,b)}{hk}. \end{aligned} \tag{5.1.2}$$

So we are trying to prove that

$$\lim_{h\to 0} \lim_{k\to 0} \frac{G(h,k)}{hk} = \lim_{k\to 0} \lim_{h\to 0} \frac{G(h,k)}{hk},$$

where

$$G(h,k) = F(a+h,b+k) - F(a,b+k) - F(a+h,b) + F(a,b)$$

and a and b are held constant. Notice we're just taking the limit of the same quantity in two different ways.[1] To make this easier, we're going to use the Mean Value Theorem (in one variable) to reduce a difference to a derivative.

For a fixed value of k, define a new one-variable function by $f(x) = F(x, b+k) - F(x,b)$; then it's easy to see that $f(a+h) - f(a) = G(h,k)$. By the Mean Value Theorem 5.1.8, we know that $f(a+h) - f(a) = hf'(\alpha)$ for some α with $a < \alpha < a+h$. Since $f'(\alpha) = F_x(\alpha, b+k) - F_x(\alpha, b)$ and $G(h,k) = f(a+h) - f(a)$, we conclude

$$G(h,k) = hF_x(\alpha, b+k) - hF_x(\alpha, b),$$

for some α (which depends on both h and k and is between a and $a+h$).

Now for fixed α, consider the function $g(y) = hF_x(\alpha, y)$. Then $G(h,k) = g(b+k) - g(b)$. Again using the Mean Value Theorem, we know there is a β with $b < \beta < b+k$ and $g(b+k) - g(b) = kg'(\beta)$. Since $g'(\beta) = hF_{xy}(\alpha, \beta)$, we finally obtain

$$G(h,k) = g(b+k) - g(b) = kg'(\beta) = hkF_{xy}(\alpha, \beta), \tag{5.1.3}$$

where $a < \alpha < a+h$ and $b < \beta < b+k$.

If we do the exact same procedure but in the other order[2], we would get $G(h,k) = khF_{yx}(\gamma, \delta)$ where $a < \gamma < a+h$ and $b < \delta < b+k$. But we don't need this, because we have a shortcut: we use (5.1.3) in the second line of (5.1.2) to relate F_{yx} directly to F_{xy}:

$$F_{yx}(a,b) = \lim_{h\to 0} \lim_{k\to 0} \frac{G(h,k)}{hk} = \lim_{h\to 0} \lim_{k\to 0} F_{xy}(\alpha, \beta). \tag{5.1.4}$$

Notice that we don't know anything about the dependence of α or β on h or k (they might not be continuous, for example), but it doesn't matter: since $b < \beta < b+h$ always, we have $\lim_{k\to 0} F_{xy}(\alpha, \beta) = F_{xy}(\alpha, b)$ by the squeeze theorem, since F_{xy} is continuous in each variable. Similarly we have $\lim_{h\to 0} F_{xy}(\alpha, b) = F_{xy}(a,b)$, and using this in (5.1.4) we get what we were looking for. □

The Chain Rule 5.1.10 is also extremely useful, especially its Corollary 5.1.12. We will use it again and again when proving that a particular expression is independent of coordinates or in changing expressions from one coordinate system to another. Get very, very familiar with it: the first step to getting good at differential geometry is knowing the Chain Rule by heart.[3]

Theorem 5.1.10 (The Chain Rule)**.** *Suppose $F = (f_1, \cdots, f_m) \colon \mathbb{R}^n \to \mathbb{R}^m$ and $G = (g_1, \cdots, g_k) \colon \mathbb{R}^m \to \mathbb{R}^k$ are both smooth. Then $G \circ F \colon \mathbb{R}^n \to \mathbb{R}^k$ is also smooth, and its derivative at any point $a \in \mathbb{R}^n$ can be computed by $D(G \circ F)(a) = DG\big(F(a)\big) \cdot DF(a)$. Thus its partial derivatives can be computed by*

$$\frac{\partial (G \circ F)_j}{\partial x_i} = \sum_{\ell=1}^{m} \frac{\partial g_j}{\partial y_\ell}\Big|_{y=F(x)} \frac{\partial f_\ell}{\partial x_i}. \tag{5.1.5}$$

Example 5.1.11. Let $\gamma \colon \mathbb{R} \to \mathbb{R}^2$ be the curve $\gamma(t) = (2\cos t, \sin t)$, and let $F \colon \mathbb{R}^2 \to \mathbb{R}$ be $F(x,y) = 3xy^2 - 2x^3$. Then $F \circ \gamma \colon \mathbb{R} \to \mathbb{R}$ is

$$F \circ \gamma(t) = 6\cos t \sin^2 t - 16 \cos^3 t,$$

[1] It shouldn't seem too obvious that this works: for example $\lim\limits_{h\to 0^+} \lim\limits_{k\to 0^+} h^k \neq \lim\limits_{k\to 0^+} \lim\limits_{h\to 0^+} h^k$.

[2] Do it yourself; you won't really understand this proof unless you do.

[3] No joke.

so

$$(F \circ \gamma)'(t) = 60 \sin t \cos^2 t - 6 \sin^3 t.$$

On the other hand, we have

$$DF\big(\gamma(t)\big) = \begin{pmatrix} 3y^2 - 6x^2 & 6xy \end{pmatrix}\Big|_{(x,y)=(2\cos t, \sin t)} = \begin{pmatrix} 3\sin^2 t - 24\cos^2 t & 12 \cos t \sin t \end{pmatrix}$$

along with

$$\gamma'(t) = \begin{pmatrix} -2\sin t \\ \cos t \end{pmatrix},$$

so it is easy to check here that

$$(F \circ \gamma)'(t) = DF\big(\gamma(t)\big) \cdot \gamma'(t).$$

☺

Corollary 5.1.12. *If $x = (x_1, \cdots, x_n) \colon \mathbb{R} \to \mathbb{R}^n$ and $f \colon \mathbb{R}^n \to \mathbb{R}$ are both smooth, then so is $f \circ x \colon \mathbb{R} \to \mathbb{R}$, and we have*

$$(f \circ x)'(t) = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}\big(x_1(t), \cdots, x_n(t)\big) \frac{dx_j}{dt}.$$

Proof. Theorem 5.1.10 is proved in more or less the same way as the Chain Rule for functions of a single variable: we simply end up multiplying the derivative matrices (5.1.1) together rather than multiplying the derivatives together. The proof of the Corollary is essentially the same (since we are only differentiating with respect to one variable at a time, we might as well consider functions of only one variable), and it makes the notation simpler. We will also assume $n = 2$ to simplify the notation.

So assume $F \colon \mathbb{R}^2 \to \mathbb{R}$ and we have a curve given by $x = \eta(t)$ and $y = \xi(t)$. We want to prove that the function $f(t) = F\big(\eta(t), \xi(t)\big)$ has

$$f'(t) = \frac{\partial F}{\partial x}\big(\eta(t), \xi(t)\big) \frac{d\eta}{dt} + \frac{\partial F}{\partial y}\big(\eta(t), \xi(t)\big) \frac{d\xi}{dt}.$$

Like so many proofs in analysis, it's basically just the old add-and-subtract trick.

$$\begin{aligned} f'(t) &= \lim_{h \to 0} \frac{f(t+h) - f(t)}{h} = \lim_{h \to 0} \frac{F\big(\eta(t+h), \xi(t+h)\big) - F\big(\eta(t), \xi(t)\big)}{h} \\ &= \lim_{h \to 0} \frac{F\big(\eta(t+h), \xi(t+h)\big) - F\big(\eta(t), \xi(t+h)\big)}{h} \\ &\qquad + \lim_{h \to 0} \frac{F\big(\eta(t), \xi(t+h)\big) - F\big(\eta(t), \xi(t)\big)}{h}. \end{aligned}$$

The term on the last line is obviously $F_y(\eta(t), \xi(t))\xi'(t)$ by the one-dimensional chain rule. To prove the term on the middle line is $F_x(\eta(t), \xi(t))\eta'(t)$, it's sufficient to prove

(5.1.6)

$$\lim_{h \to 0} \frac{F(\eta(t+h), \xi(t+h)) - F(\eta(t), \xi(t+h)) - F(\eta(t+h), \xi(t)) + F(\eta(t), \xi(t))}{h} = 0.$$

(Think about why.)

The technique to prove (5.1.6) is almost exactly the same as the technique in the proof of Theorem 5.1.9—just use the Mean Value Theorem once in each variable to reduce it to $\lim_{h \to 0} hF_{xy}(\alpha, \beta)$, which is zero since F has continuous second partial derivatives. We'll skip the details. □

The Product Rule (or Leibniz Rule) will be somewhat less useful to us than the Chain Rule, since it only works on real-valued functions. However in some sense it characterizes first-order differential operators, and we will need this.

Theorem 5.1.13 (The Product Rule)**.** *If $f\colon \mathbb{R}^n \to \mathbb{R}$ and $g\colon \mathbb{R}^n \to \mathbb{R}$ are both smooth, then so is $fg\colon \mathbb{R}^n \to \mathbb{R}$, and furthermore the partial derivatives satisfy*

$$\frac{\partial(fg)}{\partial x_i}\Big|_{x=a} = f(a)\,\frac{\partial g}{\partial x_i}\Big|_{x=a} + g(a)\,\frac{\partial f}{\partial x_i}\Big|_{x=a}.$$

Proof. The proof is exactly the same as in the single-variable case, involving adding and subtracting a term in the definition of the derivative. □

5.2. **The Contraction Lemma and consequences.** The Contraction Mapping Lemma will not be directly important to us, although it is used to prove both Theorems 5.2.4 and 5.2.6, which will be.

Lemma 5.2.1 (Contraction Mapping Lemma)**.** *Suppose M is a complete metric space and $f\colon M \to M$ is a contraction mapping, i.e., for some positive number $r < 1$, we have*

$$d\big(f(x), f(y)\big) \le r d(x,y) \qquad \text{for all } x, y \in M. \tag{5.2.1}$$

Then there is a unique point $z \in M$ for which $f(z) = z$. (Such a z is called a fixed point *of f.*

Proof. The basic idea of the proof is to start with an arbitrary point $z_o \in M$ and keep applying f to get a recursive sequence $z_{n+1} = f(z_n)$. The contraction property (5.2.1) implies that (z_n) will be a Cauchy sequence (by comparison with the geometric series in r), and thus it must converge since M is complete. Then if $z = \lim_{n\to\infty} z_n$, the equation $z_{n+1} = f(z_n)$ for all n and the continuity of f imply $z = f(z)$.

The only tricky bit is to prove that z_n is a Cauchy sequence. The trick is to relate not z_{n+1} to z_n, but rather relate the distance $d(z_{n+2}, z_{n+1})$ to the distance $d(z_{n+1}, z_n)$. We have

$$d(z_{n+2}, z_{n+1}) = d\big(f(z_{n+1}), f(z_n)\big) \le r d(z_{n+1}, z_n).$$

Inductively this implies $d(z_{n+1}, z_n) \le Ar^n$ where $A = d(z_0, z_1)$.

Hence by the triangle inequaity, for any integer k and any $m > k$ we have

$$d(z_m, z_k) \le \sum_{n=k}^{m-1} d(z_n, z_{n+1}) \le \sum_{n=k}^{m-1} Ar^n = \frac{A(r^k - r^m)}{1-r}.$$

If both m and k are sufficiently large, this can be made as small as we want, so (z_n) is a Cauchy sequence.

□

The Contraction Mapping Lemma is useful mainly because we don't have to assume anything about M except that it's complete. It gets used most often in infinite-dimensional spaces, like spaces of functions, where compactness usually fails but completeness still works.

Now we come to two theorems which are generally proved in multivariable calculus, but which don't really get used until differential geometry: the Inverse and Implicit Function Theorems. Either one is a consequence of the other.

- The Inverse Function Theorem is used mainly to relate coordinate changes, to put them all on an equal footing. (For example, in Chapter 6, we will use it to go back and forth between rectangular coordinates and polar coordinates.)
- The Implicit Function Theorem is basically used to solve equations like $F(x,y) = 0$ for y in terms of x, and most importantly, to establish that the solution $y(x)$ depends smoothly on x. Although the statement of it is rather involved, it will later make it very easy to prove certain sets are smooth manifolds.

We will prove the Implicit Function Theorem first. The condition is stated in a more cumbersome way so that we can get a more explicit result; the point is that the rank of $DF(a,b)$, as a linear operator from $\mathbb{R}^{n+k}$ to $\mathbb{R}^k$, is maximal (i.e., equal to k). If it is, then by Proposition 3.3.6, there is some $k \times k$ submatrix which is nonsingular, and we can reorder the variables without loss of generality so that the leftmost $k \times k$ submatrix is nonsingular.

Theorem 5.2.2 (Implicit Function Theorem)**.** *Suppose $F = (f_1, \cdots, f_k) \colon \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^k$ is smooth, and that $F(a,b) = 0$ for some $a \in \mathbb{R}^n$ and $b \in \mathbb{R}^k$. Furthermore suppose that the function $f \colon \mathbb{R}^k \to \mathbb{R}^k$ defined by $f(y) = F(a,y)$ has $Df(b)$ invertible.*

Then there is an open set $V \subset \mathbb{R}^n$ with $a \in V$ and a smooth function $G \colon V \to \mathbb{R}^k$ such that $G(a) = b$ and $F\big(x, G(x)\big) = 0$ for every $x \in V$.

Proof. We are trying to solve $F(x,y) = 0$ for y in terms of x, and then we'll call it $y = G(x)$. The idea is basically to set up the solution y of $F(x,y) = 0$ via Newton's method. Write the Taylor polynomial of F as $F(x,y) = C(x-a)+D(y-b)+R(x,y)$, where C is the $n \times k$ matrix of partials in the x direction, and D is the $k \times k$ matrix of partials in the y direction (assumed to be invertible), and the remainder R satisfies

$$\lim_{(x,y)\to(a,b)} \frac{R(x,y)}{\sqrt{(x-a)^2 + (y-b)^2}} = 0.$$

Then the solution y satisfies $y = K(x,y)$ where

$$K(x,y) = b - D^{-1}\big(C(x-a) + R(x,y)\big),$$

and this is a fixed-point problem for y. When x and y are close to a and b, we expect $R(x,y)$ to be small, so that for x fixed, the function $y \mapsto K(x,y)$ should be a contraction in the y-variables. Then we can use the Contraction Mapping Lemma 5.2.1 to get a unique solution y, and we worry about how exactly y depends on x later. So for the time being we will hold x fixed and make sure we can solve for y.

First we show $y \mapsto K(x,y)$ is a contraction. Let y_1 and y_2 be in $\mathbb{R}^k$, and define $h(t) = K(x, ty_1 + (1-t)y_2)$, the values of K along the straight-line segment between y_1 and y_2. Then by the one-variable Mean Value Theorem 5.1.8, we know

$$K(x,y_1) - K(x,y_2) = h(1) - h(0) = h'(\tau) = D^{-1} R_y(x,\xi)(y_1 - y_2),$$

where $0 < \tau < 1$ and $\xi = \tau y_1 + (1-\tau)y_2$.

Now since $F_y(x,y) = D + R_y(x,y)$ and $D = F_y(a,b)$, we know $R_y(a,b) = 0$. So by continuity we can make $R_y(x,y)$ as small as we want if x and y are sufficiently close to a and b. Let's choose $\delta > 0$ and $\varepsilon > 0$ such that we have

$$|K(x,y_1) - K(x,y_2)| \leq \frac{1}{2}|y_1 - y_2| \text{ whenever } |x-a| < \delta \text{ and } |y-b| \leq \varepsilon. \tag{5.2.2}$$

This comes from making $|R_y(x,y)| \leq 1/(2|D^{-1}|)$. Notice that the condition on x is an open set while the condition on y is a closed set: I need this because I want to work in a *complete* space and the open ball is not complete.

Then for any fixed x within δ of a, the map $y \mapsto K(x,y)$ is a contraction when restricted to the closed ball $y \in \overline{B}_\varepsilon(b)$. Hence since the closed ball is complete, we can use the Contraction Mapping Lemma 5.2.1 to get a unique fixed point $y \in \overline{B}_\varepsilon(b)$ satisfying $y = K(x,y)$ (and hence $F(x,y) = 0$). Let's call this fixed point $G(x)$. So we've proved there is a unique solution $G(x)$ of $F(x,G(x)) = 0$ for all x with $|x-a| < \delta$.

The only thing left to prove is that G is smooth. At the moment we don't even know G is continuous, but since $G(x) = K(x,G(x))$, we see that (using the old add-and-subtract trick)

$$\begin{aligned}|G(x_1) - G(x_2)| &= |K(x_1,G(x_1)) - K(x_2,G(x_2))| \\ &\leq |K(x_1,G(x_1)) - K(x_2,G(x_1))| + |K(x_2,G(x_1)) - K(x_2,G(x_2))| \\ &\leq M|x_1 - x_2| + \tfrac{1}{2}|G(x_1) - G(x_2)|,\end{aligned}$$

where we use equation (5.2.2), and the Mean Value Theorem for $x \mapsto K(x,G(x_1))$, setting M to be the supremum of $K_x\big(x,G(y)\big)$ over the set $x \in B_\delta(a) \times B_\varepsilon(b)$. At first this doesn't seem to help since it looks like circular reasoning, but solving the inequality for $|G(x_1) - G(x_2)|$ we get

$$|G(x_1) - G(x_2)| \leq 2M|x_1 - x_2|.$$

Hence G is Lipschitz, so it's continuous.

Since $F(x,G(x)) = 0$ for all $x \in B_\delta(a)$, we can compute the derivative of G using the Chain Rule (Theorem 5.1.10). We get

$$F_x(x,G(x)) + F_y(x,G(x))DG(x) = 0, \tag{5.2.3}$$

and we know $F_y(x,G(x))$ is nonsingular if x is sufficiently close to a. Thus we can solve to figure out what $DG(x)$ has to be, and once we have the candidate for the derivative, we can prove this must actually *be* the derivative using the definition. We can iterate the Chain Rule and keep computing higher derivatives of G in order to prove that it is C^∞ as long as F is; if you work out a couple for yourself, you'll see the only thing you ever have to worry about to solve for the higher derivatives is that F_y is invertible. □

Example 5.2.3. The simplest situation is the circle: suppose $F\colon \mathbb{R}^2 \to \mathbb{R}$ with $F(x,y) = x^2 + y^2 - 1$. Then $DF(a,b) = (2a \quad 2b)$. The hypotheses in Theorem 5.2.2 are that $F(a,b) = 0$ and $F_y(a,b) \neq 0$, which means that $a^2 + b^2 = 1$ and $b \neq 0$. Hence in a neighborhood of any a other than $a = \pm 1$, there is an open interval V containing a and a smooth function $G\colon V \to \mathbb{R}$ such that $F(x,G(x)) = 0$. In this case, of course, we know exactly what it is: the largest set V is $(-1,1)$ and $G(x) = \sqrt{1-x^2}$ or $G(x) = -\sqrt{1-x^2}$, depending on whether $b > 0$ or $b < 0$. Of course, when $a = 1$ or $a = -1$, there is *no* function $G(x)$ defined on an open interval containing a such that $F(x,G(x)) = 0$. So the nondegeneracy assumption $DF_y(a,b) \neq 0$ really is essential.

Now let's imagine we didn't know what $G(x)$ was. How would we compute $G'(x)$, $G''(x)$, etc.? First of all, $F(x,G(x)) = 0$ implies $x^2 + G(x)^2 - 1 = 0$ for all x in the open set V. Differentiating this as in equation (5.2.3), we get $2x + 2G(x)G'(x) = 0$,

which gives $G'(x) = -x/G(x)$. Since $G(x) \neq 0$ by assumption (that's exactly what the nondegeneracy condition translates into here), we know what $G'(x)$ must be.

How about higher derivatives? Differentiating the equation $x + G(x)G'(x) = 0$ again gives $1 + G'(x)^2 + G(x)G''(x) = 0$, which we could obviously solve for $G''(x)$ since we know $G'(x)$. Another differentiation yields $3G'(x)G''(x)+G(x)G'''(x) = 0$, and what you should notice is that no matter how many times we differentiate, the coefficient of the highest derivative is $G(x)$. So as long as $G(x) \neq 0$, we can solve for $G''(x)$, $G'''(x)$, etc. This happens in general: the nondegeneracy condition that $DF_y(x, y)$ is necessary for $DG(x)$ to exist, but once we have that, the same condition will ensure that all higher derivatives of G exist for free. ☺

The Inverse Function Theorem is a really easy consequence of the Implicit Function Theorem.

Theorem 5.2.4 (Inverse Function Theorem)**.** *If $F\colon U \subset \mathbb{R}^n \to \mathbb{R}^n$ is a smooth function in an open set U, and if for some $a \in U$ the derivative $DF(a)$ is an invertible matrix, then there is an open set $V \subset \mathbb{R}^n$ containing $F(a)$ and a function $G\colon V \subset \mathbb{R}^n \to \mathbb{R}^n$ which is smooth and has $F\big(G(y)\big) = y$ and $G\big(F(x)\big) = x$ for all $y \in V$ and all $x \in G[V]$.*

Proof. Define $F\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ by $F(x, y) = f(y) - x$. Then you can check (and you really should!) that all the hypotheses of the Implicit Function Theorem 5.2.2 are satisfied, and that the theorem gives you what you're looking for. □

Example 5.2.5. I mentioned above that one important way we use the Inverse Function Theorem is to get smoothness of inverses. Here is an example where even though we have an explicit formula for an inverse function, it's still hard to prove directly that it's differentiable or even continuous, but very easy to prove from the Inverse Function Theorem. We will return to this situation in Example 6.2.3.

Consider polar coordinates given by $(x, y) = F(r, \theta) = (r\cos\theta, r\sin\theta)$. Of course this is not invertible everywhere in the plane: for example when $r = 0$ we get $(x, y) = (0, 0)$ no matter what θ is. Furthermore F is periodic in θ which means for example that $F(r, -\pi) = F(r, \pi)$. So let's consider F as defined on the open rectangle $(r, \theta) \in (0, \infty) \times (-\pi, \pi)$. Then we certainly expect there to be an inverse function $G(x, y) = \big(h(x, y), j(x, y)\big)$ which gives $(r, \theta) = G(x, y)$ except on the negative x-axis.

Clearly we have $r = h(x, y) = \sqrt{x^2 + y^2}$, and this is a C^∞ function except at the origin. But θ is more complicated: we have $\tan\theta = y/x$, which is singular when $x = 0$ and does not uniquely determine θ even when $x \neq 0$. The only way to proceed is to consider the four quadrants separately: we get the formula

$$\theta = j(x, y) = \operatorname{atan2}(y, x) \equiv \begin{cases} \arctan\left(\frac{y}{x}\right) & x > 0 \\ \pi + \arctan\left(\frac{y}{x}\right) & x < 0, y > 0 \\ -\pi + \arctan\left(\frac{y}{x}\right) & x < 0, y < 0 \\ \frac{\pi}{2} & x = 0, y > 0 \\ -\frac{\pi}{2} & x = 0, y < 0 \end{cases} \tag{5.2.4}$$

It is far from clear from this formula that $j(x, y)$ should be continuous or smooth.

On the other hand, if we put these together into $G(x, y) = \big(h(x, y), j(x, y)\big)$, then $F(G(x, y)) = (x, y)$ and $G(F(r, \theta)) = (r, \theta)$ as long as the domains are properly

restricted. So G is the function obtained from Theorem 5.2.4, which is continuous and smooth because DF is nonsingular: we have

$$DF(r,\theta) = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix},$$

and it's trivial to compute that the determinant of this matrix is r, which is always positive on the domain we care about. Thus the theorem tells us that G must be smooth in a small open set around any point, which means it's smooth globally. ☺

You may not have seen the following theorem in a standard analysis class; it often gets featured as the main theorem in a differential equations class. Its proof also uses the Contraction Mapping Lemma 5.2.1, though in a totally different way from the Inverse Function Theorem.

Theorem 5.2.6 (The Fundamental Theorem of Ordinary Differential Equations)**.** *Suppose $F = (f_1, \cdots, f_n)\colon \mathbb{R}^n \to \mathbb{R}^n$ is a smooth function, and we seek a solution of the system of differential equations $\frac{dx}{dt} = F\big(x(t)\big)$:*

$$\begin{aligned} \frac{dx_1}{dt} &= f_1\big(x_1(t), \cdots, x_n(t)\big) \\ &\vdots \\ \frac{dx_n}{dt} &= f_n\big(x_1(t), \cdots, x_n(t)\big), \end{aligned} \tag{5.2.5}$$

with initial condition $\big(x_1(0), \cdots, x_n(0)\big) = (a_1, \cdots, a_n)$.

For any $b \in \mathbb{R}^n$, there is a smooth function $\Gamma = (\gamma_1, \cdots, \gamma_n)\colon (-\delta, \delta) \times U \to \mathbb{R}^n$ defined in some neighborhood U of b and for some small $\varepsilon > 0$ such that for any $a \in U$, the curve $t \mapsto \Gamma(t, a)$ is the unique solution of the differential equation (5.2.5) *with initial condition $\Gamma(0, a) = a$.*

In simpler language: for any initial condition, there is a unique solution of the system (5.2.5)*, defined for a possibly short time, and furthermore the solution depends smoothly on the initial condition.*

Proof. The basic idea is Picard's iteration procedure, which takes some initial guess $\eta_0\colon (-\varepsilon, \varepsilon) \to \mathbb{R}^n$ with $\eta_0(0) = a$, and defines new functions $\eta_n(t)$ by the recursion $\eta_{n+1} = T(\eta_n)$, where the map T is given explicitly by the integral formula

$$T(\eta)(t) = a + \int_0^t F\big(\eta(s)\big)\, ds.$$

Now T is a map from the complete metric space of continuous functions $C\big((-\varepsilon, \varepsilon), \mathbb{R}^n\big)$ (with metric defined by the supremum norm of differences) to itself. If ε is small enough, then one can prove that T will be a contraction mapping, and the unique fixed point of T will be the solution $\gamma(t)$ satisfying $\gamma(0) = a$ and $\frac{d\gamma}{dt} = F\big(\gamma(t)\big)$.

To prove T is a contraction, just compute

$$|T(\eta)(t) - T(\xi)(t)| = \left| \int_0^t F\big(\eta(s)\big) - F\big(\xi(s)\big)\, ds \right| \le \int_0^t \big|F\big(\eta(s)\big) - F\big(\xi(s)\big)\big|\, ds.$$

Now since F is smooth, there is a closed neighborhood $\overline{B}_\delta(0) \times \overline{B}_\varepsilon(a)$ of $(0, a)$ on which $|F_y(t, y)| \le L$, for some number L. Then

$$|T(\eta)(t) - T(\xi)(t)| \le L \int_0^t |\eta(s) - \xi(s)|\, ds \qquad \text{if } |t| \le \delta,$$

from which we can conclude that if $\|\cdot\|$ denotes the supremum norm on $[0,\delta]$, then

$$\|T(\eta) - T(\xi)\| \leq L\delta \|\eta - \xi\|.$$

So possibly shrinking δ, we can ensure $L\delta < 1$ so that T is a contraction. The only thing we need to check is that T actually maps some closed set of continuous functions *into itself*, which is easy to check: if $|\eta(t) - a| \leq \varepsilon$ for every $t \in [-\delta,\delta]$, then the same is true for $T\eta$.

So T is a contraction on the complete space of continuous functions in the supremum norm, and hence there is a unique fixed point γ, which satisfies

$$\gamma(t) = a + \int_0^t F\big(\gamma(s)\big)\, ds \tag{5.2.6}$$

for all $t \in (-\delta,\delta)$. We know γ is continuous, and since F is also continuous, we can take the derivative of both sides of (5.2.6) and get $\gamma'(t) = F\big(\gamma(t)\big)$, which shows γ' is also continuous. Iterating and using the Chain Rule, we conclude that all derivatives of γ exist.

Finally the proof that γ depends smoothly on the initial conditions works as follows. Define $\Gamma(t,b) = \gamma(t)$, where $\gamma(t)$ solves $\gamma'(t) = F(\gamma(t))$ with $\gamma(0) = b$. Since the time of existence depends only on the size of L, there is an open set of values b for which we have solutions all on the *same* time interval $(-\delta,\delta)$. So Γ is defined on some open neighborhood of any particular condition $(0,a)$. We have $\Gamma_t(t,a) = F(t,\Gamma(t,a))$, and so if it existed, the function $H(t,a) = \Gamma_a(t,a)$ would have to satisfy the differential equation $H_t(t,a) = F_a(t,\Gamma(t,a))H(t,a)$. This differential equation for H, with initial condition $H(0,a) = 1$, certainly has a unique solution (it's even easier than the above, since the differential equation is linear). And now that we have a candidate for the derivative, we can prove this actually is the derivative. □

Let's do an example to see exactly how Picard iteration works.

Example 5.2.7. Consider the initial value problem $\frac{dx}{dt} = 3 - x$ with $x(0) = a$. You can check that the exact solution is $x(t) = 3 + (a-3)e^{-t}$, so the map above is $\Gamma(t,a) = 3 + (a-3)e^{-t}$. We can see explicitly how the Picard iteration works: for simplicity let's suppose $a = 5$. Start with $\eta_0(t) = 5$, and construct the sequence

$$\eta_{n+1}(t) = 5 + \int_0^t \big(3 - \eta_n(s)\big)\, ds = 5 + 3t - \int_0^t \eta_n(s)\, ds.$$

We obtain the sequence

$$\begin{aligned}
\eta_0(t) &= 5 \\
\eta_1(t) &= 5 - 2t \\
\eta_2(t) &= 5 - 2t + t^2 \\
\eta_3(t) &= 5 - 2t + t^2 - \tfrac{1}{3}t^3 \\
\eta_4(t) &= 5 - 2t + t^2 - \tfrac{1}{3}t^3 + \tfrac{1}{12}t^4 \\
\eta_5(t) &= 5 - 2t + t^2 - \tfrac{1}{3}t^3 + \tfrac{1}{12}t^4 - \tfrac{1}{60}t^5 \qquad \vdots
\end{aligned}$$

and once we see the pattern, inductively we can show that

$$\eta_n(t) = 5 + \sum_{k=1}^{n} \frac{2(-1)^k}{k!} t^k,$$

and in the limit we get the solution

$$\gamma(t) = 5 + 2(e^{-t} - 1) = 3 + 2e^{-t}.$$

Clearly this sequence actually converges for all values of t, although the theorem would only guarantee this for $|t| < 1$ since the Lipschitz constant is $L = 1$. ☺

For our second example, let's consider a nonlinear differential equation to see how the dependence on parameters works.

Example 5.2.8. Consider the differential equation

$$\frac{dx}{dt} = t^2 + x^2.$$

There is no elementary solution of this equation, although the theorem tells us there are local solutions on a sufficiently small interval around $t = 0$ for any given value of $a = x(0)$.

Let $\Gamma(t, a)$ be the solution operator. Then we have

$$\frac{\partial \Gamma}{\partial t}(t, a) = t^2 + \Gamma(t, a)^2. \tag{5.2.7}$$

Letting $H(t, a) = \Gamma_a(t, a)$ as in Theorem 5.2.6, differentiating (5.2.7) with respect to a, and using the fact that mixed partials commute from Theorem 5.1.9, we obtain

$$\frac{\partial H}{\partial t}(t, a) = 2\Gamma(t, a)H(t, a).$$

This is a *linear* differential equation, which means we can write the solution explicitly for $H(t, a)$ in terms of the unknown function $\Gamma(t, a)$. Since $\Gamma(0, a) = a$, we know $H(0, a) = 1$, and we get the formula

$$H(t, a) = \frac{\partial \Gamma}{\partial a}(t, a) = \exp\left(2\int_0^t \Gamma(s, a)\, ds\right).$$

Now if we know $\Gamma(t, a)$ has infinitely many derivatives in the t direction, this formula tells us that we can differentiate as many times as we like in the a direction as well. Hence Γ has partial derivatives of all orders in both the time t and the initial condition parameter a. ☺

5.3. **Integration.** We can define the integral of a smooth function of several variables over a subset using essentially the same definition as Darboux: one takes a partition of the subset, compares the supremum and the infimum over each element of the partition, and proves that as the partition shrinks in size, the upper sum and lower sum converge to the same thing. Practically, however, this definition is not very useful; instead we want to get a multivariable integral in terms of several single-variable integrals, since those are easy to compute by the Fundamental Theorem of Calculus. Fubini's Theorem enables us to do this. The following simple version is all we'll need.

Theorem 5.3.1 (Fubini's Theorem)**.** *Suppose we have a cube $[0, 1]^n \subset \mathbb{R}^n$ and a smooth function $f\colon \mathbb{R}^n \to \mathbb{R}$. Then the integral $\displaystyle\int_{[0,1]^n} f\, dV$ (defined in terms of partitions of the cube) can be computed as the iterated integral*

$$\int_{[0,1]^n} f\, dV = \int_0^1 \left(\int_0^1 \cdots \left(\int_0^1 f(x_1, x_2, \cdots, x_n)\, dx_1\right) \cdots dx_{n-1}\right) dx_n.$$

Furthermore it does not matter in which order we perform the iterated integrals.

Proof. It is sufficient to prove this for two variables at a time, so we want to show

$$\int_0^1 \int_0^1 f(x,y)\,dx\,dy = \int_0^1 \int_0^1 f(x,y)\,dy\,dx \tag{5.3.1}$$

if f is smooth. We will derive this from the fact that mixed partial derivatives commute via Theorem 5.1.9. The idea comes from a paper of Aksoy and Martelli.[4]

Define $g(x,y) = \int_0^y f(x,v)\,dv$, and $h(x,y) = \int_0^x g(u,y)\,du$. Then h is smooth and $\dfrac{\partial^2 h}{\partial y \partial x} = \dfrac{\partial g}{\partial y} = f$. Therefore $\dfrac{\partial^2 h}{\partial x \partial y} = f$ as well since mixed partials commute.

Now $h(1,1)$ is the right side of (5.3.1) by definition. But also the left side of (5.3.1) is (by the one-dimensional version of the Fundamental Theorem of Calculus)

$$\begin{aligned}\int_0^1 \int_0^1 f(x,y)\,dx\,dy &= \int_0^1 \int_0^1 \frac{\partial}{\partial x}\frac{\partial h}{\partial y}\,dx\,dy \\ &= \int_0^1 \frac{\partial h}{\partial y}(1,y) - \frac{\partial h}{\partial y}(0,y)\,dy \\ &= h(1,1) - h(0,1) - h(1,0) + h(0,0).\end{aligned}$$

From the definition of h we can see that $h(0,1) = h(0,0) = 0$, while $h(1,0) = 0$ from the definition of g, So we just have $h(1,1)$ remaining, as we wanted, which shows that the left side of (5.3.1) equals the right side. □

The change of variables theorem for integrals will be very important to us, for it will allow us to define the integral in a coordinate-invariant way at the end of these notes. It generalizes the Substitution Rule for one-variable integrals. Again we only need a simple case.

The change of variables theorem is usually stated in more generality than this (in particular involving integration on more general regions than squares), but from the perspective of differential geometry, it's a lot easier to just have one definition for squares and define everything else in terms of that via parametrizations. (For example, spherical coordinates give a parametrization of the unit sphere in terms of the rectangle $[0,\pi] \times [0,2\pi]$, and it's on that rectangle that everyone actually does spherical integrations.) We just have to check consistency if we have two ways of describing the same integral on the same square.

The technique of proof is basically to reduce each side to an integral over the boundary, using something like Green's Theorem. We will discuss Green's Theorem later (as a special case of Stokes' Theorem), though here we just need the one-dimensional Fundamental Theorem of Calculus.

Theorem 5.3.2 (Change of Variables for Integrals)**.** *Suppose we have a smooth function $f\colon \mathbb{R}^n \to \mathbb{R}$ and a smooth function $\Psi\colon \mathbb{R}^n \to \mathbb{R}^n$ with smooth inverse Ψ^{-1} such that $\Psi\big([0,1]^n\big) = [0,1]^n$.*

Then

$$\int_{[0,1]^n} f\big(\Psi(\mathbf{u})\big) \det D\Psi(\mathbf{u}) dV(\mathbf{u}) = \int_{[0,1]^n} f(\mathbf{x})\,dV(\mathbf{x}). \tag{5.3.2}$$

[4]Mixed partial derivatives and Fubini's Theorem, The College Mathematics Journal, vol. 33, no. 2 (Mar. 2002), pp. 126–130.

Proof. I'll prove this in the two-dimensional case, which gives the easiest proof I've seen. There is an only slightly more complicated proof of the n-dimensional result due to Peter Lax which uses the same basic idea.[5]

Write $\mathbf{x} = (x, y) = \Psi(\mathbf{u}) = \big(\varphi(u, v), \psi(u, v)\big)$, and suppose that Ψ maps the boundary segments into themselves (so that for example $\psi(u, 0) = 0$ for all $u \in [0, 1]$, which corresponds to mapping the bottom boundary into itself). Then

$$\det(D\Psi)(\mathbf{u}) = \varphi_u(u, v)\psi_v(u, v) - \varphi_v(u, v)\psi_u(u, v).$$

We are then trying to prove that

$$\int_0^1 \int_0^1 \Big(f\big(\varphi, \psi\big)[\varphi_u\psi_v - \varphi_v\psi_u]\Big)(u, v)\, du\, dv = \int_0^1 \int_0^1 f(x, y)\, dx\, dy. \tag{5.3.3}$$

The trick is to define a new function

$$g(x, y) = \int_0^x f(r, y)\, dr, \tag{5.3.4}$$

so that $f(x, y) = g_x(x, y)$ for all x and y in the square. This way we are basically integrating a derivative and can reduce the double integral to an integral over the boundary. It's easier to make everything work on the boundary since it's just the one-dimensional substitution rule.

First note that the right side of (5.3.3) is obviously

$$\int_0^1 \int_0^1 \frac{\partial g}{\partial x}(x, y)\, dx\, dy = \int_0^1 g(1, y)\, dy - \int_0^1 g(0, y)\, dy = \int_0^1 g(1, y)\, dy \tag{5.3.5}$$

since $g(0, y) = 0$.

Let us write $p(u, v)$ for the integrand on the left side of (5.3.3): we have

$$p(u, v) = g_x\Big(\varphi(u, v), \psi(u, v)\Big)\Big[\varphi_u(u, v)\psi_v(u, v) - \varphi_v(u, v)\psi_u(u, v)\Big], \tag{5.3.6}$$

and we are now trying to show that

$$\int_0^1 \int_0^1 p(u, v)\, du\, dv = \int_0^1 g(1, y)\, dy. \tag{5.3.7}$$

The trick is that $p(u, v)$ can be written in the form

$$p(u, v) = \frac{\partial q}{\partial u}(u, v) - \frac{\partial r}{\partial v}(u, v)$$

for some functions q and r, and doing this allows us to write the left side of (5.3.3) as a boundary integral. Indeed, if

$$q(u, v) = g\Big(\varphi(u, v), \psi(u, v)\Big)\psi_v(u, v), \tag{5.3.8}$$

$$r(u, v) = g\Big(\varphi(u, v), \psi(u, v)\Big)\psi_u(u, v), \tag{5.3.9}$$

[5]Change of variables in multiple integrals, The American Mathematical Monthly, vol. 106 (1999), pp. 497–501.

then using the Chain Rule (5.1.5) gives

$$\begin{aligned}\frac{\partial q}{\partial u}(u,v) - \frac{\partial r}{\partial v}(u,v) &= \Big[g_x\Big(\varphi(u,v),\psi(u,v)\Big)\varphi_u(u,v) + g_y\Big(\varphi(u,v),\psi(u,v)\Big)\psi_u(u,v)\Big]\psi_v(u,v)\\ &\quad - \Big[g_x\Big(\varphi(u,v),\psi(u,v)\Big)\varphi_v(u,v) - g_y\Big(\varphi(u,v),\psi(u,v)\Big)\psi_v(u,v)\Big]\psi_u(u,v)\\ &\quad + g\Big(\varphi(u,v),\psi(u,v)\Big)\psi_{vu}(u,v) - g\Big(\varphi(u,v),\psi(u,v)\Big)\psi_{uv}(u,v)\\ &= g_x\Big(\varphi(u,v),\psi(u,v)\Big)\Big[\varphi_u(u,v)\psi_v(u,v) - \varphi_v(u,v)\psi_u(u,v)\Big]\\ &= p(u,v)\end{aligned}$$

by the definition (5.3.6). Here we used equality of mixed partials: $\psi_{uv} = \psi_{vu}$.

The consequence is that the left side of (5.3.7) can be computed using Fubini's Theorem 5.3.1 to get

$$\begin{aligned}\int_0^1\int_0^1 p(u,v)\,du\,dv &= \int_0^1\int_0^1 \frac{\partial q}{\partial u}(u,v)\,du\,dv - \int_0^1\int_0^1 \frac{\partial r}{\partial v}(u,v)\,dv\,du\\ &= \int_0^1 \big[q(1,v) - q(0,v)\big]\,dv - \int_0^1 \big[r(u,1) - r(u,0)\big]\,du.\end{aligned}$$

To simplify this, recall that we assumed (φ,ψ) mapped the square to itself, so that $\varphi(0,v) = 0$, $\varphi(1,v) = 1$, $\psi(u,0) = 0$, and $\psi(u,1) = 1$. Hence using the definitions (5.3.8)–(5.3.9), we get

$$\begin{aligned}q(1,v) &= g\big(\varphi(1,v),\psi(1,v)\big)\psi_v(1,v) = g\big(1,\psi(1,v)\big)\psi_v(1,v),\\ q(0,v) &= g\big(\varphi(0,v),\psi(0,v)\big)\psi_v(0,v) = g\big(0,\psi(0,v)\big)\psi_v(0,v),\\ r(u,1) &= g\big(\varphi(u,1),\psi(u,1)\big)\psi_u(u,1) = 0,\\ r(u,0) &= g\big(\varphi(u,0),\psi(u,0)\big)\psi_u(u,0) = 0.\end{aligned}$$

The third and fourth lines are zero since $\psi(u,0)$ and $\psi(u,1)$ are both constant in u. The second line is also zero since $g(0,y) = 0$ by definition (5.3.4). So all that's left is

$$\int_0^1\int_0^1 p(u,v)\,du\,dv = \int_0^1 g\big(1,\psi(1,v)\big)\psi_v(1,v)\,dv.$$

To relate this to $\int_0^1 g(1,y)\,dy$, just use the one-dimensional change of variables $y = \psi(1,v)$, so that $dy = \psi_v(1,v)\,dv$. We thus obtain (5.3.7), as desired. □

Example 5.3.3. The most basic example is the polar coordinate transformation $x = r\cos\theta$, $y = r\sin\theta$. The Jacobian of this transformation is

$$J(r,\theta) = \begin{vmatrix}\frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta}\\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta}\end{vmatrix} = \begin{vmatrix}\cos\theta & -r\sin\theta\\ \sin\theta & r\cos\theta\end{vmatrix} = r.$$

Therefore the area element transforms into $dA = dx\,dy = r\,dr\,d\theta$, in the sense that the integral of a function over a region Ω can be evaluated using either of

$$\int_\Omega f(x,y)\,dx\,dy = \int_\Omega h(r,\theta)\,r\,dr\,d\theta,$$

where $h(r,\theta) = f(r\cos\theta, r\sin\theta)$. (Strictly speaking we did not actually prove this, since the only coordinate transformations we considered were those that mapped the unit square to itself, but the general proof works in more or less the same way.)

In three dimensions using spherical coordinates (ρ, θ, ϕ) defined by $x = \rho \sin\theta \cos\phi$, $y = \rho \sin\theta \sin\phi$, $z = \rho \cos\theta$, the Jacobian determinant is easily computed to be $J(\rho, \theta, \phi) = \rho^2 \sin\theta$, so that the volume element is $dV = dx\, dy\, dz = \rho^2 \sin\theta\, d\rho\, d\theta\, d\phi$. ☺

In Chapter 15 we will give the change-of-variables formula a meaning independently of its application in iterated integrals, by viewing it instead as a formula involving n-forms. In fact the reason n-forms end up being so important in this subject is because the determinant happens to show up in the change-of-variables formula and also shows up in the change-of-basis formula for n-forms in Proposition 4.3.10.

We have one last result: differentiating under an integral. The fact that this works is occasionally useful, but perhaps not of fundamental importance. We'll use it only once, to prove the Poincaré Lemma.

Theorem 5.3.4. *Suppose $F\colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ is smooth, and consider the function $f\colon \mathbb{R} \to \mathbb{R}$ given by*

$$f(s) = \int_{[0,1]^n} F(s,x)\, dV(x).$$

Then f is smooth, and

$$f'(s) = \int_{[0,1]^n} \frac{\partial F}{\partial s}(s,x)\, dV(x).$$

Proof. It works the same whether $n = 1$ or not, so we might as well do one dimension to simplify the proof. Fix a particular s, and let h be any fixed number. If $f(s) = \int_0^1 F(s,x)\, dx$, then

(5.3.10)
$$\frac{f(s+h) - f(s)}{h} - \int_0^1 F_s(s,x)\, dx = \int_0^1 \left[\frac{F(s+h,x) - F(s,x)}{h} - F_s(s,x) \right] dx.$$

By the Mean Value Theorem 5.1.8, we know for every fixed x that $F(s+h,x) - F(s,x) = hF_s(\xi(h,x),x)$ for some $\xi(h,x)$ between s and $s+h$. We don't assume that $\xi(h,x)$ is continuous in either variable or anything.

Thus

$$\frac{F(s+h,x) - F(s,x)}{h} - F_s(s,x) = F_s(\xi(h,x),x) - F_s(s,x).$$

Using the Mean Value Theorem again on the right side of this, we get

$$F_s(\xi(h,x),x) - F_s(s,x) = (\xi(h,x) - s) F_{ss}(\zeta(h,x),x)$$

for some $\zeta(h,x)$ between s and $s+h$. Now F_{ss} is continuous, so it's bounded by some constant L on $[s-1, s+1] \times [0,1]$. Thus as long as $|h| < 1$ we know

$$\left| \frac{F(s+h,x) - F(s,x)}{h} - F_s(s,x) \right| \leq |\xi(h,x) - s|\, |F_{ss}(\zeta(h,x),x)| \leq Lh.$$

This being true for every $x \in [0,1]$, the integral on the right side of (5.3.10) is bounded in size by Lh, and so as h goes to zero the left side converges to 0 as well. □

6. Coordinates

"Luminous beings are we, not this crude matter."

6.1. **Concepts of coordinates.** In this Chapter I want to discuss classical notions of coordinate charts, which predate manifolds by hundreds of years. Classically one was primarily interested in coordinate charts as a way of solving problems in the plane or in space; only later was it realized that all the same complications show up in trying to set up manifolds. For now I am going to follow the historical approach, which means I will be talking about an abstract space M, by which I usually mean two- or three-dimensional space. It will help if you think of it as pre-Descartes. That is, you have a plane but there's no special horizontal or vertical line that you call an "axis." Imagine a plane like a Greek geometer would, and think of Descartes spitefully, the way his old-fashioned contemporaries might have. Unlearn the idea that the plane is $\mathbb{R}^2$. The plane is the plane, and $\mathbb{R}^2$ is a list of two numbers. Now let's proceed.

Much of the motivation for the tensor calculus (the main tool for computations in differential geometry) came originally from the desire of physicists to have their equations expressed in a coordinate-invariant way.[6] To some degree, this is a philosophical issue: ancient mathematicians viewed points as pre-existing objects, and their "coordinates" as just an occasionally convenient way of specifying them. Geometric constructions of points were often more important than their distances from fixed lines. After Descartes, the algebraic idea of points as certain coordinates gained prominence. In analysis, for example, a plane is fundamentally a set of pairs of reals that you can occasionally draw pictures on. In differential geometry, on the other hand, we view a coordinate system only as a convenience: the points exist abstractly, and any set of coordinates is as good as any other in telling you where one is in relation to another.

To give a concrete example, we might view the plane $\mathbb{R}^2$ as the set $\mathbb{C}$ of complex numbers. We can specify a complex number by separating it into its real and imaginary components, $z = x + iy$, or we can write it in polar coordinates as $z = re^{i\theta}$. Depending on the application, one or the other might be more useful. We can describe continuous real-valued functions on $\mathbb{C}$ independently of coordinates; for example, $f\colon \mathbb{C} \to \mathbb{R}$ defined by $f(z) = \operatorname{Re}(z^2)$ makes sense independently of any particular coordinate system. In rectangular coordinates, the function is represented by $f_{\text{rect}}(x, y) = x^2 - y^2$, and in polar coordinates the function is represented by $f_{\text{polar}}(r, \theta) = r^2 \cos 2\theta$, but the function itself is defined without reference to coordinates.

This is generally how we will want to think of spaces and functions on them: we have a topological space M (which for now is homeomorphic to $\mathbb{R}^n$) and we have continuous functions $f\colon M \to \mathbb{R}$. The space M may be given in any number of ways:

[6] The importance of this philosophical point can't be underestimated. In physics, the fact that Newtonian mechanics stayed the same when one observer moved at a constant speed in some direction was profound and fundamental. As were Maxwell's beautiful equations of electromagnetism. The fact that Maxwell's equations looked hideously ugly as soon as you added a uniform velocity to a frame was a profound disappointment in the late 1800s, and was one of the primary motivations for special relativity.

for example the plane might be the complex numbers, or it might be the sphere with a point removed, or it might be the set of intervals in $\mathbb{R}$. Our method of representing continuous functions on M will depend on how M is defined, but what's essential is that we *don't* think of M as being defined by any particular coordinates. Doing this will make generalizing everything to manifolds much easier. So we will distinguish between M, the abstract Euclidean space where any coordinate system is as good as any other, and $\mathbb{R}^n$, the standard Euclidean space with Cartesian coordinates. To specify when we're thinking this way, I'll write $M \cong \mathbb{R}^n$.

This is a big conceptual leap to make, and you should think carefully about it. The point is that we need to do most computations in coordinates, but on the other hand, we can't pick any preferred coordinate system. It's similar to how, when working on a finite-dimensional vector space (as in Chapter 3), we need to do most computations in a basis, although the abstract vectors we're dealing with have some meaning independently of any basis: as long as you and I always know how to transform results in your basis to results in my basis and vice versa, we can work this way. I remind you again of Plato's suggestion in *Republic* that all we see are just the shadows of reality dancing on the walls of a cave: points and such in the abstract Euclidean space exist in some deeper sense, but we can only perceive them by looking at their shadows (their coordinate representations). As long as we understand this, we will know how to transform between different coordinate representations (i.e., the shadow you see from your position vs. the shadow I see from mine).

6.2. **Examples of coordinate systems.** Historically, the first coordinate system on $\mathbb{R}^n$ was obviously the Cartesian system, used in some sense by ancient mathematicians and formalized by Descartes in the 1600s. (For historical notes, I will generally refer to Julian Coolidge, *A history of geometrical methods*.)

Example 6.2.1 (Cartesian coordinates)**.** Let us suppose M is homeomorphic in some way to $\mathbb{R}^n$; then Cartesian coordinates on M will be written as $\{x^1, x^2, \cdots, x^n\}$. We think of x^1, x^2, etc., as being functions $x^k \colon M \to \mathbb{R}$, and we can write $\mathbf{x} \colon M \to \mathbb{R}^n$ for the collection of these functions. This is essentially the identity map, although philosophically we are thinking of the domain as being some abstract space and the range as being a set of n numbers. The reason we are doing these convoluted things will be clearer when we talk about other coordinates.

The superscript notation is used in differential geometry to make certain formulas more convenient; it should not be confused with exponents. If the dimension of M is two or three, we can use (x, y) or (x, y, z) instead of the superscripts. ☺

Remark 6.2.2. This is as good a time as any to apologize for the notation. The notation $\mathbf{x}$ will usually mean the function that takes a point in an abstract space to its coordinates in $\mathbb{R}^n$. But sometimes it will refer to a particular point in $\mathbb{R}^n$, as in vector calculus. The notation $x^1, \cdots, x^n$ will sometimes mean the individual component functions of the function $\mathbf{x}$, but more often it will mean a particular point $(x^1, \cdots, x^n) \in \mathbb{R}^n$. This is obviously not logically consistent. Worse still, we will often use notation such as $f \circ \mathbf{x}^{-1}(x^1, \cdots, x^n)$, which means $f(p)$ where $\mathbf{x}(p) = (x^1, \cdots, x^n)$. What I actually mean should be clear from the context. But this is one of those situations where I think it's more confusing to use good notation than to settle for bad notation.

Example 6.2.3 (Polar coordinates)**.** The first alternative coordinate system consisted of polar coordinates in the plane, invented by Jacob Bernoulli in the late 1600s (but like Cartesian coordinates, also used less systematically by earlier mathematicians). Standard polar coordinates (r, θ) are related to standard Cartesian coordinates (x, y) by the usual formulas

$$(x, y) = (r\cos\theta, r\sin\theta). \tag{6.2.1}$$

As discussed in Example 5.2.5, the inverse function is

$$(r, \theta) = \big(\sqrt{x^2 + y^2}, \operatorname{atan2}(y, x)\big),$$

where atan2 has a rather complicated formula. However the Inverse Function Theorem guarantees that these complicated formulas give C^∞ functions of the coordinates (x, y) on some open set.

It is important to notice that we *cannot* define $\theta(x, y)$ in such a way as to be continuous on $\mathbb{R}^2$ or even on $\mathbb{R}^2$ minus the origin: we need to eliminate an entire ray (corresponding to $\theta = \pi$, although eliminating the ray $\theta = 2\pi$ is common as well). Even in this simple example, we see that we cannot expect a general coordinate system to be defined on the entire space. The map from $(0, \infty) \times (-\pi, \pi)$ to the plane is best visualized as in Figure 6.1. The fact that θ can be defined on the plane minus a ray but *not* on the plane minus the origin has profound importance later on: this situation is the most basic and fundamental example in the general theory of de Rham cohomology, which we will discuss in Chapter 18.

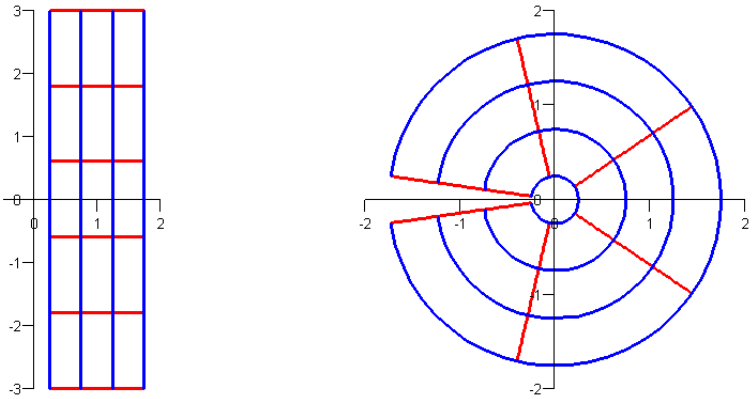

FIGURE 6.1. Coordinate curves in the $r\theta$-plane on the left, and their image under (6.2.1) in the xy-plane on the right.

Now let us think of $M \cong \mathbb{R}^2$ as being the abstract Euclidean space, with no preferred system of coordinates. Let $\mathbf{x}\colon M \to \mathbb{R}^2$ be the Cartesian system with $\mathbf{x} = (x, y)$, and $\mathbf{u}\colon U \to \mathbb{R}^2$ be the polar coordinate system $\mathbf{u} = (r, \theta)$, with the open set U defined as M minus the leftward ray corresponding to $\theta = \pi$:

$$U = \mathbb{R}^2 \backslash \{(x, 0) \,|\, x \leq 0\}.$$

Then the equations above say that the *transition map* is given by

$$\mathbf{x} \circ \mathbf{u}^{-1}(r, \theta) = (r\cos\theta, r\sin\theta) \tag{6.2.2}$$

while its inverse is given by

$$\mathbf{u} \circ \mathbf{x}^{-1}(x, y) = \big(\sqrt{x^2 + y^2}, \operatorname{atan2}(y, x)\big). \tag{6.2.3}$$

The domain of $\mathbf{x} \circ \mathbf{u}^{-1}$ is the image of $\mathbf{u}$, which is $(0, \infty) \times (-\pi, \pi)$, a half-strip. The domain of $\mathbf{u} \circ \mathbf{x}^{-1}$ is restricted by the domain of $\mathbf{u}$, so it is U. Notice that by using *two* coordinate systems and just worrying about their transition maps, we have basically managed to avoid even talking about the abstract Euclidean space M. (In other words, we're just dealing directly with the shadows on the cave wall, regardless of what they really represent.) In this notation, if $M = \mathbb{C}$ with $f(z) = \operatorname{Re}(z^2)$, then $f \circ \mathbf{x}^{-1}(x, y) = f_{\text{rect}}(x, y) = x^2 - y^2$ and $f \circ \mathbf{u}^{-1}(r, \theta) = f_{\text{polar}}(r, \theta) = r^2 \cos 2\theta$. So the notation allows us to separate the *actual* function from its representation in coordinates.

Why go through all this trouble of domain-restricting? The Calculus II approach to polar coordinates often allows θ to be any value at all, identifying those that differ by integer multiples of 2π, and it allows r to be zero and sometimes even negative! In so doing we cover the entire Euclidean plane with polar coordinates. The problem is that when this approach leads to difficulties, we have to get around them using some ad hoc technique. This might be OK if polar coordinates were the only coordinate system we'd ever change to, but that's not the case. We want to treat both Cartesian and polar coordinates as equals, and doing this forces us to work on some restricted subset of the plane, where the coordinate transformations work equally well (and have good continuity and differentiability properties) both ways.

To make things more explicit, suppose we have a function $f_{\text{polar}}(r, \theta) = r^3$. This looks like a perfectly smooth function, and it is on the set $(0, \infty) \times (-\pi, \pi)$, but it's *not* as a function on the Euclidean plane $M \cong \mathbb{R}^2$. The reason is obvious if we assume that $f_{\text{polar}} = f \circ \mathbf{u}^{-1}$ for some function $f\colon M \to \mathbb{R}^2$. Write it in Cartesian coordinates, and we get $f_{\text{rect}}(x, y) = f \circ \mathbf{x}^{-1}(x, y) = (x^2 + y^2)^{3/2}$. This function does not have a continuous third derivative and hence no fourth derivative at all: we can compute that

$$\partial_x^3 f_{\text{rect}}(x, y) = \frac{3x(2x^2 + 3y^2)}{(x^2 + y^2)^{3/2}},$$

which has no limit as $(x, y) \to (0, 0)$. Specifically,

$$\lim_{y \to 0} \partial_x^3 f_{\text{rect}}(0, y) = 0 \quad \text{while} \quad \lim_{x \to 0} \partial_x^3 f_{\text{rect}}(x, 0) = 6.$$

We don't want to deal with functions that look smooth in one coordinate system and not-smooth in another coordinate system. Instead we can say that since f depends smoothly on r, the function $f \circ \mathbf{u}^{-1}\colon U \subset M \to \mathbb{R}$ is smooth on a subset U of M (but not on M itself). By doing this, we never have to worry about the strange behavior of a coordinate system at a singularity. Instead we only work with the coordinate system at the points in the open set where it is smooth; if we want to work with other points, we do so using some other coordinate chart. Think of a coordinate singularity as being, in Plato's analogy, a hole or crack in the cave wall that we have to learn to look past, knowing that the real objects have no holes or cracks. The functions we really care about will be defined properly, somehow, on the entire space. Thus they will *automatically* look smooth in any coordinate system. ☺

Make sure you understand all of this before going on.

Example 6.2.4 (Other coordinate systems)**.** Although they aren't traditionally taught in vector calculus, there are many other coordinate systems that are sometimes useful. They are generally defined by their level curves, and they are useful when the geometry of a particular problem involves the same level curves. For example, Cartesian coordinates are useful in a domain bounded by a rectangle, while polar coordinates are useful in domains bounded by circles and rays. If a boundary is more complicated, it is usually better to use a coordinate system suited to the boundary. Here are some classical examples.

- **Parabolic coordinates** These planar coordinates are related to Cartesian coordinates by the transformation

$$(x, y) = \left(\sigma\tau, \tfrac{1}{2}(\tau^2 - \sigma^2)\right), \tag{6.2.4}$$

which can be inverted without much difficulty to obtain

$$(\tau, \sigma) = \left(\pm\sqrt{y + \sqrt{x^2 + y^2}}, \frac{\pm x}{\sqrt{y + \sqrt{x^2 + y^2}}} \right).$$

From the explicit formulas, we see that this is a smooth and invertible transformation in the region $\tau > 0$, $\sigma \in \mathbb{R}$, and furthermore all (x, y) except for the lower ray $L = \{(x, y) \,|\, x = 0, y \leq 0\}$ can be obtained from such τ and σ. Curves of constant σ or constant τ are parabolas, as shown in Figure 6.2. You might notice that the parabolas all cross at right angles: this is a useful property in Riemannian geometry, and is a typical feature of the most popular coordinate charts.[7]

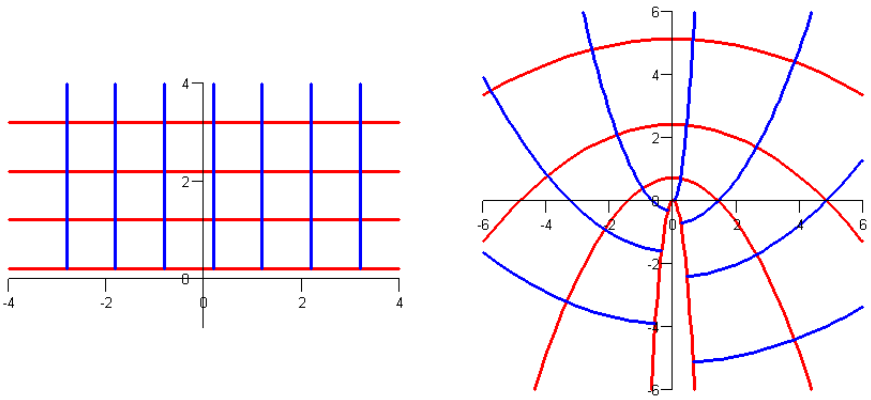

FIGURE 6.2. Coordinate curves in the $\sigma\tau$-plane on the left, and their image under (6.2.4) in the xy-plane on the right.

- **Elliptical coordinates** These coordinates are related to Cartesian coordinates by the transformation

$$(x, y) = (\cosh\mu \cos\nu, \sinh\mu \sin\nu). \tag{6.2.5}$$

[7]One of the reasons complex analysis finds so many applications is that it gives an easy way to construct such orthogonal coordinate charts: in this case you might notice that $x+iy = -\frac{i}{2}(\sigma+i\tau)^2$ is a complex-analytic function.

By demanding $\mu > 0$ and $\nu \in (-\pi, \pi)$ for example, we obtain a genuine coordinate system which is invertible. This coordinate system is convenient because the level curves are hyperbolas (when ν is held constant) and ellipses (when μ is held constant), and all of these level curves have the same foci, at $(-1, 0)$ and $(1, 0)$. See Figure 6.3. Observe that the coordinate system fails to cover the left half of the x-axis (because of the angle restriction) and also fails to cover the portion of the x-axis between the two foci, so that the image of the transformation is the plane minus the portion of the x-axis left of $(1, 0)$. As before, one of the reasons this coordinate chart is useful is because the level sets are orthogonal curves.[8]

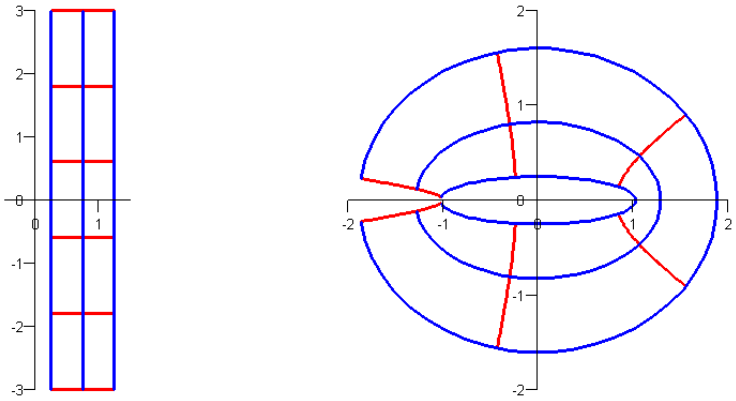

FIGURE 6.3. Coordinate curves in the $\mu\nu$-plane on the left, and their image under (6.2.5) in the xy-plane on the right.

☺

As with many mathematical fields (such as analysis and algebra), differential geometry historically proceeded from fairly simple foundations, which were gradually expanded using special techniques to solve interesting problems, until the theory was rebuilt from scratch to accommodate the greater generality. With differential geometry, the use of ever more complicated coordinate systems, which often led to non-explicit formulas for solutions of problems, demanded the construction of a theory which was genuinely independent of any particular coordinate system. This is where we are going.

Based on the issues observed above with polar coordinates, and on our philosophy that any coordinate system is as good as any other, we now define a coordinate chart formally.

Definition 6.2.5. Consider $\mathbb{R}^n$ as an abstract space M. A *coordinate chart* on M is a pair (ϕ, U), where U is an open subset of M and ϕ is a homeomorphism from U to a subset of $\mathbb{R}^n$ (in other words, ϕ^{-1} is a continuous map from $\phi[U]$ to M). Two coordinate charts (ϕ, U) and (ψ, V) are called C^∞*-compatible* if the functions $\phi \circ \psi^{-1} \colon \psi[U \cap V] \subset \mathbb{R}^n \to \mathbb{R}^n$ and $\psi \circ \phi^{-1} \colon \phi[U \cap V] \subset \mathbb{R}^n \to \mathbb{R}^n$ are both C^∞.

[8]And again, orthogonality comes from complex analysis: we have $x + iy = \cosh(\mu + i\nu)$. It's fair to say that if quaternions in three or four dimensions generated orthogonal coordinate charts as easily as complex numbers did in two dimensions, they'd have become as famous.

There are several things to notice about this definition.

- We assume only a topological structure on M, and we don't differentiate anything on M directly. The only things we're allowed to differentiate are the transition functions between coordinate charts, not the actual coordinate charts themselves. (Again, for motivation you might imagine that M represents the complex numbers, or just some abstract topological space that happens to be homeomorphic to the plane.)
- We explicitly incorporate the open set U into the definition of the coordinate chart, which avoids difficulties due to singular points. Even if ϕ and ψ are the same map, if the open sets are different, then their charts are different. (However they are compatible, since the transition function is the identity, which is C^∞.)
- If two coordinate charts do not cover the same portion of M (i.e., $U \cap V = \emptyset$) then they are trivially compatible, since there is no transition function for which to check derivatives.
- The notion of C^∞-compatible is a not quite an equivalence relation. We would like to say that if (ϕ, U) and (ψ, V) are compatible, and (ψ, V) and (ξ, W) are compatible, then (ϕ, U) and (ξ, W) are compatible. The problem is that we'll check smoothness of $\phi \circ \psi^{-1}$ on $\psi[U \cap V]$ and smoothness of $\psi \circ \xi^{-1}$ on $\xi[V \cap W]$, and this will prove smoothness of $\phi \circ \xi^{-1}$ on $\xi[V \cap W] \cap \xi[U \cap V]$, but this may be a proper subset of $\xi[U \cap W]$, which is what we'd really care about.
- We use C^∞ as our requirement for the transition functions instead of real analytic (i.e., having a convergent multivariable Taylor series). This is done mostly to avoid worrying about a radius of convergence, as well as to allow "bump functions" which are usually C^∞ but never real analytic. We will discuss this in more depth later, in Chapter 13.

7. Manifolds

"That's no moon. It's a space station!"

7.1. **Motivation and definition.** Once you get comfortable with the idea of coordinate charts on $\mathbb{R}^n$ (in particular, the idea that a typical coordinate system will not cover the entire space, but just an open subset), it becomes natural to look at manifolds more generally. To do this properly, one needs to understand a bit of point-set topology.

The basic object in topology is the *open set*. If M is a set, then a *topology* on M is a collection $\mathcal{T}$ of sets satisfying three properties:

- $\mathcal{T}$ contains the empty set $\emptyset$ and the entire set M
- If U and V are in $\mathcal{T}$, then so is $U \cap V$.
- If $\{U_\alpha \mid \alpha \in I\}$ is any collection of sets that are in $\mathcal{T}$ (for an arbitrary index set I), then so is the union $\cup_{\alpha\in I} U_\alpha$.

Every set in $\mathcal{T}$ is called an *open subset of* M. The basic example is the open sets you are familiar with in $\mathbb{R}$ or $\mathbb{R}^n$: sets for which every point is an interior point, or in other words sets for which every point has a radius r such that the entire ball of radius r is contained in the open set. Since just about all the most important concepts of real analysis can be abstracted into a statement about open sets, almost all the most important results of analysis are actually results about topology. Examples include compactness, connectedness, convergence, continuity, and even a few that don't start with 'c.' Intuitively you should think of topological spaces as very general things, and of manifolds as the simplest possible generalization of the Euclidean topologies $\mathbb{R}^n$. Here are the basic topological definitions.

Definition 7.1.1. Suppose M is a topological space (i.e., a set with a topology $\mathcal{T}$ satisfying the three basic properties). Then:

(1) M is *compact* if for every family of open sets $\{U_\alpha \mid \alpha \in I\}$ for some index set I with $M = \cup_{\alpha\in I} U_\alpha$, there is a finite subset $\{\alpha_1, \ldots, \alpha_m\}$ such that $M = \cup_{k=1}^m U_{\alpha_k}$.
(2) M is *connected* if the only way to write $M = U \cup V$ with $U \cap V = \emptyset$ is if one set is M and the other set is $\emptyset$.
(3) A sequence (x_n) *converges* to x if for every open $U \ni x$, the tail $\{x_n \mid n \geq N\}$ is contained in U for N large enough.
(4) If N is also a topological space, then $f\colon M \to N$ is *continuous* if $f^{-1}[V]$ is open in M whenever V is open in N.

The basic idea of a manifold is that near every point, if you zoom in closely enough, it should look like $\mathbb{R}^n$. So for example a circle in $\mathbb{R}^2$ would be manifold, since on a sufficiently small segment it looks like an interval in $\mathbb{R}$. On the other hand a figure '8' would not be a manifold, since no matter how close you get to the center point, it looks like $\times$. See Figure 7.1. Here I'm using "looks like" in the sense of topological equivalence, which means there is a 1-1 correspondence from one to the other which is continuous in both directions. So in particular we don't care about the fact that a circle is round, just that it closes up: an oval or even a square will be considered an equivalent topological manifold. (The fact that the square has corners will be dealt with in a bit when we talk about smooth manifolds,

but for right now we're just talking about topology.) See Figure 7.2. Note also that an open interval is a manifold while a closed interval is not; no neighborhood of an endpoint looks like the real line. A nice exercise is to identify which capital letters on your keyboard are one-dimensional manifolds, using a sans-serif font like Verdana or Arial.

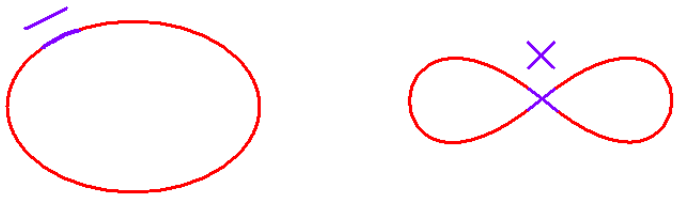

FIGURE 7.1. The oval on the left is a manifold; the small purple segment can be identified with an interval. The figure eight on the right is not a manifold because every neighborhood of the center point (shown in purple) looks like an $\times$ and is topologically distinct from an interval.

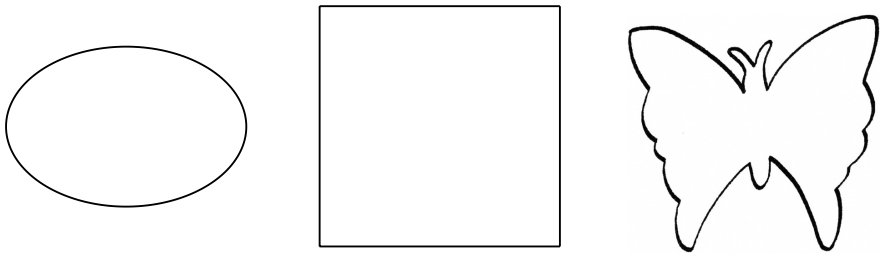

FIGURE 7.2. Topologically, these are all the same curve and might as well be a circle.

Start building up an intuition for topological invariance. Much later we will start worrying about actually measuring things, but for now any kind of stretching or bending will be considered to have no effect whatsoever on the manifold. However gluing things together or poking holes *will* change the topology. So for example we get the famous joke, "A topologist is someone who doesn't know the difference between a coffee cup and a doughnut," since she could easily deform the coffee cup into a doughnut topologically as long as she doesn't try to cut off the handle. (See Figure 7.3.) More crudely, there's the less famous joke that a topologist doesn't know the difference between his ass and a hole in the ground, but he knows the difference between his ass and *two* holes in the ground. The topological property that distinguishes them is that a hole is distinguished from no hole by the fact that you can draw a loop around the hole which can't be contracted without leaving the space, while any loop in the plane with no hole can be contracted to a point; with two holes, you can draw two noncontractible loops.

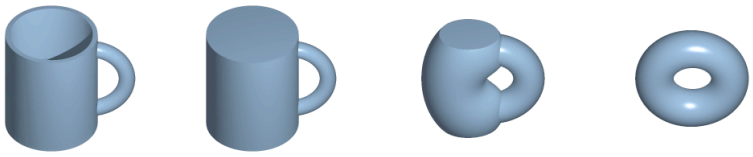

FIGURE 7.3. A coffee mug topologically transformed into a doughnut.

Roughly speaking, there is only one "closed" manifold in one dimension, which is the circle. In two dimensions, there is a countable family of closed manifolds: the sphere, the (one-holed) torus, the two-holed torus, the three-holed torus, etc. These can be parametrized by food, as shown in Figure 7.4. Remember, the manifold is the surface of the food, not the substance of it. The actual food as a three-dimensional object has a boundary and thus cannot be a manifold.

FIGURE 7.4. Mmmm, manifolds. The surfaces of these pastries are all two-dimensional closed manifolds, arranged by number of holes: none, one, two, three, and many. The manifold is the surface of the pastry, not the pastry itself. (The objects are *not* three-dimensional manifolds: points inside the pastry think they are the origin of some $\mathbb{R}^3$, while points on the boundary know they are not.)

As yet another intuitive way of understanding topological manifolds, pretend you live in the manifold and are unable to get out or even imagine what's outside. You have no compass or ruler for measuring things geometrically; all you have is a whole lot of breadcrumbs.[9] You could tell an 'H' from a 'T' (just count the forks

[9]We're assuming that nothing eats the breadcrumbs.

in the road) but not an 'I' from an 'S' (since all you could do is start at one end and go to the other end). Similarly you could tell apart the surface of sphere from the surface of a torus, and both of those from the surface of a two-holed torus. See Figure 7.4. Books like Edwin Abbott's classic *Flatland*[10] may help with this sort of intuition, if you're having trouble.

It's almost time for a definition. What we're trying to capture is the idea that for every point p in a manifold M, there should be an open set U containing p and a coordinate chart $\mathbf{x}\colon U \to \mathbb{R}^n$ such that $\mathbf{x}$ is a homeomorphism onto its image. (In other words, $\mathbf{x}$ is continuous and invertible, and $\mathbf{x}^{-1}\colon \mathbf{x}[U] \to M$ is also continuous.) This is the same definition we gave in Definition 7.1.1, and at first seems to capture what we want.

It's not good enough though. In abstract topology open sets can look kind of strange, so for example it's quite possible that U may be homeomorphic to something like the closed half-plane in $\mathbb{R}^n$, and the above definition would claim that the closed half-plane is a manifold. So we need to demand that the image $\mathbf{x}[U]$ in $\mathbb{R}^n$ is actually an open set we're familiar with.

Definition 7.1.2. (Preliminary definition) An *n-dimensional generalized topological manifold* is a topological space M such that for every point $p \in M$, there is an open set $U \ni p$ and a homeomorphism $\phi\colon U \to \mathbb{R}^n$ which maps U onto an open set in $\mathbb{R}^n$.

We will use symbols like ϕ to denote coordinate charts when we are interested in them abstractly; when we have concrete manifolds and coordinate charts in mind, we will generally use symbols like $\mathbf{x}$.

Now the most familiar open sets in $\mathbb{R}^n$ are open balls around the origin,

$$B_r(0) = \{x \in \mathbb{R}^n \mid \|x\| < r\},$$

and the entire space $\mathbb{R}^n$ itself. Conveniently these are homeomorphic to each other: you just take the map $\varphi\colon B_r(x) \to \mathbb{R}^n$ defined by $\varphi(x) = \frac{x}{r-\|x\|}$, the inverse of which is $\varphi^{-1}(y) = \frac{ry}{1+\|y\|}$. Furthermore by definition any open set in $\mathbb{R}^n$ contains an open ball and thus a set homeomorphic to $\mathbb{R}^n$. So the following seemingly more restrictive definition is actually equivalent to Preliminary Definition 7.1.2: M is an n-dimensional manifold if for every $p \in M$ there is an open $U \ni p$ and a homeomorphism $\phi\colon U \to \mathbb{R}^n$ which maps U onto *all* of $\mathbb{R}^n$.

Unfortunately, although we have a pretty good idea of what we intuitively want manifolds to look like, this definition allows for too much weird behavior. The next couple of examples will only really make sense if you have had a topology course already, so don't worry about them if not.

Example 7.1.3. "The long line." To understand this example, it helps to know something about ordinals. Richard Koch at the University of Oregon has an excellent and easy-to-understand writeup[11] of it if you're curious.

Roughly the idea is to take $\omega_1 \times [0,1)$, where ω_1 is the first uncountable ordinal, and treat it as though it's an uncountable number of copies of $[0,1)$ laid end to end. (We use ordinals because they are a well-ordered set, which implies that every element has an immediate predecessor.) Specifically we define an ordering by $(x,\alpha) < (y,\beta)$ if either $\alpha < \beta$, or $\alpha = \beta$ and $x < y$, and then give this space

[10] http://www.geom.uiuc.edu/∼banchoff/Flatland/

[11] http://www.math.ucsd.edu/∼nwallach/LongLine[1].pdf accessed January 18, 2013.

the order topology, a subbasis of which is any interval of the form $((x, \alpha), (y, \beta))$. Clearly around every point except $(0, 0)$ we can find an open interval homeomorphic to $\mathbb{R}$: just take all points less than (x, α): this is a union of countably many intervals $(0, 1)$.

So if we just glue together two copies (one on the left, one on the right), we get a one-dimensional manifold. See Figure 7.5 for a heuristic picture. The problem with it is that it's way too long (hence the name) and so for example it's impossible to define a distance on it which would generate the topology. It's also impossible to embed it into any Euclidean space. This comes from the fact that its topology is not "second countable" (i.e., there is no countable number of open sets which you can use, via taking arbitrary unions, to get all other open sets).

Now the reason one wants a manifold to have a second countable topology is that intuitively one doesn't want to have all that many coordinate charts. Ideally a manifold would be compact, and hence covered by finitely many coordinate charts. Failing that, the next best thing to ask is that we need only countably many coordinate charts, which would allow us to view the manifold as a countable union of compact sets. We'll find in general that compactness is quite useful, and so we don't want to be too far from a compact set.

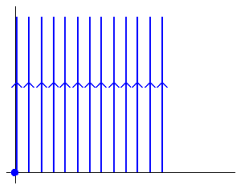

FIGURE 7.5. The "long ray" consists of a square with the lexicographic ordering. "Increasing" in the lexicographic ordering means going up along the vertical segments until the top, when one moves to the next vertical segment to the right. (Hence the need for well-ordering.) The long line is two copies of the long ray with the origin identified.

☺

Example 7.1.4. "The line with two origins."

In this example, we take the line and put an extra point O' next to the origin O. The line with two origins is $L = \mathbb{R} \cup O'$. See Figure 7.6. To define the topology, we take all of the usual open sets from $\mathbb{R}$, and for every open set U containing the origin O, we add the new set $U' = (U \backslash \{O\}) \cup O'$.

Since each U is homeomorphic to $\mathbb{R}$, so is U'. Thus L is a manifold. However, it is not Hausdorff: in other words, any open set containing O must intersect any open set containing O'. This is a very nonintuitive sort of property: two distinct points O and O' should have disjoint open sets enclosing them.

☺

Having seen these counterexamples, we now want to forget them. So now we give the actual definition of a manifold.

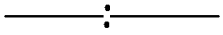

FIGURE 7.6. The line with two origins. On top is the actual point, and below is its identical evil twin. Or is it the other way around!?

Definition 7.1.5. (Actual definition) An *n-dimensional topological manifold* is a topological space M which is Hausdorff and second countable, and such that at each point $p \in M$ there is an open set $U \ni p$ and a homeomorphism ϕ from U onto $\mathbb{R}^n$.

The easiest way to understand these extra requirements in the definition is that they force the manifold to be metrizable. In other words, there is a distance function such that all open sets are the union of open balls in the metric. This is a consequence of the Urysohn metrization theorem. We will never actually care what the distance function is, but you can always imagine it being there in the background when we are talking about open sets. That is, you can always imagine that a subset U of M is open if and only if for every point $p \in U$, there is a number $r > 0$ such that

$$B_r(p) \equiv \{q \in M \,|\, d(p,q) < r\} \subseteq U,$$

where d is some distance function; you just have to realize that there are lots of different distance functions that will generate the exact same open sets, so you can't take any one of them too seriously.

As in Chapter 6, each pair (ϕ, U) or $(\mathbf{x}, U)$ is called a *coordinate chart*. Typically one needs more than one coordinate chart to cover the manifold, since the only manifold which can be covered by one chart is $\mathbb{R}^n$ itself.

Definition 7.1.6. A collection of coordinate charts whose domains cover M is called an *atlas* of M.

I've used a lot of topology in discussing this, which you don't really need to know. In fact one could (and historically many people did) define manifolds without saying anything about their topology in an abstract sense. Instead one requires the maps ϕ to be invertible, and continuity comes in an indirect way by looking at the transition maps $\phi \circ \psi^{-1}$, which are maps from one open subset of $\mathbb{R}^n$ to another. In the alternative definition, the requirement is that every such transition map be continuous, and this is easier since continuity of such maps is well-understood without any topology background. If one had a space satisfying such conditions, one could define the topology by declaring a set in M to be open if and only if its image under any coordinate chart was open in $\mathbb{R}^n$.

Definition 7.1.5 gives us topological manifolds, but we don't yet have a concept of smooth manifold. Now continuity can be put into fairly abstract terms, but for smoothness we really want to work strictly in terms of maps on $\mathbb{R}^n$, where we understand things. Thus smoothness will be defined only in terms of the transition functions.

Example 7.1.7. Consider the unit circle

$$S^1 = \{(x, y) \in \mathbb{R}^2 \,|\, x^2 + y^2 = 1\}.$$

Let's write down some coordinate charts and look at the transitions. I mentioned in Chapter 1 that there were a few common ways to set up maps of the 2-sphere, and in Figure 1.1 I demonstrated this schematically in the case of the 1-sphere. Now we're going to take seriously the stereographic projection, and redefine it slightly to make the formulas simpler.

Let $U = S^1 \backslash \{(0, -1)\}$; that is, delete the south pole from the circle. You should imagine cutting the circle there and then unwrapping it onto a line. We now construct a map $\phi \colon U \to \mathbb{R}$ using the following idea: draw a line from $(0, -1)$ through the point $(x, y) \in S^1$; this line eventually crosses the x-axis at coordinates $(u, 0)$, and the map is $\phi(x, y) = u$. Concretely ϕ is the restriction of the map $\Phi \colon \mathbb{R}^2 \backslash [\mathbb{R} \times \{-1\}] \to \mathbb{R}$ to S^1, where

$$u = \Phi(x, y) = \frac{x}{y + 1}. \tag{7.1.1}$$

Obviously ϕ is undefined at $(0, -1)$ since a line is determined by two *distinct* points.

We need another coordinate chart to cover all of S^1, and so we use the same idea the other way around: let $V = S^1 \backslash \{(0, 1)\}$ and let $\psi \colon V \to \mathbb{R}$ by $\psi = \Psi|_{S^1}$, where

$$v = \Psi(x, y) = \frac{x}{1 - y}. \tag{7.1.2}$$

Since $U \cup V = S^1$, the charts (ϕ, U) and (ψ, V) form an atlas of S^1.

On the intersection $U \cap V$, which is homeomorphic to two disjoint intervals in $\mathbb{R}$, there are two coordinate charts, and they need to be compatible. Notice that the point $(0, -1)$ that is missing from U has coordinates $\psi(0, -1) = 0$ in V. Similarly $\phi(0, 1) = 0$ in U. The transition map is thus $\psi \circ \phi^{-1}$ defined on $(-\infty, 0) \cup (0, \infty) \subset \mathbb{R}$. Let's compute it algebraically, as shown in Figure 7.7.

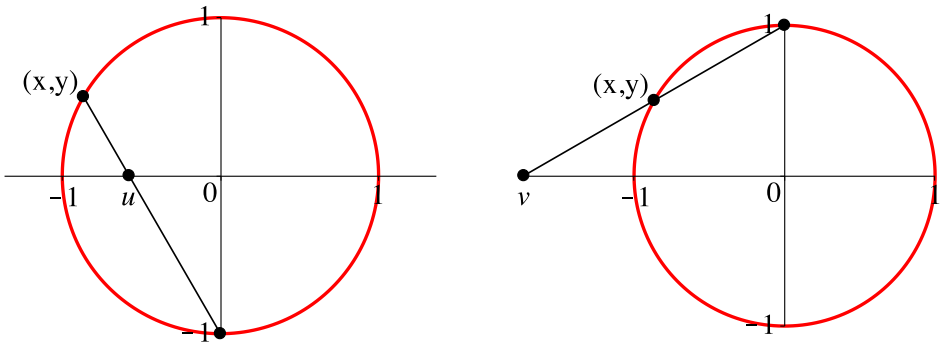

FIGURE 7.7. On the left, the chart (ϕ, U) defining the south-pole stereographic coordinate. On the right, the chart (ψ, V) defining the north-pole stereographic coordinate. We can check that $v = 1/u$.

From formula (7.1.1) we compute ϕ^{-1}: we just have to solve the equations $x^2+y^2=1$ and $x/(y+1)=u$ for (x,y), which yields

$$(x,y)=\phi^{-1}(u)=\left(\frac{2u}{1+u^2},\frac{1-u^2}{1+u^2}\right).$$

Plugging into (7.1.2), we obtain

$$v=\psi\circ\phi^{-1}(u)=\frac{\frac{2u}{1+u^2}}{1-\frac{1-u^2}{1+u^2}}=\frac{1}{u}. \tag{7.1.3}$$

As expected, this is defined for $\{u\in\mathbb{R}\,|\,u\neq 0\}$, which is $\phi[U\cap V]$. And certainly on this set it is not merely continuous but infinitely differentiable. Finally it is clear that $\phi\circ\psi^{-1}(v)=1/v$ since it must be the inverse function of (7.1.3). Thus (ϕ,U) and (ψ,V) are *smoothly compatible*, and every other acceptable coordinate chart on M should have smooth transition functions with one (and hence both) of these two. ☺

Definition 7.1.8. An *n-dimensional smooth manifold* is an n-dimensional topological manifold with an atlas of charts satisfying the following compatibility property: for each charts (ϕ,U) and (ψ,V) on M, the map

$$\phi\circ\psi^{-1}\colon \psi[U\cap V]\subset\mathbb{R}^n\to\phi[U\cap V]\subset\mathbb{R}^n$$

is C^∞.

We may start with a small number of charts, and add new ones that are compatible with the old ones in this sense; frequently we think of a smooth manifold as having a *maximal atlas* consisting of all possible coordinate charts which have smooth transition functions.

It's worth asking whether, given a topological manifold, there's always a smooth atlas for it. It seems like this should be true, but unfortunately it's not. If the dimension of the manifold is $n=1$, $n=2$, or $n=3$, this is true. (In fact we can classify all such manifolds, thanks to recent work of Perelman; we will discuss this in the next Chapter.) However in four dimensions there is a topological manifold called "the E_8 manifold" found by Freedman in 1982 which has no smooth structure. One can also ask whether a smooth structure is unique, if it is known to exist. This is also false, even if one asks for uniqueness only up to equivalence classes. The first counterexample found was the possibility of multiple structures on the 7-dimensional sphere S^7 (by Milnor in the 1950s); however it's now known that there are uncountably many nonequivalent smooth manifold structures even on $\mathbb{R}^4$.

Because of this, we will generally not worry about topological manifolds beyond this point. Any manifold I define will have a natural smooth structure, and the continuous structure will fall out of that.

7.2. **Practicalities.** In the next Chapter I'll discuss the best-known examples along with the classification results in low dimensions. For now I just want to discuss how to determine whether a particular set is a manifold. One pretty much never actually constructs the charts explicitly. Instead, the most common techniques are using the Inverse Function Theorem 5.2.4 and the Implicit Function Theorem 5.2.2.

The first way to define a manifold is via parametrization. We start with a function $F\colon U \subset \mathbb{R}^n \to \mathbb{R}^{n+k}$ which is meant to "embed" an n-dimensional manifold into an $(n+k)$-dimensional Euclidean space. To be precise:

Definition 7.2.1. If $F\colon U \subset \mathbb{R}^n \to \mathbb{R}^{n+k}$ is a C^∞ map on an open set U, it is called an *immersion* if for every $x \in U$ the differential $DF(x)$ has maximal rank as a linear map from $\mathbb{R}^n$ to $\mathbb{R}^{n+k}$.

We would like to say that the image of an immersion is a manifold, but this is not always true. An easy example is a self-intersecting curve in the plane, such as the lemniscate of Jacobi in Figure 7.8a, which is the image of

$$F(t) = \left(\frac{\cos t}{1+\sin^2 t}, \frac{\sin t \cos t}{1+\sin^2 t} \right)$$

for $t \in \mathbb{R}$. This curve has a self-intersection at the points corresponding to $t = \frac{\pi}{2}$ and $t = \frac{3\pi}{2}$, where it has two different tangent lines.

A trickier example is the folium of Descartes in Figure 7.8b, given by the image of

$$F(t) = \left(\frac{3t}{1+t^3}, \frac{3t^2}{1+t^3} \right)$$

for $t \in (-1, \infty)$. There's no obvious self-intersection one would see by looking at the equations algebraically, but in the limit as $t \to \infty$ the loop closes up to meet the point where $t = 0$, at the origin. Hence the image is not a manifold, since in any neighborhood of $(0,0)$ the image looks like a 'T.'

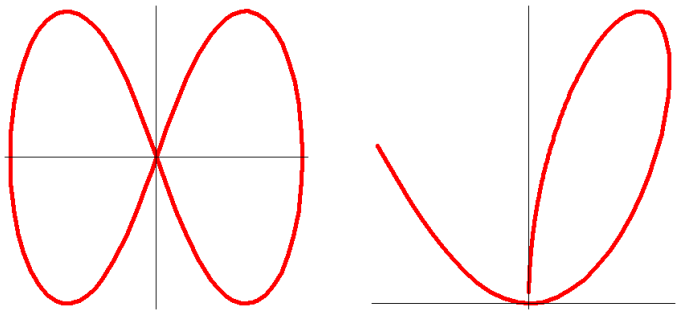

FIGURE 7.8. On the left, the lemniscate of Jacobi, which fails to be a manifold because of a self-intersection with different tangent lines. On the right, the folium of Descartes, the image of which is not a manifold because of the asymptotic self-intersection.

Still one wants to be able to get a manifold by imposing the right conditions. The immersion condition is meant to accomplish most of this (it will be used, via the Inverse Function Theorem 5.2.4, to construct a coordinate chart in a neighborhood of each point in the image), and it always gives good local behavior, but some sort of global condition is also needed to avoid the problem of self-intersections. Even with the above examples in mind, this is harder than it seems. One might allow for intersections—they're impossible to avoid if one wants to get any interesting topology out of a subset of the plane—but just require that the derivatives match up. This would prevent crossing of skew tangent lines, which seems to be mucking

things up above. But then one could imagine first derivatives matching and not higher derivatives. Even matching derivatives of all orders doesn't work, as the following famous example shows.

Example 7.2.2. Let $f\colon \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} 0 & x = 0, \\ e^{-1/x^2} & x \neq 0. \end{cases} \tag{7.2.1}$$

You can compute directly that $\lim_{x\to 0} f(x) = 0$, that $f'(0)$ exists from the definition, that $f'(x)$ is continuous, that $f''(0)$ exists, etc. In this way you can see that f is a C^∞ function. However all derivatives of f at $x = 0$ are zero.

This is an extremely important example and will be used later for various purposes. We will discuss it in much greater depth in Example 13.2.1. For right now, just notice that it's a function that is C^∞ but not real analytic: its Maclaurin series converges to $g(x) = 0$ rather than $f(x)$.

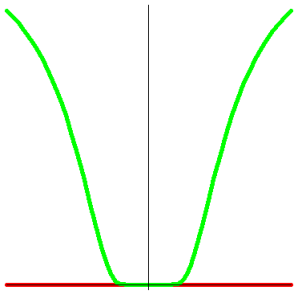

FIGURE 7.9. The graph of $y = e^{-1/x^2}$, shown in green. It is impossible to distinguish this from the graph of $y = 0$ just by computing derivatives at $x = 0$.

☺

There's no great solution to this problem, and whether a particular parametrization gives an actual manifold has to be worked out case by case. Here's an example.

Example 7.2.3. (The Möbius band) Consider $U = \mathbb{R} \times (-1, 1) \subset \mathbb{R}^2$, an infinite horizontal band. Let $F\colon U \to \mathbb{R}^3$ be given by

$$(x, y, z) = F(u, v) = \left(\left(1 + \tfrac{v}{2}\cos\tfrac{u}{2}\right)\cos u, \left(1 + \tfrac{v}{2}\cos\tfrac{u}{2}\right)\sin u, \tfrac{v}{2}\sin\tfrac{u}{2}\right).$$

The image of this is called the Möbius band (or Möbius strip) and is shown in Figure 7.10. Notice that the boundary is $v = 1$ and $v = -1$, which are not included (otherwise it obviously couldn't be a manifold).

It looks like it's a manifold, but how would we check? The first thing to do is verify that F is an immersion. So we compute:

$$DF(u, v) = \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \\ \frac{\partial z}{\partial u} & \frac{\partial z}{\partial v} \end{bmatrix} = \begin{bmatrix} -\frac{v}{4}\sin\frac{u}{2}\cos u - \left(1 + \frac{v}{2}\cos\frac{u}{2}\right)\sin u & \frac{1}{2}\cos\frac{u}{2}\cos u \\ -\frac{v}{4}\sin\frac{u}{2}\sin u + \left(1 + \frac{v}{2}\cos\frac{u}{2}\right)\cos u & \frac{1}{2}\cos\frac{u}{2}\sin u \\ \frac{v}{4}\cos\frac{u}{2} & \frac{1}{2}\sin\frac{u}{2} \end{bmatrix}$$

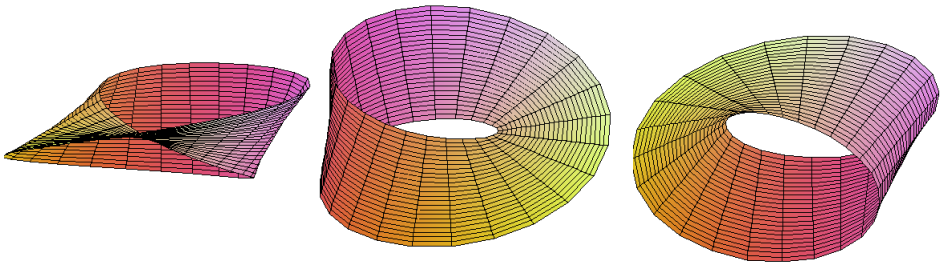

FIGURE 7.10. The Möbius band seen from a few different angles.

To compute the rank of this matrix, we compute the upper 2×2 determinant, which yields $-\frac{1}{4}\cos\frac{u}{2}\left(2 + v\cos\frac{u}{2}\right)$. This is nonzero as long as $\cos(u/2) \neq 0$, which is true as long as u is not an odd integer multiple of π. On the other hand if $u = (2k+1)\pi$ for some integer k, then

$$DF((2k+1)\pi, v) = \begin{bmatrix} \frac{v}{4} & 0 \\ -1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix},$$

and the last two rows clearly give a 2×2 invertible matrix. Thus DF always has rank two, using Proposition 3.3.6. So F is an immersion.

Now the map F is certainly not bijective from $\mathbb{R}^2$ to $\mathbb{R}^3$ since $F(u + 4\pi, v) = F(u, v)$ for any (u, v). This is a trivial self-intersection, in much the same way that the standard parametrization of the unit circle intersects itself. There is also a nontrivial intersection: $F(u+2\pi, v) = F(u, -v)$ for any (u, v). It is not hard to check that these are essentially the only possibilities: $F(u, v) = F(p, q)$ for $-1 < v, q < 1$ implies that either $(u, v) = (p + 4m\pi, q)$ or $(u, v) = (p + 2\pi(2m+1), -q)$ for some integer m.

From this we obtain the following: if we restrict the domain to $U_1 = (0, 2\pi) \times (-1, 1)$, then $F|_{U_1} : U_1 \to \mathbb{R}^3$ is one-to-one and of maximal rank. Hence there is a coordinate chart on the subset $V_1 = F[U_1]$ of the Möbius band M back to U_1 which is smooth by the Inverse Function Theorem 5.2.4. The portion missing is

$$M \backslash V_1 = \{F(0, v) \mid -1 < v < 1\} = \left\{(x, 0, 0) \mid \tfrac{1}{2} < x < \tfrac{3}{2}\right\}.$$

Similarly if we restricted the domain to $U_2 = (-\pi, \pi) \times (-1, 1)$, then $F|_{U_2} : U_2 \to \mathbb{R}^3$ generates a coordinate chart on another part $V_1 = F[U_2]$ of the Möbius band, and the missing portion is the set

$$M \backslash V_2 = \{F(\pi, v) \mid -1 < v < 1\} = \left\{(x, 0, 0) \mid -\tfrac{3}{2} < x < -\tfrac{1}{2}\right\}.$$

Hence $V_1 \cup V_2 = M$, and these two charts cover the entire Möbius band.

Let's now look at the transition function and verify that it is smooth. Let $\zeta(u, v) = F|_{U_2}^{-1} \circ F|_{U_1}(u, v)$. Its domain Ω is the inverse image of $V_1 \cap V_2$ under $F|_{U_1}$, which means it's the set $(0, 2\pi) \times (-1, 1)$ with the set $\{\pi\} \times (-1, 1)$ deleted, since $F[\{\pi\} \times (-1, 1)] = M \backslash V_2$. In other words the domain consists of the disjoint rectangles $\Omega = \Omega_1 \cup \Omega_2$ where $\Omega_1 = (0, \pi) \times (-1, 1)$ and $\Omega_2 = (\pi, 2\pi) \times (-1, 1)$. The difference between these two sets is that Ω_1 is a subset of both U_1 and U_2, while Ω_2 is only a subset of U_1. We thus have $\zeta(u, v) = (u, v)$ if $(u, v) \in \Omega_1$, while if

$(u, v) \in \Omega_2$ then $\zeta(u, v) = (u - 2\pi, -v)$. Obviously these formulas are smooth and invertible, so we get a differentiable manifold.

To visualize what's going on, consider the rectangle $X = [0, 2\pi] \times (-1, 1)$. It's easy to see that the image $F[X]$ is all of M. Furthermore, since $F(0, v) = F(2\pi, -v)$ on X, we can think of F as gluing the left and right sides of X together with a twist, where for example point $(0, \frac{1}{2})$ gets glued to $(2\pi, -\frac{1}{2})$. See the diagram in Figure 7.11.

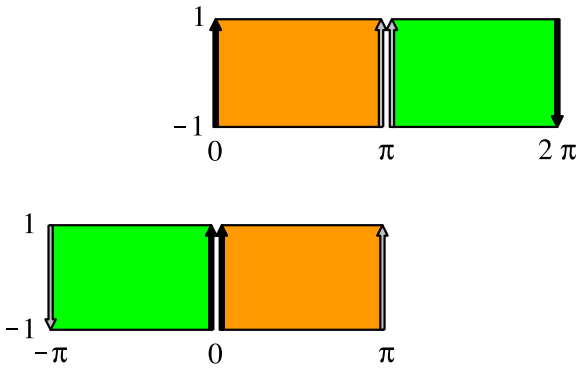

FIGURE 7.11. The Möbius band coordinate charts: the set U_1 is on top and the set U_2 is on the bottom. Imagine gluing the set above along the gray arrows first to get the set $X = [0, 2\pi] \times (-1, 1)$, then pulling around and twisting to glue along the black arrows to get the Möbius strip in Figure 7.10. The transition from U_1 to U_2 corresponds to separating U_1 at the midpoint, sliding the green rectangle around to the other side, then flipping it over vertically to get the black arrows to match up. Hence the transition map ζ is the identity on the orange set and $(u, v) \mapsto (u - 2\pi, -v)$ on the green set.

☺

The method in Example 7.2.3 is obviously rather cumbersome, although it gives us a nice explicit picture of coordinate charts and their transitions in a simple but not trivial case. The essential features which made this work are that F is an immersion (its derivative has maximal rank everywhere), and that F is invariant under the action of a discrete group, generated in this case by $(u, v) \mapsto (u+2\pi, -v)$. This works because the universal cover of the Möbius strip is the strip $\mathbb{R} \times (-1, 1)$. We will do the same thing more generally in Theorem 9.1.7.

Clearly the parametrization technique is fraught with problems, although the idea of finding a fundamental domain ends up being quite powerful. And if it works, the parametrization automatically gives convenient coordinate charts on a subset of the original domain. Still it's much more complicated for most cases. Imagine trying to prove the same sort of thing with the sphere (where a parametrization by spherical coordinates has actual singularities).

So we have the following alternative technique.

Definition 7.2.4. Suppose $F\colon \mathbb{R}^{n+k} \to \mathbb{R}^k$ is a C^∞ map. Then $r \in \mathbb{R}^k$ is a *regular value* if for every $x \in F^{-1}(r) \subset \mathbb{R}^{n+k}$, the rank of $DF(x)\colon \mathbb{R}^{n+k} \to \mathbb{R}^k$ is k. If r is not a regular value, it is called a *singular value*.

For example, if $F\colon \mathbb{R}^3 \to \mathbb{R}$ is $F(x,y,z) = x^2+y^2+z^2$, then $DF(x,y,z) = (2x \quad 2y \quad 2z)$, and this will have rank one iff not all of x, y, z are zero. If $r > 0$ then $F^{-1}(r)$ consists of points with $x^2+y^2+z^2 = r$, so that $DF(x,y,z)$ has rank one for all such points. Thus any $r > 0$ is a regular value. If $r = 0$ then $F^{-1}(r)$ consists only of $(0,0,0)$ and then $DF(0,0,0)$ has rank zero. So $r = 0$ is a singular point. If $r < 0$ then $F^{-1}(r)$ is empty, but then the regular value condition is trivially satisfied. So any $r < 0$ is also a regular value but for a different reason.

Theorem 7.2.5. *Suppose $F\colon \mathbb{R}^{n+k} \to \mathbb{R}^k$ is a C^∞ map. Suppose also that $r \in \mathbb{R}^k$ is a regular value of F. If $F^{-1}(r)$ is not empty, then it is a C^∞ n-dimensional manifold.*

Proof. This is just the Implicit Function Theorem. Let $p \in F^{-1}(r)$ be any point; then $DF(p)$ has rank k. So by rotating the domain $\mathbb{R}^{n+k}$ as in Proposition 3.3.6 and the discussion before Theorem 5.2.2, we can assume that the right $k \times k$ submatrix of $DF(p)$ is nonsingular. Writing $p = (a,b)$, the hypotheses of Theorem 5.2.2 are satisfied, so there is an open set $V \subset \mathbb{R}^n$ containing a and a smooth function $G\colon V \to \mathbb{R}^k$ such that $G(a) = b$ and $F\big(x, G(x)\big) = r$ for every $x \in V$.

The inverse of the coordinate chart will now be $g\colon V \to \mathbb{R}^{n+k}$ given by

$$g(u_1, \cdots, u_n) = \big(u_1, \cdots, u_n, G_1(u_1, \cdots, u_n), \cdots, G_k(u_1, \cdots, u_n)\big),$$

and g maps V into $F^{-1}(r)$, giving a parametrization of $F^{-1}(r)$ in a neighborhood of p. □

Example 7.2.6. The standard unit sphere S^2 is defined by

$$S^2 = \{(x,y,z) \in \mathbb{R}^3 \,|\, x^2+y^2+z^2 = 1\}.$$

There are two ways to put a manifold structure on S^2: one is to use north-pole and south-pole stereographic coordinates as in the one-dimensional case from Example 7.1.7. This gives us two coordinate charts that together cover S^2, which is certainly sufficient to define a smooth manifold structure.

The other way is to use the Implicit Function Theorem as in Theorem 7.2.5. Define $F\colon \mathbb{R}^3 \to \mathbb{R}$ by $F(x,y,z) = x^2+y^2+z^2$. Then the derivative operator is

$$DF(x,y,z) = \begin{pmatrix} 2x & 2y & 2z \end{pmatrix},$$

and as long as at least one of the components is nonzero, this matrix has rank one, which is maximal. If $F(x,y,z) = r = 1$, i.e., $x^2+y^2+z^2 = r = 1$, then not all components are zero, so $r = 1$ is a regular value. What charts do we obtain from Theorem 7.2.5?

Suppose (x_0, y_0, z_0) is a point in S^2 with $z_0 \neq 0$. Then the right 1×1 submatrix of $DF(x_0, y_0, z_0)$ is nonsingular, and there is an open set $V \subset \mathbb{R}^2$ containing the point (x_0, y_0) and a smooth function $G\colon V \to \mathbb{R}$ such that $G(x_0, y_0) = z_0$ and $x^2+y^2+G(x,y)^2 = 1$. Clearly if $z_0 > 0$ then $G(x,y) = \sqrt{1-x^2-y^2}$ on the open set $V = \{(x,y) \in \mathbb{R}^2 \,|\, x^2+y^2 < 1\}$. This gives the usual parametrization of the top hemisphere, and if $z_0 < 0$ we would get the parametrization of the bottom hemisphere.

We then obtain a map $g\colon V \to \mathbb{R}^3$ given by

$$g(u,v) = \big(u, v, \sqrt{1-u^2-v^2}\big).$$

The actual coordinate chart $\phi\colon S^2 \to \mathbb{R}^2$ is the restriction of $\Phi\colon \mathbb{R}^3 \to \mathbb{R}^2$ given by $\Phi(x,y,z) = (x,y)$: we have $\phi = \Phi\big|_{S^2 \cap [\mathbb{R}^2 \times (0,\infty)]}$. (Obviously ϕ makes sense on all of S^2, but it is only invertible on the upper half or lower half.)

The assumption $z_0 \neq 0$ thus yields two coordinate charts that cover the open top hemisphere and the open bottom hemisphere, which together leave out the "equator" circle when $z = 0$. To cover the rest of S^2 with the Implicit Function technique, we need to be able to switch the coordinates when $z = 0$. This process yields a pair of coordinate charts for $x > 0$ and $x < 0$, and another pair of charts when $y > 0$ and $y < 0$. Thus in total we obtain six coordinate charts using this technique. See Figure 7.12 for an illustration of these six charts.

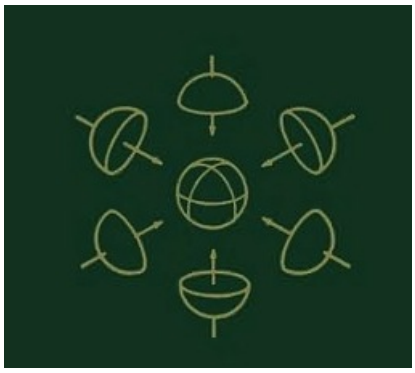

FIGURE 7.12. The six hemispherical coordinate charts arising from applying Theorem 7.2.5. This illustration is from the cover of M.P. do Carmo's *Riemannian Geometry*.

Smoothness of the transition functions follows from Theorem 7.2.5, but let's check explicitly between the hemisphere $x > 0$ and the hemisphere $z < 0$. Set $U = S^2 \cap (0,\infty) \times \mathbb{R}^2$ and $V = S^2 \cap \mathbb{R}^2 \times (-\infty, 0)$. Then the charts are (ϕ, U) with $\phi(x,y,z) = (y,z)$ and (ψ, V) with $\psi(x,y,z) = (x,y)$, and we have

$$(p,q) = \psi \circ \phi^{-1}(u,v) = \psi\big(\sqrt{1-u^2-v^2}, u, v\big) = \big(\sqrt{1-u^2-v^2}, u\big)$$

which is defined on the set $\{(u,v) \in \mathbb{R}^2 \,|\, u^2+v^2 < 1, v < 0\}$ and has smooth inverse

$$(u,v) = \phi \circ \psi^{-1}(p,q) = \big(q, -\sqrt{1-p^2-q^2}\big).$$

☺

Our final example uses a little of both techniques: the parametrization and quotient method of Example 7.2.3 and the Implicit Function method of Example 7.2.6.

Example 7.2.7. The 2-torus is denoted by $\mathbb{T}^2$. There are a variety of ways to define it, but the simplest is as the subset

$$\mathbb{T}^2 = \{(w,x,y,z) \in \mathbb{R}^4 \,|\, w^2 + x^2 = 1, y^2 + z^2 = 1\}.$$

If $F\colon \mathbb{R}^4 \to \mathbb{R}^2$ is given by $F(w,x,y,z) = (w^2+x^2, y^2+z^2)$, then $\mathbb{T}^2 = F^{-1}(1,1)$, and it is easy to verify that $DF(w,x,y,z)$ always has rank two on $\mathbb{T}^2$. We then use Theorem 7.2.5 to show that $\mathbb{T}^2$ is a smooth manifold.

Alternatively we can define $G\colon \mathbb{R}^2 \to \mathbb{R}^4$ by the formula

$$G(u,v) = (\cos u, \sin u, \cos v, \sin v).$$

We check that $DG(u,v)$ always has rank two, so it's an immersion, and that $G(u+2m\pi, v) = G(u, v+2n\pi) = G(u,v)$ for any integers m and n. We then build coordinate charts using sets like $(0,2\pi)\times(0,2\pi)$ as in Example 7.2.3.

Of course it's hard to visualize $\mathbb{T}^2$ in $\mathbb{R}^4$, so a more common visualization is through a parametrization like $H\colon \mathbb{R}^2 \to \mathbb{R}^3$ given by

$$H(u,v) = \big((a+b\cos u)\cos v, (a+b\cos u)\sin v, b\sin u\big),$$

where $a > b$. This gives the usual doughnut picture, although geometrically it's not nearly so nice. For example, near the end of the text we will be able to show that the embedding in $\mathbb{R}^4$ gives a flat surface while the embedding in $\mathbb{R}^3$ is intrinsically curved. ☺

8. Low-dimensional examples of manifolds

"I don't like sand. It's coarse and rough and irritating, and it gets everywhere. Not like here. Here everything's soft, and smooth."

In this Chapter and the next, we will go over the basic examples that every student of differential geometry is expected to see.

In each case the classification of manifolds relies on various topological properties. This will only work when the manifold is one-, two-, or three-dimensional, since as mentioned above every topological manifold is a smooth manifold in those dimensions but not in higher dimensions. Furthermore in higher dimensions there are too many possibilities to even construct an algorithm for distinguishing manifolds, and therefore there cannot be a classification of them.[12]

Generally the easier case is when the manifold is compact. In the usual (classical) terminology, a compact connected manifold is called *closed*.[13]

8.1. **One dimension.** There are only two connected one-dimensional manifolds satisfying Definition 7.1.5. The distinguishing feature is compact vs. noncompact, so we get either S^1 or $\mathbb{R}$. Note that the Hausdorff property and second-countability are essential for this classification to work—otherwise the long line (Example 7.1.3) and the line with two origins (Example 7.1.4) would violate the classification. We will see in the proof exactly where these assumptions are used.

First we need a classification theorem which is fairly well-known in one-variable real analysis.

Lemma 8.1.1. *Every open set Ω in $\mathbb{R}$ is a countable union of disjoint open intervals: $\Omega = \bigcup_{n=1}^{N}(a_n, b_n)$ where N is either a natural number or infinity.*

Proof. We will just sketch the proof. Given any point $x \in \Omega$, let $b_x = \sup\{y \,|\, (x, y) \subset \Omega\}$ and $a_x = \inf\{y \,|\, (y, x) \subset \Omega\}$. Show that $(a_x, b_x) \subset \Omega$, and that if $y \in (a_x, b_x)$, then $a_y = a_x$ and $b_y = b_x$. Hence either $(a_x, b_x) = (a_y, b_y)$ or $(a_x, b_x) \cap (a_y, b_y) = \emptyset$ for all $x, y \in \Omega$. Every open interval contains a rational number, so there can't be more than countably many such open intervals. Finally Ω must actually equal the union of all these intervals. □

First we use Lemma 8.1.1 to prove a Lemma about coordinate charts on a one-dimensional (Hausdorff) manifold.

Lemma 8.1.2. *Let (ϕ, U) and (ψ, V) be two coordinate charts on a one-dimensional Hausdorff manifold M with $\phi[U] = \psi[V] = \mathbb{R}$. Suppose U overlaps V but that neither U nor V is a subset of the other. Then $U \cap V$ is either homeomorphic to a single open interval or two disjoint open intervals. In the first case $U \cup V$ is homeomorphic to $\mathbb{R}$, and in the second case $U \cup V$ is homeomorphic to S^1.*

Proof. Let W be one component of $U \cap V$, so that W is homeomorphic to an open interval; then both $I = \phi[W]$ and $J = \psi[W]$ are open intervals in $\mathbb{R}$. I claim these intervals must be half-infinite. Assume to get a contradiction that $I = (a, b)$

[12] A.A. Markov, "Insolubility of the Problem of Homeomorphy," English translation at http://www.cs.dartmouth.edu/∼afra/goodies/markov.pdf

[13] Of course, as a topological space, every manifold is closed, so this is a genuine difference in terminology.

for finite a, b. Let $J = (c, d)$ where c or d may be infinite. If J is all of $\mathbb{R}$, then $\psi[W] = \psi[V]$ which means $V = W \subset U$, contradicting the fact that neither U nor V is a subset of the other. So either d is finite or c is finite; assume without loss of generality that d is finite.

The map $\psi \circ \phi^{-1} \colon (a, b) \to (c, d)$ is a homeomorphism, which is either strictly increasing or strictly decreasing. We can assume it's increasing. Then we must have

$$\lim_{t \to b^-} \psi \circ \phi^{-1}(t) = d. \tag{8.1.1}$$

Let $\delta = \psi^{-1}(d) \in V$ and $\beta = \phi^{-1}(b) \in U$. The point δ cannot be in U; if it were, then we'd have $\delta \in U \cap V$, so that $\delta \in W$ and thus $d \in J$, contradicting $J = (c, d)$. Similarly β cannot be in V, because if it were then $\beta \in W$ and thus $b \in I$. I want to say that $\beta = \delta$ to get a contradiction. See Figure 8.1.

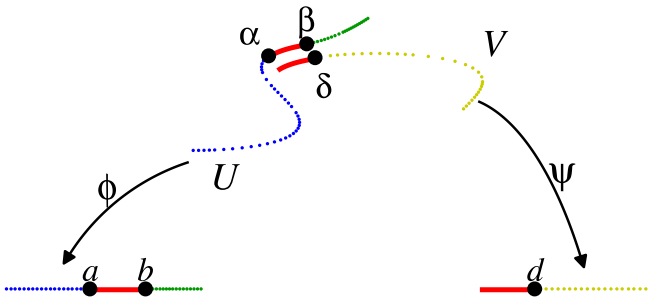

FIGURE 8.1. A possible schematic of what must happen if $\phi[U \cap V]$ is a bounded subset of $\mathbb{R}$. Here there are two copies of the open set $W = U \cap V$ shown in red; β is an endpoint of W in U and δ is an endpoint of W in V. Since $\beta \in U$ and $\delta \in V$, and β must equal δ by the Hausdorff property, we see that $\beta = \delta \in W$, which contradicts the fact that these are endpoints of W.

If $\beta \neq \delta$, then by the Hausdorff property, there would be disjoint open sets B and D in M such that $\beta \in B$ and $\delta \in D$. Now $\phi[B]$ contains b, so it has points slightly smaller than b; by (8.1.1) all points sufficiently close to b and smaller than b must end up as close as we want to d under the map $\psi \circ \phi^{-1}$. But $\psi[D]$ is an open set containing d, which means there is at least one point $x \in \phi[B]$ such that $\psi \circ \phi^{-1}(x) \in \psi[D]$. In other words, the point $\phi^{-1}(x)$ is in both B and D, contradiction. Thus we actually have $\beta = \delta$, and this contradicts the fact that $\beta \in U \backslash V$ and $\delta \in V \backslash U$.

We derived all this from the assumption that $\phi[W] = (a, b)$ where a and b are both finite. Thus if U and V are overlapping charts on M, neither of which is a subset of the other, then any component of $U \cap V$ must map under both ϕ and ψ to at least a half-infinite interval in $\mathbb{R}$. Hence in particular there can't be more than two components of $U \cap V$.

This tells us that exactly one of the following happens for the set $I = \phi[U \cap V]$:

(1) I is all of $\mathbb{R}$,

(2) $I = (b, \infty)$ for some finite b,
(3) $I = (-\infty, a)$ for some finite a,
(4) $I = (-\infty, a) \cup (b, \infty)$ for $-\infty < a \leq b < \infty$.

Case (1) is impossible since it implies $U \cap V = U$, i.e., that $U \subset V$.

In cases (2) or (3), we can easily see that $U \cup V$ must be homeomorphic to $\mathbb{R}$. For example in case (2) we know that $J = \psi[U \cap V]$ must be homeomorphic to (b, ∞) and must be half-infinite, so it must look like $(-\infty, a)$ or (c, ∞). Since any two open intervals in $\mathbb{R}$ are homeomorphic whether finite or half-infinite or infinite, we can compose φ and ψ with homeomorphisms to obtain homeomorphisms $\tilde{\varphi}\colon U \to (-\infty, 1)$ and $\tilde{\psi}\colon V \to (0, \infty)$ such that $\tilde{\phi}[U \cap V] = \tilde{\psi}[U \cap V] = (0, 1)$. We then define the map $\zeta\colon U \cup V \to \mathbb{R}$ by setting $\zeta = \phi$ on U and $\zeta = \psi$ on $V \backslash U$; then ζ is continuous since the limit of ϕ and the value of ψ match up at the point $\beta = \psi^{-1}(1)$. See Figure 8.2.

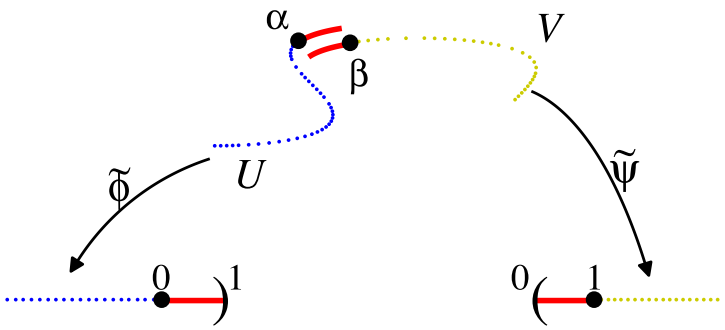

FIGURE 8.2. As in Figure 8.1, if we have charts that overlap in exactly one interval, then we can arrange it so that U maps to $(-\infty, 1)$ and V maps to $(0, \infty)$. Combining the maps gives a chart on $U \cup V$.

Finally in case (4) I claim $U \cup V$ is homeomorphic to S^1, as in Example 7.1.7. Obviously $\psi[U \cap V]$ must also be at least half-infinite, and since it is homeomorphic to two disjoint open sets, it must also consist of two disjoint open sets; call them $\psi[U \cap V] = (-\infty, c) \cup (d, \infty)$ with $c \leq d$. We have $\alpha = \phi^{-1}(a) \notin V$ and $\beta = \phi^{-1}(b) \notin V$, along with $\gamma = \psi^{-1}(c) \notin U$ and $\delta = \psi^{-1}(d) \notin U$. See Figure 8.3 for an illustration, which should make it clear how to write down a homeomorphism from $U \cup V$ to S^1.

□

Theorem 8.1.3. *The only connected one-dimensional topological manifolds are S^1 and $\mathbb{R}$.*

Proof. Obviously $\mathbb{R}$ is a smooth manifold with the identity map as coordinate chart. That S^1 is a smooth manifold follows from our computation of the coordinate charts in Example 7.1.7; alternatively we can use its expression as $F^{-1}(1)$ where $F(x, y) = x^2 + y^2$, along with Theorem 7.2.5, exactly as we did with S^2. The circle is clearly connected, since it's the image of $\mathbb{R}$ under the parametrization $(x, y) = (\cos t, \sin t)$.

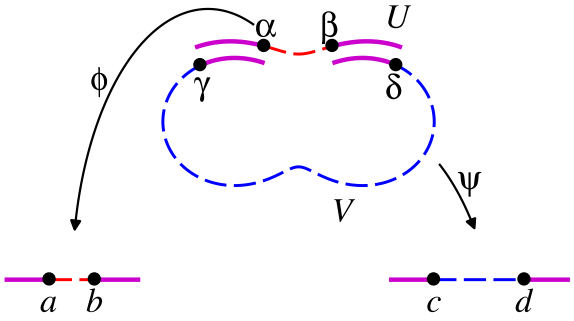

FIGURE 8.3. In case (4), the union $U \cup V$ must be homeomorphic to a circle. Here the purple parts represent $U \cap V$.

Now suppose M is a connected one-dimensional manifold. We will prove M is homeomorphic to S^1 or $\mathbb{R}$. Since it's a manifold, it has coordinate charts consisting of open sets homeomorphic to $\mathbb{R}$. Our goal will be to reduce the number of coordinate charts actually needed to either one or two; in the former case we should get $\mathbb{R}$ and in the latter we should get S^1.

Now we put it all together. Since M is second countable, it is Lindelöf; in other words every open cover has a countable subcover. Thus we need only countably many of the coordinate charts to cover all of it. So we can write $M = \cup_i U_i$ where there are finitely many or countably many i. Start with U_1; if $U_1 \cap U_j$ is empty for every $j > 1$, then we can write M as the disjoint union of U_1 and $\cup_{j>1} U_j$, which contradicts the assumption that M is connected. So there is a smallest j such that $U_1 \cap U_j \neq \emptyset$, and we might as well call this U_2. From above, either $U_1 \cup U_2$ is homeomorphic to $\mathbb{R}$ or $U_1 \cup U_2$ is homeomorphic to S^1. In the former case we can set $U_1' = U_1$ and $U_2' = U_1 \cup U_2$, then look for the next set (suppose it's U_3) which intersects it. If $U_2' \cup U_3$ is not homeomorphic to S^1, set $U_3' = U_2' \cup U_3$. Continuing in this way, we either get a countable nested union of intervals each of which is homeomorphic to $\mathbb{R}$, in which case the entire manifold is homeomorphic to $\mathbb{R}$, or we stop at some point because we have found a copy of S^1 inside M. (If that happens, then all other open sets which intersect U_n' must be proper subsets.) □

In the proof above we can see clearly where the assumptions that M is second countable and Hausdorff enter, as they must: otherwise we could get the line with two origins or the long line as in Examples 7.1.3–7.1.4. But already the proof is rather long. In higher dimensions it will be impossible to give a direct proof like this, and instead we have to apply techniques from algebraic topology.

8.2. **Two dimensions.** The two-dimensional manifolds have been completely classified. The easiest to understand are the closed (i.e., compact) manifolds, of which the foods in Figure 7.4 are the simplest examples.

The essential idea for this classification relies on the idea of triangulation, which is a way of systematically approximating a manifold by a simplicial complex. The way in which we do this is very similar to the representation of the Möbius band in Example 7.11, as a polygon in $\mathbb{R}^2$ with certain sides identified. An example of what

we're going for is shown in Figure 8.4. Notice the following properties which are going to make counting things much easier: the intersection of any pair of triangles is either empty, exactly one edge, or exactly one vertex. Notice also that if we unhinge the triangle at a few of the edges, we can unfold it into a planar figure, as garishly illustrated below. (Of course, after cutting it at various edges, we have to remind ourselves to glue them back together.) The planar figure is substantially easier to work with than an embedded polyhedron in three (or four) dimensions. Once we have a planar figure for a triangulation, we can eliminate some edges while still keeping the figure planar. In this way we end up representing a triangulated surfaces as a polygon with some sides identified.

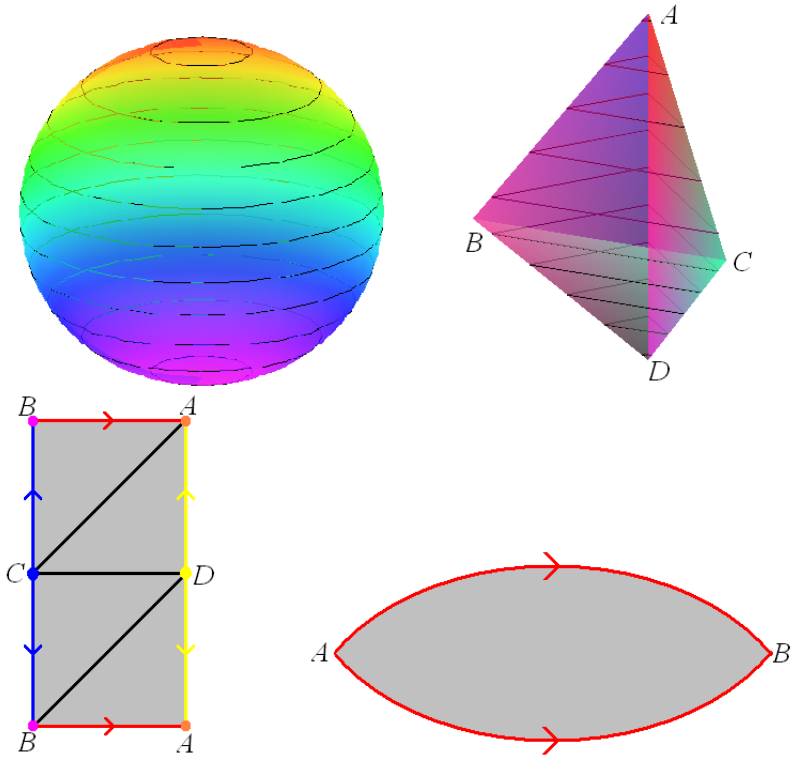

FIGURE 8.4. On top, the round sphere and its triangulation as a polyhedron in $\mathbb{R}^3$. On the bottom, a planar diagram of the triangulation, and a simplified version: the coin-purse model, which we obtain by zipping up the edges on the triangulation diagram to remove the C and D vertices. (If we zipped up the edges here, it would close up to give a topological sphere.)

The smallest triangulation of the torus has 14 faces, as shown in Figure 8.5. The simplified planar model is shown next to it. You can check that any faces intersect in only one edge or only one point. If you try to use fewer triangles (for example, if you tried dividing the planar model into four squares and cut each square in half diagonally), the triangles end up having too many intersections.

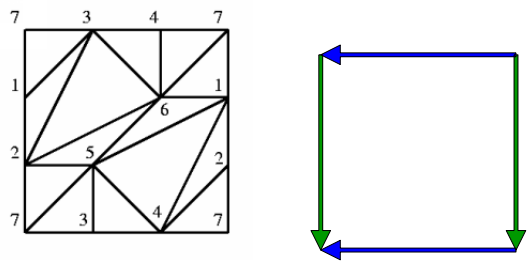

FIGURE 8.5. The minimal-face triangulation of the torus due to Möbius, and the simplified version.

Now that we have an intuitive idea of what a triangulation is, it's time for some definitions.

First is the notion of quotient space. I used this already in Example 7.2.3, to think of the Möbius band as a certain quotient space of $[0, 2\pi] \times (-1, 1)$ modulo the equivalence $(0, v) \cong (2\pi, -v)$, or alternatively as $\mathbb{R} \times (-1, 1)$ modulo the equivalence

$$(x, y) \cong (x + 2n\pi, (-1)^n y) \quad \text{for any } n \in \mathbb{Z}. \tag{8.2.1}$$

(The former is convenient since there is only one pair of edges to be identified; the latter is convenient since the equivalence is generated by a group action.)

Definition 8.2.1. An *equivalence relation* is a relation $\cong$ satisfying the properties

(1) $p \cong p$ for all $p \in M$;
(2) $p \cong q \Leftrightarrow q \cong p$, for all $p, q \in M$;
(3) If $p \cong q$ and $q \cong r$, then $p \cong r$.

The *equivalence class* of a point p is

$$[p] = \{q \in M \mid q \cong p\}.$$

We often write the projection from the set of all points to the set of all equivalence classes as $\pi(p) = [p]$.

The *quotient space* is a topological space Q consisting of all equivalence classes; a set U in Q is open if and only if $\pi^{-1}(U) = \{p \in M \mid [p] \in U\}$ is open in M.

Example 8.2.2. The Möbius band quotient relation is actually an equivalence relation. Explicitly, first $(x, y) \cong (x, y)$ since $n = 0 \in \mathbb{Z}$. Second, if $(x, y) \cong (p, q)$ then $p = x + 2n\pi$ and $q = (-1)^n y$ for some $n \in \mathbb{Z}$, and therefore $(p, q) \cong (p - 2n\pi, (-1)^{-n} q) = (x, y)$ since $-n$ is in $\mathbb{Z}$ whenever n is. Third, if $(x, y) \cong (p, q)$ and $(p, q) \cong (c, d)$, then for some integers m and n we know $p = x + 2m\pi$, $c = p + 2n\pi$, $q = (-1)^m y$, and $d = (-1)^n q$. Therefore $c = x + 2(m + n)\pi$ and $d = (-1)^{m+n} y$, so $(x, y) \cong (c, d)$.

Similarly the torus quotient relation $(x, y) \cong (x + 2m\pi, y + 2n\pi)$ for any $m, n \in \mathbb{Z}$ is an equivalence relation.

However, the quotient space of a manifold doesn't have to be a manifold. For example in $\mathbb{R}^2$, if $(x, y) \cong (p, q)$ whenever $x^2 + y^2 = p^2 + q^2$, then the equivalence classes are circles along with the origin, and the quotient space is $[0, \infty)$, which is not a manifold. One needs extra assumptions to ensure that it is. We will discuss this again in Theorem 9.1.7. ☺

Definition 8.2.3. Let T be any standard triangle in $\mathbb{R}^2$; to be specific we can use the one with vertices $(0,0)$, $(0,1)$, and $(1,0)$. The "edges" are the closed subsets each of which can be identified with the closed interval $[0,1]$, while the "face" is the entire closed set.

A *triangulation* of a two-dimensional manifold M is a finite or countable collection of closed subsets T_i of M, such that there is a homeomorphism $\phi_i \colon T \to T_i$. Via this homeomorphism we can identify the three edges E_{i1}, E_{i2}, and E_{i3}, as well as the three vertices V_{i1}, V_{i2}, and V_{i3}. We require it to satisfy the following properties:

- The triangles cover M, i.e.., $\cup_i T_i = M$.
- For any i and j, the intersection $T_i \cap T_j$ is either exactly one vertex, exactly one entire edge, or empty.
- An edge of any one triangle is an edge of exactly one other triangle.

The triangulations of the torus and sphere shown above satisfy all three criteria. The second condition is for convenience, while the third condition is clearly necessary to make the space a manifold.

Note that we only ask that T_i be homeomorphic to the triangle, so it doesn't have to look much like one. For example you could pick any three distinct points on the boundary of a disc and call that a triangle, as in Figure 8.6. In particular we don't care if the internal angles add up to π, and thus we don't care for example if three triangles in the manifold meet at a single vertex and the sum of angles at the vertex is more than 2π. This might affect smoothness but not anything topological. For example, cutting each of the sides of a cube in half diagonally gives a decent triangulation of the sphere.

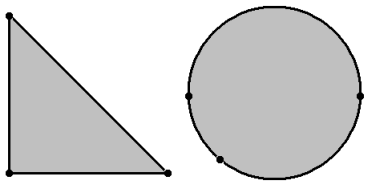

FIGURE 8.6. Either one is just as good a triangle.

It turns out every two-dimensional topological manifold has a triangulation; we will sketch the proof below. The first rigorous proof of this was in 1925 by Tibor Radó[14] It shouldn't be too hard to believe this intuitively, although a rigorous proof is somewhat involved. All proofs I've seen depend on the Jordan-Schoenflies theorem, which states that a simple closed curve separates a plane into two disjoint regions, and there is a homeomorphism from the plane to itself which maps that closed curve onto a circle. (This fact itself has a rather involved proof.) The corresponding result in three dimensions using tetrahedra is even harder[15] The

[14]See for example Ahlfors and Sario, *Riemann surfaces*, for the original proof in English, or Doyle and Moran, "A short proof that compact 2-manifolds can be triangulated," Inventiones mathematicae, **5** pp. 160–162 (1968).

[15]E.E. Moise, *Affine structures in* 3*-manifolds. V. The triangulation theorem and Hauptvermutung*, Annals of Mathematics (2), **56** pp. 96–114.

result in four dimensions is false due to the E_8 manifold example of Freedman mentioned above. The result in higher dimensions is unknown, although for *smooth* manifolds it *is* known to be true in any dimension.[16]

Theorem 8.2.4. *Any two-dimensional topological manifold has a triangulation. If the manifold is compact, then it has a finite triangulation.*

Proof (Sketch). The idea of the proof is to set up a covering by coordinate charts, each of which intersects only finitely many other coordinate charts. (For a compact manifold this is obvious, since we only need finitely many coordinate charts. Generally we need second-countability and local compactness.) One then establishes that the boundary of each chart is a simple closed curve, and the Jordan-Schoenflies theorem shows that the curves divide the manifold into regions homeomorphic to a closed disc, overlapping other such sets only at the boundary. Then, again using finiteness, we subdivide each of these disclike regions into triangles by introducing new vertices in the interior. □

Our first task is to classify compact surfaces, which is clearly easier since there are only finitely many triangles to worry about.

The most important step is to reduce a compact surface to a diagram as for the sphere or torus: a polygon with sides identified in pairs. Examples include those shown in Figure 8.7, which we will discuss in detail later.

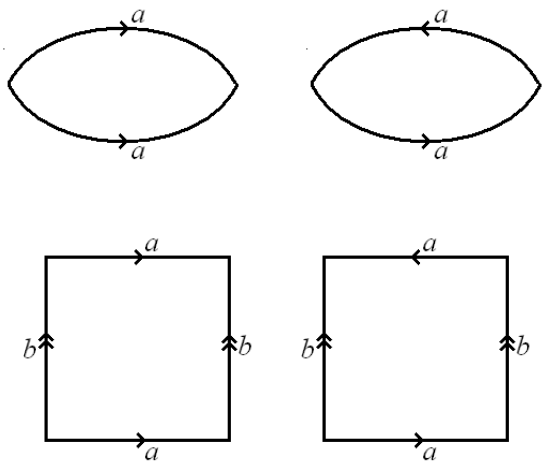

FIGURE 8.7. Several surfaces expressed as the quotient space of a polygon. Following the perimeter counterclockwise gives the "word" which completely describes the manifold structure. The sphere, expressed as aa^{-1}; the projective plane, expressed as aa; the torus, expressed as $aba^{-1}b^{-1}$; and the Klein bottle, expressed as $abab^{-1}$.

[16]S.S. Cairns, *On the triangulation of regular loci*, Annals of Mathematics (2), **35** pp. 579–587 (1934).

Theorem 8.2.5. *Every connected compact two-dimensional manifold M is homeomorphic to the quotient of a polygon in the plane, with an even number of sides, and the sides identified pairwise.*

Proof. From Theorem 8.2.4, we know that M is homeomorphic to a space consisting of finitely many triangles. Pick one, and call it T_1. We know T_1 intersects some other triangle in an edge, for otherwise the space wouldn't be a manifold. So pick one that intersects T_1 in an edge and call it T_2. Either we are done, or there are more triangles in the list. If there are more, then at least one of them must intersect $T_1 \cup T_2$ in an edge, because if the intersection were empty then the manifold would not be connected, and if the intersection were just one point then it would not be a manifold. (It would have to look like the union of two cones joined at a point.) We keep going like this, listing the triangles as $T_1, \ldots, T_N$ where each T_i shares an edge with at least one triangle that's listed earlier. Write the specific edges used as E_2 (the edge that T_1 and T_2 share), E_3 (the edge that T_3 shares with either T_1 or T_2), etc., up to E_N.

Now we arrange the triangles in the plane. Technically what we're going to do is consider the disjoint union of planar triangles with edges identified as our equivalence relation (that is, if a point is interior to a triangle, it is equivalent only to itself; if it is on an edge of a triangle, it is equivalent to exactly one other point on some other edge). Then the manifold will be homeomorphic to the quotient space of this union of triangles.

To actually get a workable model, though, we're going to explicitly eliminate the edges listed as E_i by actually gluing the triangles together there (and then forgetting about them); in other words, replace the first triangle with a quadrilateral, add another triangle to get a pentagon, etc. All the other edges are going to end up on the outside of the polygon, but there will still be some guide by which they'll be identified.

The things to check precisely are that

(1) the union of two polygons homeomorphic to a disc, with one side on each identified in the quotient topology, is still homeomorphic to a disc;
(2) that quotient spaces can be done sequentially; that is, if side E and side F are to be identified, we can first identify side E, then identify side F, and the result will be topologically the same as if we did them both simultaneously.

The reason the second property matters is that we will only glue a triangle onto the polygon in a way that keeps the diagram planar. Clearly if we already had a planar polygon and we glued just one new triangle along some edge, the graph would still be planar; hence we glue edges if they keep the graph planar, and otherwise just add them to the outside of the polygon.

The fact that an even number of edges are left over on the perimeter is a consequence of the fact that there have to be an even number of edges to start with, since every edge must belong to exactly two triangles. We eliminate an even number of edges through identification, and so we have an even number left over. □

So at this point we have a polygon with some edges that we will call a_1 through a_n, each one repeated once. (So $2n$ edges in all.) By going counterclockwise and identifying the edges in order, we can uniquely specify the polygon. Each edge has a direction, so if we encounter a_j facing counterclockwise we list it as a_j, and if it's facing clockwise we write it as a_j^{-1}. Then the entire path will have a "word"

like $aba^{-1}ccb^{-1}$, which specifies it up to rotation. The motivation for writing it this way is because this is how it would show up in the fundamental group. See Theorem 8.2.12 a bit later.

Now as you can imagine, there is more than one way to triangulate a surface, and it's quite possible we could have the same surface with two different polygons and two different words. (There's no reason to have the same number of sides, for example, and even if they did it's easy to rearrange the figure so that the words are different.) So how can we get a classification? Well there's a nice, easy-to-visualize procedure for reducing these things to a standard form.

Lemma 8.2.6. *(First reduction) If the word obtained by following the polygon boundary contains a term of the form aa^{-1}, and these are not the only edges, then we can eliminate this expression.*

Proof. If the entire word is aa^{-1}, then we have a sphere. Otherwise, we just fold that part over and zip it up, and we end up with two fewer edges. See Figure 8.8. □

Notice this does *not* work if you see aa: this represents a projective plane which cannot be eliminated.

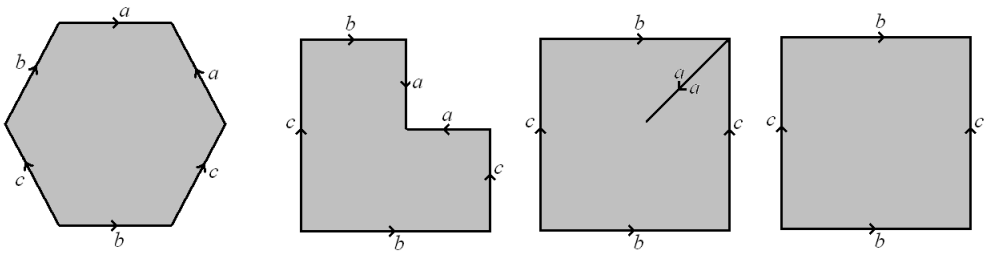

FIGURE 8.8. Lemma 8.2.6 example: eliminating a word that contains aa^{-1}. Fold the polygon over between those edges and glue them together. Then you can ignore them completely.

Lemma 8.2.7. *(Second reduction) Any word formed by the boundary of a polygon can be set up so that there is only one vertex (i.e., every edge is actually a circle).*

Proof. Again I'll illustrate with a simple explicit example. In Figure 8.9, we start with a polygon with word $abcabc$. It has three different vertices: P which is always the start of a and the end of c, Q which is the start of b and the end of a, and R which is the start of c and the end of a. Q appears twice here, and let's say I want to get rid of it. I look at two edges coming out of Q, and I draw a triangle taking a shortcut past Q. This gives a new side, here called d, which I can then cut along. I take the newly created triangle and detach it, then flip/rotate/slide it across so I can join it onto another side. I join it so that the Q points match up, and in so doing I've now reduced the number of Q from two to one. (I've also added one more R.) This works generally: no matter how many times a point Q appears, I can cut a triangle along a diagonal such that Q is the opposite end, then glue that triangle up against another Q appearing somewhere else, and in so doing I have

reduced the number of Q points by one (and increased some other type of point by one). Eventually I get down to one Q remaining, as in this case, and when that happens we *must* get to something as shown: Q has two edges facing out of it going in the same direction, which we can then zip up as in Lemma 8.2.6 to eliminate Q entirely from the perimeter. So instead of three different vertices we end up with two. Then we can go after one of those using the same construction and eliminate it, until all vertices are identified with the same point. □

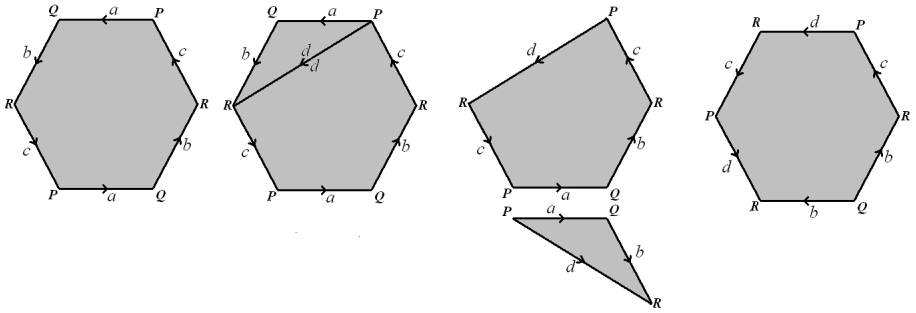

FIGURE 8.9. Lemma 8.2.7 example: eliminating one vertex in favor of another. Cut out the offending vertex with a triangle, slide it over to match it with another copy of it, then reattach it there.

Lemma 8.2.8. *(Third reduction) Suppose an edge shows up twice in the word of a polygon boundary in the same direction both times, such as $\cdots a \cdots a \cdots$, then we can replace the pair by another pair of edges in the same direction and also consecutive, i.e., $\cdots ee \cdots$.*

Proof. The idea is shown in Figure 8.10. We draw a line from the end of one copy of a to the same end of the other copy of a. The shortcut arrow e goes in the same direction as a. We now cut along e and end up with two polygons (there's no reason either has to be a triangle), which we now paste along a. Since both a arrows were going in the same direction, we need to flip over one of the polygons in order to glue it to the other. Now in what remains of the original polygon there is a sequence $a \to e$, and in the new cut polygon we have a and e both having the same tail and facing the same direction. So when we glue along a, we get a sequence ee with both arrows in the same direction. □

Lemma 8.2.9. *(Fourth reduction) Suppose we have two edges a going in opposite directions and two edges b going in opposite directions, such as $\cdots a \cdots b \cdots a^{-1} \cdots b^{-1} \cdots$. Then we can assume they appear consecutively as $aba^{-1}b^{-1}$.*

Proof. See Figure 8.11. We draw a line from one end of a to the other and cut there, calling the new edge e. We then slide this piece across the diagram so that we can glue it along b. Notice that since the b edges were in opposite directions, we don't have to flip this piece. Now we end up with a sequence aea^{-1}. Furthermore, since we didn't have to flip the second polygon, the two e edges are now also going in opposite directions. So we can perform the same trick, drawing a line from one endpoint of e to the other and cutting along the new edge f. Gluing along a, it is

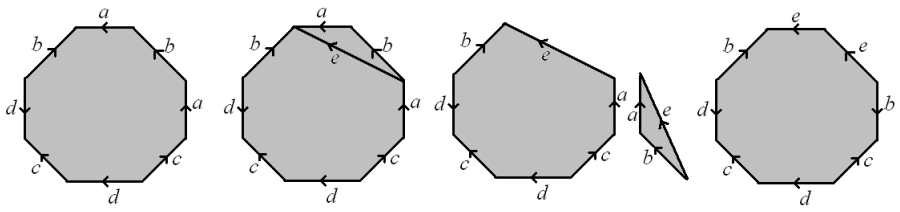

FIGURE 8.10. Lemma 8.2.8 example: replacing two arrows going in the same direction with two *consecutive* arrows in the same direction. Cut from the tail of one to the tail of the other, flip over, slide and reattach along the arrows you want to replace.

easy to see that we end up with the two f edges now interspersed between the e edges and in opposite directions. □

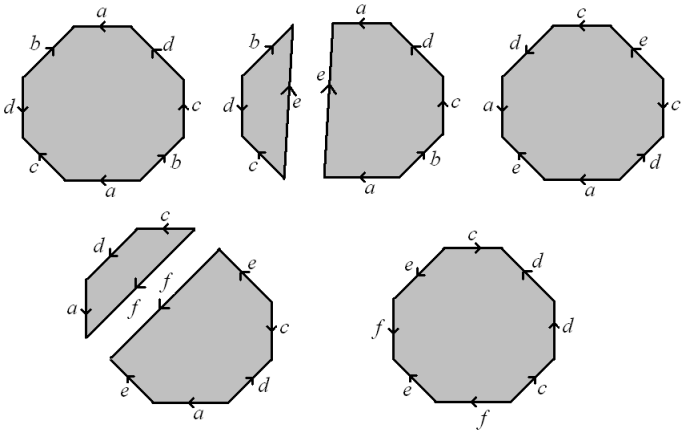

FIGURE 8.11. Lemma 8.2.9 example: if we have two pairs of edges a and b, both of which are in opposite directions and not adjacent, we can eliminate them and get new edges e and f such that they appear in the pattern $efe^{-1}f^{-1}$ in the word. We just do the same operation twice: cut a line from the tail of one to the tail of the other, then glue along the original edge.

Putting this all together, we see that we can separate all pairs into those going the same way (in which case they become consecutive and look like cc) or pairs going the opposite way (in which case they either cancel out, or end up in the form $aba^{-1}b^{-1}$).

Now let's see what we've got. Notice that cc (by itself) is the word corresponding to the projective plane shown in Figure 8.7 (upper right corner), while $aba^{-1}b^{-1}$ is the word corresponding to the torus (lower left corner). Looking at the diagram in Figure 8.12, we see that we can separate all a, b sides from all c sides by cutting a new d side. d is topologically a circle (remember, all vertices are going to be identified to the same point by Lemma 8.2.7, so that any segment from one vertex to another is actually a circle). Hence we can think of $ccaba^{-1}b^{-1}$ as coming from two disjoint manifolds which are connected by cutting a disc out of each and then gluing the manifolds along that disc. As long as we only do one at a time, this will still be a manifold (in a neighborhood of the boundary circle, we have half of $\mathbb{R}^n$ coming from one manifold and the other half coming from the other manifold). In this way we can break up any complicated two-dimensional manifold into "prime factors."

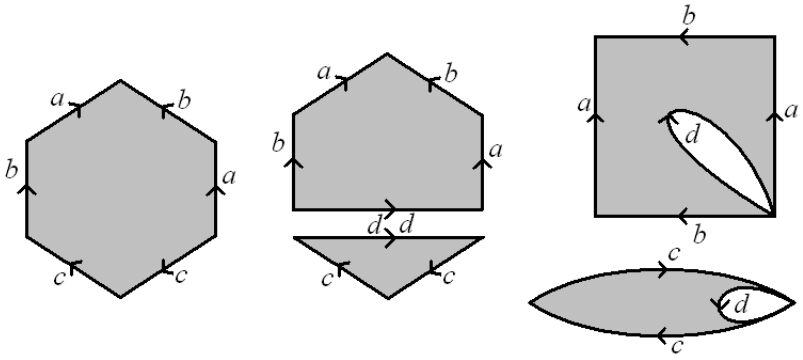

FIGURE 8.12. Decomposing a manifold with word $ccaba^{-1}b^{-1}$ into the connected sum of the torus $\mathbb{T}^2$ and the projective plane $\mathbb{P}^2$.

Again it's time for a general definition. Connected sums are very intuitive and easy to visualize, and they are also one of the most useful ways we have of finding interesting manifolds.

Definition 8.2.10. Suppose M and N are two 2-dimensional topological manifolds. The *connected sum* S, denoted by $M\#N$, is a new topological space constructed by the following procedure. Take any subsets $D \subset M$ and $E \subset N$ which are homeomorphic to the closed unit ball

$$B_1(0) = \{x \in \mathbb{R}^2 \mid \|x\| \leq 1\}.$$

Remove the interiors of D and E to get just the boundaries ∂D and ∂E. Let $\gamma\colon S^1 \to \partial D$ and $\delta\colon S^1 \to \partial E$ be homeomorphisms. Put an equivalence relation on the disjoint union $(M \setminus \operatorname{int} D) \cup (N \setminus \operatorname{int} E)$ by saying $\gamma(\theta) \cong \delta(\theta)$ for every $\theta \in S^1$. Then $S = M\#N$ is the quotient space, and it is a topological manifold. See Figure 8.13.

We can also define higher-dimensional connected sums; for example we might take two copies of $\mathbb{R}^3$ and cut out a solid torus from each, then glue the spaces together along the boundary (a 2-torus). Clearly this operation can get much more

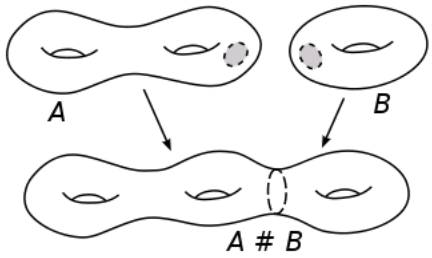

FIGURE 8.13. An illustration stolen form the "Connected Sum" article on Wikipedia. Here a 2-holed torus has been combined with a standard 1-holed torus to give a 3-holed torus.

complicated than in the two-dimensional case, but it's still considered the basis for any possible classification of three-dimensional manifolds.

One can prove (using the polygonal representations) that the connected sum operation is commmutative, i.e., $M\#N = N\#M$ homeomorphically, and that it is associative, i.e., $(M\#N)\#P = M\#(N\#P)$ homeomorphically. Hence it really is appropriate to think of it as a kind of multiplication. It's not hard to see that the sphere S^2 serves as the identity in this multiplication, since cutting a disc out of the sphere gives another disc left over, and gluing that in to replace a disc in M just gives M back again.

The Lemmas above, along with the triangulation result, have now almost proved the following famous classification.

Theorem 8.2.11. *Every two-dimensional compact topological manifold is either S^2, or a connected sum of n copies of the torus $\mathbb{T}^2$ for some positive integer n, or the connected sum of n copies of the projective plane $\mathbb{P}^2$ for some positive integer n.*

To see this, just write the word as some number of terms of the form $a_i b_i a_i^{-1} b_i^{-1}$ and some number of terms of the form $c_i c_i$, and separate all the terms by cutting. The manifold is then a connected sum of the pieces we have separated.

The only issue is what happens if we have $\mathbb{T}^2\# \cdots \#\mathbb{T}^2\#\mathbb{P}^2\# \cdots \#\mathbb{P}^2$. Actually we can reduce all the terms to projective planes, just using the following formula: $\mathbb{T}^2\#\mathbb{P}^2 = \mathbb{P}^2\#\mathbb{P}^2\#\mathbb{P}^2$. (You'll prove this in the homework.)

To prove these are all actually different, one uses the fundamental group. This computation can basically be done in a standard topology course using the Seifert-van Kampen theorem, so we will not prove it here.

Theorem 8.2.12. *Suppose M is an compact two-dimensional manifold that comes from a polygon with word reduced to the form $a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_n b_n a_n^{-1} b_n^{-1}$. Then the fundamental group of M is the group generated by all possible combinations of the elements a_k and b_k modulo the single relation*

$$a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_n b_n a_n^{-1} b_n^{-1} = 1.$$

Alternately if M is a compact two-dimensional manifold that comes from a polygon with word reduced to the form $a_1 a_1 \cdots a_n a_n$, then the fundamental group of M is the group generated by all possible combinations of a_k modulo the single relation

$$a_1 a_1 \cdots a_n a_n = 1.$$

For example in case $n = 1$ with word $aba^{-1}b^{-1}$, we have $M \cong \mathbb{T}^2$, and the fundamental group is generated by a and b with relation $aba^{-1}b^{-1} = 1$, which translates to $ab = ba$. So it's commutative and must be $\mathbb{Z}^2$. In case $n = 1$ with word aa, the fundamental group is generated by a with relation $a^2 = 1$, so it's $\mathbb{Z}_2$.

The fundamental groups are pretty nasty in most other cases, but an easy way to distinguish them is to abelianize; that is, demand that all the elements commute with each other in addition to satisfying the relevant relation in Theorem 8.2.12. (This is equivalent to considering the first homology group rather than the first homotopy group.) Then if M is a connected sum of n tori, the abelianized fundamental group is $\mathbb{Z}^{2n}$ (the single relation is automatically true). And if M is a connected sum of n projective planes, the abelianized fundamental group is $\mathbb{Z}^{n-1} \times \mathbb{Z}_2$, since the one relation is equivalent in the abelianized case to $(a_1 \cdots a_n)^2 = 1$.

Two concepts are used to distinguish compact surfaces: genus and orientability.

Definition 8.2.13. If M is a two-dimensional compact manifold which is the connected sum of n tori or the connected sum of n projective planes, then n is called the *genus* of M. The sphere S^2 has genus zero.

The genus is the number of disjoint simple closed curves (i.e., homeomorphic to S^1) that can be removed from the surface without making it disconnected. Thus it is a topological invariant, and surfaces with different genus are not homeomorphic.

Definition 8.2.14. If there is a family (ϕ_i, U_i) of C^∞ coordinate charts covering a smooth manifold M such that whenever $U_i \cap U_j \neq \emptyset$, the map $\phi_i \circ \phi_j^{-1} \colon \phi_j[U_i \cap U_j] \subset \mathbb{R}^n \to \phi_i[U_i \cap U_j] \subset \mathbb{R}^n$ has $\det(D(\phi_i \circ \phi_j)^{-1}) > 0$ everywhere, then M is called *orientable*.

Note that since $\phi_i \circ \phi_j^{-1}$ is C^∞ and its inverse is as well, the derivative must be nonsingular everywhere, and hence the determinant is either always positive or always negative on any connected subset. There is a purely topological definition of orientability, but it's a bit more complicated, and we won't need it.

The simplest orientable manifold is the sphere.

Example 8.2.15. The 2-sphere S^2 is orientable. For example we can use the explicit coordinate charts given by hemispherical charts in Example 7.2.6. Even more simply we can use the two-dimensional analogue of the stereographic coordinates from Example 7.1.7; then there are only two coordinate charts, and you can check that the transition map is $(p, q) = \left(\frac{u}{u^2+v^2}, \frac{v}{u^2+v^2}\right)$. Unfortunately we have $p_u q_v - p_v q_u = -1/(u^2 + v^2)^2$, which is negative! However we could just switch the coordinates to get $(\tilde{p}, \tilde{q}) = \left(\frac{v}{u^2+v^2}, \frac{u}{u^2+v^2}\right)$, which yields $p_u q_v - p_v q_u = 1/(u^2 + v^2)^2 > 0$. ☺

We can also check that the torus $\mathbb{T}^2$ is orientable.

Example 8.2.16. The torus $\mathbb{T}^2$ is orientable. The basic idea is to use the exact same method as in Example 7.2.3 to parametrize the torus, such as by a map

$$F(u, v) = \big((2 + \cos u) \cos v, (2 + \cos u) \sin v, \sin u\big).$$

The image of this map is shown in Figure 8.16. The parametrization is defined on $\mathbb{R}^2$, but restricting to certain open sets, we obtain bijections that yield coordinate charts. The transition maps all end up being formulas of the form $(u, v) \mapsto (u +$

$2m\pi, v + 2n\pi)$ where m and n are either 0, 1, or -1, and these transition maps all have positive Jacobian determinant. ☺

It can be shown that the connected sum of two orientable manifolds is also orientable, and thus Example 8.2.16 shows that an n-holed torus is always orientable.

Example 8.2.17. The Möbius band defined in Example 7.2.3 is not orientable. We already showed there how to cover it with two coordinate charts which intersect in two disjoint open sets. The transition map is given by either $(u, v) \mapsto (u, v)$ or $(u, v) \mapsto (u - 2\pi, -v)$; the first of these has positive Jacobian determinant, while the second one has negative Jacobian determinant. No trick like switching variables in Example 8.2.15 will change this, since it will correct the sign on one component and break it on the other component.

Furthermore even if we had some totally different family of coordinate charts, they could not possibly satisfy the positive-Jacobian compatibility condition: they would have to be compatible with the charts we already have, and the charts we have are not compatible with each other. ☺

Finally one can prove that an open subset of an orientable manifold is also orientable. This fact shows that any manifold which contains a subset homeomorphic to the Möbius strip is not orientable. Examples include the projective plane $\mathbb{P}^2$ (see Figure 8.14) and the connected sum $\mathbb{P}^2 \# \mathbb{T}^2 \# \cdots \# \mathbb{T}^2$.

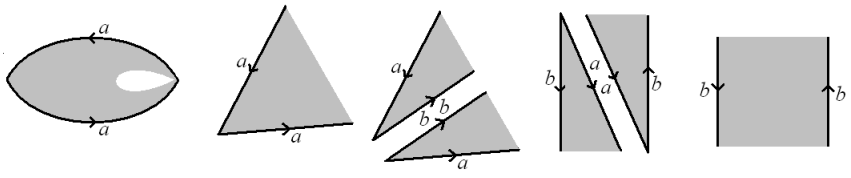

FIGURE 8.14. Deleting a closed disc from a projective plane yields a Möbius strip. Hence the projective plane cannot be orientable.

Thus a two-dimensional manifold is completely determined by its genus and whether it is orientable.

Having classified the two-dimensional manifolds using polygons, we now analyze these manifolds in more detail.

- The standard representation of the 2-sphere S^2 is $\{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}$, as in Example 7.2.6. Standard spherical coordinates (θ, ϕ) cover all but a closed half of a great circle: the formulas

$$(x, y, z) = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$$

  are smooth and invertible on the open set $(0, \pi) \times (0, 2\pi)$, and cover the sphere except for the points $(\sqrt{1 - z^2}, 0, z)$ for $-1 \leq z \leq 1$. See Figure 8.15 for an illustration.

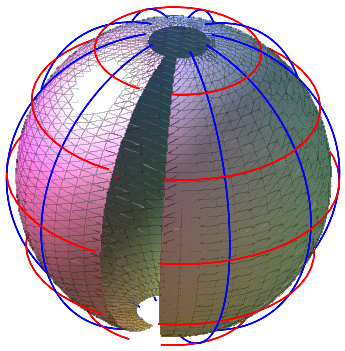

FIGURE 8.15. The image of $[\varepsilon, \pi-\varepsilon]\times[\varepsilon, 2\pi-\varepsilon]$ under the standard (physicist's) spherical coordinate chart. As $\varepsilon \to 0$ we approach the sphere but are always missing a closed half of a great circle.

Another common coordinate chart is stereographic projection, given by $(x, y, z) = (\frac{2u}{u^2+v^2+1}, \frac{2v}{u^2+v^2+1}, \frac{u^2+v^2-1}{u^2+v^2+1})$ for $(u, v) \in \mathbb{R}^2$, which is just a generalization of the circle charts in Example 7.1.7. Because the stereographic charts cover the entire sphere minus a single point, the sphere is often thought of as the completion of the plane by a single point at infinity.[17] So for example, if it's convenient we might think of a function on the plane, which has the same limit at infinity in all directions, as a function on the sphere instead. This is very common in complex analysis.

- The most convenient representation of the torus is as the quotient space of $\mathbb{R}^2$ under the equivalence relation $(x, y) \cong (x+2m\pi, y+2n\pi)$ for integers m and n, which is equivalent to the planar diagram. It's frequently visualized as the image of $(x, y, z) = ((a + b\cos u)\cos v, (a + b\cos u)\sin v, b\sin u)$, for constants $a > b > 0$, as in Figure 8.16. A more convenient embedding is into $\mathbb{R}^4$ given by the formula $(w, x, y, z) = (\cos u, \sin u, \cos v, \sin v)$, as discussed in Example 7.2.7.
- The projective plane, denoted by $\mathbb{P}^2$, is the quotient of the 2-sphere under the equivalence relation $(x, y, z) \cong (-x, -y, -z)$. Historically it was important as the space of lines through the origin in $\mathbb{R}^3$. Each line through the origin is determined by a single point on it, and so we might as well find the point of intersection with the unit sphere. However there are always two such intersections (and they are negatives of each other), and since we want both points to represent the same object, we might as well identify them. A simpler but equivalent way to think about it is as one hemisphere of the sphere with the boundary sealed up via identifying antipodal points there.

  We cannot embed the projective plane (or any non-orientable surface) in $\mathbb{R}^3$. If we could, then it would be the boundary of a compact region which is orientable (as a subset of $\mathbb{R}^3$); the outer unit normal vector field could be

[17] If you know about the "one-point compactification," this is just saying that S^2 is the one-point compactification of $\mathbb{R}^2$. Typically one should not expect a one-point compactification of a manifold to also be a manifold, but here it works.

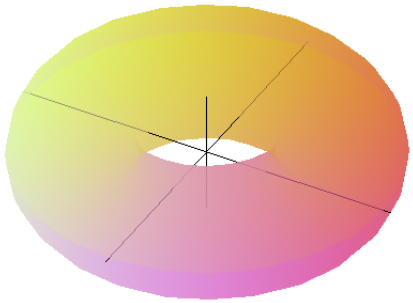

FIGURE 8.16. The standard embedding of the torus in $\mathbb{R}^3$.

used to provide an orientation, which contradicts the fact that $\mathbb{P}^2$ is non-orientable since it contains the Möbius strip as in Figure 8.14. If we try to embed in three dimensions, it will be forced to intersect itself, for exactly the same reason that a closed curve that forms a knot must intersect itself when projected onto the plane, but need not intersect itself in $\mathbb{R}^3$. As long as we don't take the self-intersections seriously, we can use these to try to visualize the projective plane. A good embedding of the projective plane is given by the image of the sphere under the map $F(x, y, z) = (xz, yz, xy, \frac{1}{2}(y^2 - z^2))$; it satisfies $F(x, y, z) = F(-x, -y, -z)$, and these are the only self-intersections, and therefore it is actually a homeomorphism of the projective plane. (See Theorem 9.1.7.) If we ignore the fourth coordinate and project onto what's left, we get Steiner's Roman surface, and if we ignore the third coordinate, we get the cross-cap. See Figure 8.17.

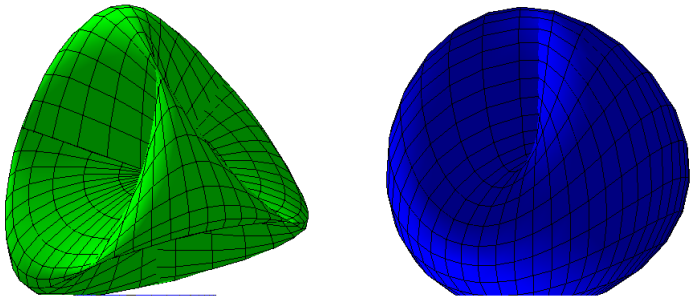

FIGURE 8.17. Two common representations of the projective plane, as self-intersecting surfaces in $\mathbb{R}^3$. On the left is the Roman surface, and on the right is the cross-cap.

- The Klein bottle is the connected sum of $\mathbb{P}^2$ and $\mathbb{P}^2$, so it has word $aabb$, which can be simplified to the square with word $aba^{-1}b$. It is constructed like the torus: we wrap up the square in the usual way to get a cylinder,

except instead of wrapping the ends of the cylinder around to get a torus, we turn one end inside out to glue it to the other circle the opposite way. This forces the bottle to cross itself if we try to embed it in $\mathbb{R}^3$, just like for the projective plane. See Figure 8.18.

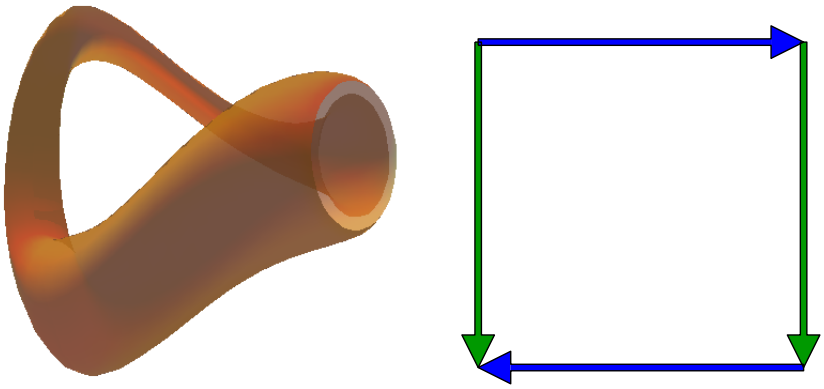

FIGURE 8.18. The usual picture of the Klein bottle as a self-intersecting surface in $\mathbb{R}^3$ on the left (although the surface does not intersect itself in higher dimensions). On the right, the Klein bottle as a quotient of a rectangle under a twist (compare to Figure 8.5 which is almost the same).

As with the torus, the best way to think about the Klein bottle is as the quotient of $\mathbb{R}^2$ under the equivalence relation $(x, y) \cong ((-1)^n x + 2m\pi, y + 2n\pi)$. We can embed the Klein bottle into $\mathbb{R}^4$, although the parametrization is not quite so simple. The embedding

$$(u, v) \mapsto (\cos u, \cos 2v, \sin 2v, \sin u \cos v, \sin u \sin v)$$

into $\mathbb{R}^5$ works and is relatively simple.

- Higher-genus surfaces are just connected sums of these. It is easy to visualize the connected sum of tori using the explicit embedding in $\mathbb{R}^3$, as in Figure 8.13.

If you know about Gaussian curvature for surfaces, you can think of genus-zero surfaces like the projective plane and sphere as having constant curvature $+1$, and you can think of genus-one surfaces like the torus and Klein bottle as having constant curvature 0. It's less obvious that higher-dimensional surfaces can be made to have constant curvature -1, but it's true. In fact the easiest way to see that the torus and Klein bottle have curvature zero is to tile the plane with copies of them: curvature is a local property, so locally you can't tell a torus or Klein bottle from the Euclidean plane.

Unfortunately you can't tile the Euclidean plane with regular versions of the higher-dimensional polygons corresponding to other surfaces. (You can tile it with hexagons, which are used to build $\mathbb{P}^2\#\mathbb{T}^2$, but you can't get the identifications to match up with the usual tiling.) The problem is that the angles at each vertex have to add up to 2π radians in order for it to smoothly correspond to Euclidean space (otherwise you get something that looks like a cone, which is fine topologically but

not so good geometrically). The interior angles of a regular octagon for example are $\frac{3\pi}{4}$ radians, and so we cannot put together an integer number of regular octagons in the flat plane all meeting at the same point.

However on the hyperbolic plane the angles of a triangle don't need to add up to π. In fact the sum of the sides can be arbitrarily small just by taking a large enough triangle. The same rules apply for other polygons. In particular we can construct a large octagon with interior angles adding up to 2π, use isometries of the hyperbolic plane to get the others, then match eight of them up at each point to get a nice regular tiling. Figure 8.19 shows what a tiling of the Poincaré disc by regular octagons looks like. Note that all the octagons are actually isometric: some of them look small because they're approaching the disc boundary, but that boundary is actually infinite distance from everything else.

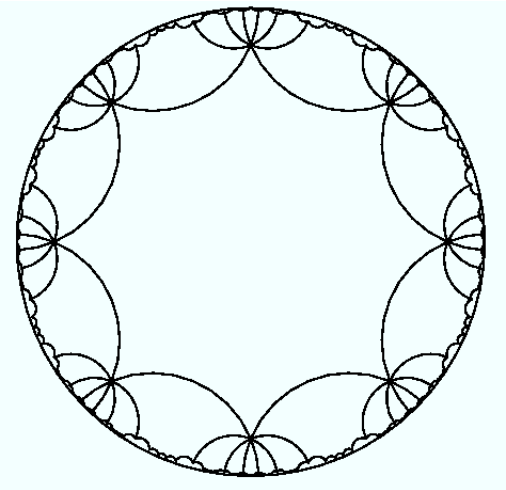

FIGURE 8.19. A hyperbolic plane (modeled as the Poincaré disc) can be tiled by regular octagons, and also by any other polygon diagram representing a surface of genus two or more.

Poincaré proved that any polygon diagram of a compact surface with equal lengths can tile the hyperbolic plane, so we can think of any surface of genus two or higher as a quotient of the hyperbolic plane, which thus has constant curvature -1. These surfaces cannot be embedded in Euclidean 3-space (no compact negative-curvature surface can), but they can be embedded in 4-space.

This theorem is called the *uniformization theorem*: every two-dimensional compact surface can have a constant-curvature geometry put on it.[18] The corresponding result for three dimensions is the substance of Thurston's conjecture, and Perelman's proof of Thurston's conjecture showed that all three-dimensional compact manifolds could be geometrized. This has not quite gotten us a full classification of three-dimensional manifolds, but most people think we are close.

The only other thing to say about two-dimensional manifolds is the case where they are noncompact. The classification here is harder, mainly because any open subset of a manifold is a manifold. So we can remove finitely many or countably

[18]Details will come in an addendum.

many discs from any compact surface and get a noncompact surface. Essentially the idea of the classification[19] is to compactify the surface by taking a connected sum with finitely many or countably many discs. To do this, we first have to figure out how many discs are needed: the set of all such discs is called the "ideal boundary," with one point in the ideal boundary for each component of the surface minus a compact subset, and it may be quite complicated.

[19]See B. Kerékjártó, Vorlesungen über Topologie. Vol. I, Springer, Berlin, 1923, or I. Richards, *On the classification of noncompact surfaces*, Trans. Amer. Math. Soc. **106** (1963), 259–269. MR 26 #746.

9. Higher-dimensional examples of manifolds

"I think my eyes are getting better. Instead of a big dark blur, I see a big light blur."

In the last Chapter, we classified all one-dimensional manifolds (there are two cases: compact and noncompact) and all two-dimensional manifolds (determined by the genus and whether they are orientable or not). There is no similar classification known in three dimensions, although the solution of Thurston's geometrization conjecture by Perelman leaves us pretty close to a complete classification (but not quite so explicit as "compute this list of algebraic quantities for the two manifolds, and all of them are the same if and only if the two manifolds are homeomorphic," which is where we stand in dimension one or two). The geometrization conjecture basically says that a simply-connected 3-dimensional manifold can be decomposed into a connected sum of simpler 3-dimensional pieces, each of which has exactly one of eight "model" geometries. Here the connected sum is taken using either 2-tori or 2-spheres, and the eight model geometries are the generalization of the three model geometries in two dimensions (that is, the sphere S^2 of constant positive curvature, the plane $\mathbb{R}^2$ of zero curvature, and the hyperbolic plane $\mathbb{H}^2$ of constant negative curvature). Once we discuss Riemannian metrics, it will be easier to describe the model spaces, but for now you can consider the standard ones to be S^3, $\mathbb{R}^3$, $\mathbb{H}^3$, $S^2 \times \mathbb{R}$, $\mathbb{H}^2 \times \mathbb{R}$, along with three that are special types of Lie groups with left-invariant metrics (which we are not ready to discuss yet; we will only begin discussing them in our final Chapter 20).

In four dimensions, no classification of compact manifolds is even possible. This comes from the fact that every manifold has a group, called the *fundamental group*, associated with it. If two manifolds are homeomorphic, then their fundamental groups must be isomorphic, and it's easier to determine the latter than the former. In two dimensions the fundamental group of a compact manifold coming from a polygon with $2n$ sides is a group with n generators satisfying one relation coming from the word around the perimeter. The easiest way to understand this is with some examples. On the torus $\mathbb{T}^2$ the word is $aba^{-1}b^{-1}$, and so the group is generated by taking powers of a and b, inverses of a and b, and multiplying as many of these together as desired. The only simplification we get comes from the fact that the word is 1: $aba^{-1}b^{-1} = 1$, which (multiplying by b and then a on the right) means that $ab = ba$. So the group is actually commutative, which forces it to be $\mathbb{Z}^2$. For a projective plane the group is generated by one element a with one relation $a^2 = 1$, which means it consists of only two elements: $\{1, a\}$, so it's $\mathbb{Z}^2$.

There aren't that many groups one can get as the fundamental group of a surface though, since a general group on finitely many generators can have all sorts of relations while the fundamental group of a surface can only have one. We get more possibilities in three dimensions, and in four dimensions we can get *any* group with finitely many generators and finitely many relations as the fundamental group of some compact manifold. Unfortunately the algebraists have proved that there cannot be an algorithm for deciding whether two such groups are the same[20], and

[20]This is known as *the word problem*, which is algorithmically undecidable. See for example the Wikipedia article `http://en.wikipedia.org/wiki/Word_problem_for_groups`

the consequence for geometry is that there is no algorithm to distinguish the fundamental groups of 4-manifolds. Thus there can't be an algorithm to distinguish between homeomorphism classes of 4-manifolds, which is exactly what we mean by a classification.

Hence we won't worry too much about trying to capture all possible manifolds, instead just working with the most popular examples. First we discuss the ways of building new manifolds from old ones.

9.1. **New manifolds from old manifolds.** We already have one method, which comes from inverse images of regular values. If $F\colon \mathbb{R}^{n+k} \to \mathbb{R}^k$ is C^∞ and $r \in \mathbb{R}^k$ is a regular value, so that whenever $F(x) = r$ the rank of $DF(x)$ is k, then $F^{-1}(r)$ is a smooth n-dimensional manifold. We can actually generalize this a bit.

Definition 9.1.1. Suppose N is an $(n+k)$-dimensional smooth manifold and K is a k-dimensional smooth manifold. Let $F\colon N \to K$ be a continuous function. We say that F is *smooth* if for any coordinate chart (U, ϕ) on N and coordinate chart (V, ψ) on K, the map $\psi \circ F \circ \phi^{-1}$ is C^∞ on its domain, which is an open subset of $\mathbb{R}^{n+k}$). We say that F *has rank k at a point $p \in N$* if the map $D(\psi \circ F \circ \phi^{-1})$ has rank k at $\phi(p)$.

Note that the only way we could possibly check smoothness and the rank of the derivative is by moving things over to Euclidean space, and there's only one logical way to do this. Note also that if we find that F has rank k in some coordinate charts, then it must have the same rank in all other compatible coordinate charts, since the transition maps all are smooth and have maximal rank.

Theorem 9.1.2. *If $F\colon N \to K$ is smooth as in Definition 9.1.1 and has rank k at every point of $F^{-1}(r)$ for some $r \in K$, then $F^{-1}(r)$ is a smooth manifold of dimension n.*

Proof. The only thing to do is check the existence of coordinate charts. But if that's all we're doing, then we might as well be doing everything locally anyway, and then we can just use the implicit function theorem in any coordinate chart as in Theorem 7.2.5. □

We have already worked this out for the 2-sphere S^2 in $\mathbb{R}^3$ in Example 7.2.6. Higher-dimensional spheres work in exactly the same way.

It is rare that one actually *proves* that a smooth manifold comes from a map from one manifold to another: usually one works with functions on a Euclidean space. However it is sometimes interesting; see Example 9.1.12 below.

Another popular way to build more complicated manifolds is to take the Cartesian product of two simpler manifolds.

Theorem 9.1.3. *If M is an m-dimensional manifold and N is an n-dimensional manifold, then the product space $M \times N$ is an $(m+n)$-dimensional manifold. If M and N are smooth, then so is $M \times N$.*

Proof. The product of Hausdorff spaces is Hausdorff, and the product of second-countable spaces is second-countable. So we just have to check the coordinate charts. If (U, ϕ) is a chart on M and (V, ψ) is a chart on N, then $U \times V$ is an open set in the product topology, and the map $\phi \times \psi$ given by $(\phi \times \psi)(p, q) = (\phi(p), \psi(q))$ is a homeomorphism from $U \times V$ to $\mathbb{R}^m \times \mathbb{R}^n \cong \mathbb{R}^{m+n}$. The transition maps

on the product are obviously C^∞ if each component's transition maps are, since $(\phi_1 \times \psi_1) \circ (\phi_2 \times \psi_2)^{-1} = (\phi_1 \circ \phi_2^{-1}) \times (\psi_1 \circ \psi_2^{-1})$. □

In this way we get manifolds like $S^m \times S^n$, which is typically considered the second-simplest type of compact manifold (after the spheres themselves). We also get tori.

Example 9.1.4. The *n-torus*[21] $\mathbb{T}^n$ is the product of n copies of S^1; or if you prefer the inductive approach, $\mathbb{T}^1 = S^1$ and $\mathbb{T}^n = \mathbb{T}^{n-1} \times S^1$. Here S^1 is either a quotient space of $[0, 2\pi]$ or $[0, 1]$ with endpoints identified, depending on what's more convenient for you. The typical choice is $[0, 1]$ unless you're doing Fourier series.

You can also think of the n-torus as the quotient of $\mathbb{R}^n$ by the lattice $\mathbb{Z}^n$ or $(2\pi\mathbb{Z})^n$, as we did in Example 7.2.7 in two dimensions. In fact this is typically how one ends up dealing with a torus: one wants to talk about a function on $\mathbb{R}^n$ which is periodic, so one descends to the torus instead because the torus is compact (and compact spaces are much more convenient than noncompact spaces for almost every purpose).

The n-torus can also be easily embedded in $\mathbb{R}^{2n}$ using the same sort of technique as in Example 7.2.7, writing it as the image of $(\cos\theta_1, \sin\theta_1, \cdots, \cos\theta_n, \sin\theta_n)$. This is nice to know, but not usually all that useful. ☺

We now generalize the notion of connected sum, which we previously defined for surfaces in Definition 8.2.10. It just involves replacing discs with balls.

Definition 9.1.5. The *connected sum* of two n-dimensional manifolds M and N is a new manifold $M\#N$ defined by removing the interiors of closed subsets of M and N which are homeomorphic to the closed unit ball, then gluing what remains along the boundary spheres S^{n-1} that remain.

As long as we consider only balls and spheres as their boundary, this doesn't get too complicated, but it's a very useful way to get interesting manifolds without much trouble. Just like in two dimensions, this operation is symmetric and associative, and the n-sphere acts like the identity.

The last and perhaps most interesting construction is a quotient space by a discrete group. We have already seen several examples of this: $\mathbb{T}^n$ is the quotient of $\mathbb{R}^n$ by the discrete group $\mathbb{Z}^n$, the projective plane $\mathbb{P}^2$ is the quotient of S^2 by the discrete group $\mathbb{Z}_2$, etc. Let's formalize this procedure and figure out some conditions under which the quotient space of a manifold is guaranteed to be another manifold.

Definition 9.1.6. A *group* G is a set of objects with a "multiplication" operation $(x, y) \mapsto x \otimes y$ satisfying:[22]

(1) Associativity: $x \otimes (y \otimes z) = (x \otimes y) \otimes z$ for all $x, y, z \in G$;
(2) Identity: there is an identity element e such that $e \otimes x = x \otimes e = x$ for all $x \in G$;
(3) Inverse: for each $x \in G$ there is an $x^{-1} \in G$ such that $x \otimes x^{-1} = x^{-1} \otimes x = e$.

[21] Some people use "n-torus" for a two-dimensional compact orientable manifold with genus n. These people are wrong. Don't do this. Almost everyone agrees that an n-torus is an n-dimensional manifold, and you'd confuse them.

[22] Don't confuse this operation with the tensor product, which only makes sense in a totally different context.

If M is a manifold, a *group action* on M is a group G such that for each $g \in G$ there is a homeomorphism $\phi_g \colon M \to M$ such that $\phi_g \circ \phi_h = \phi_{g \otimes h}$.

For example, $\mathbb{Z}$ is a group, where $m \otimes n = m + n$, the identity is $e = 0$, and $m^{-1} = -m$. A corresponding group action on $\mathbb{R}$ is $\phi_m(u) = u + m$. Another group action with the same group is $\phi_m(u) = 2^m u$. (If you're not familiar with groups, you should definitely work out all the details of this example to understand what we're doing.) This generalizes easily to $\mathbb{Z}^n$ and $\mathbb{R}^n$.

$\mathbb{Z}_2 = \{1, -1\}$ is a group under multiplication. More frequently one also uses $\{0, 1\}$ under addition and declares that $1 + 1 = 0$; think of 0 as representing even numbers and 1 as representing odds. This generalizes to $\mathbb{Z}_p$ for any integer p, where one works with addition and just declares $p = 0$. A group action on $\mathbb{R}^n$ is $\phi_1(x) = x$ and $\phi_{-1}(x) = -x$. This gives a group action on S^n as well since it clearly preserves lengths.

A group doesn't have to be discrete: we can think of $\mathbb{R}$ as a group under addition, and a group action on $\mathbb{R}^2$ by rotation $\phi_x(u, v) = (\cos xu - \sin xv, \sin xu + \cos xv)$. Alternatively we can think of the group as $S^1 = \mathbb{R}/(2\pi\mathbb{Z})$, which yields essentially the same group action.

Clearly a group action generates an equivalence relation, via the definition $p \cong q$ if and only if $\phi_g(p) = q$ for some $g \in G$. (Associativity of the group is used to prove transitivity of the equivalence relation, existence of inverses is used to prove symmetry, and existence of an identity is used to prove that $p \cong p$. Work this all out if this stuff is foreign to you!) So we can consider the quotient space, which we denote by M/G.

Because I don't care about the greatest possible generality and want proofs to be easy, I'm going to make a fairly strong assumption about the group actions we will consider.

Definition 9.1.7. A *free and proper discrete group action* is a group action for which

- for any $p \in M$ there is a neighborhood $U \ni p$ such that $U \cap \phi_g[U] = \emptyset$ for all $g \in G$ except the identity.
- for any distinct $p, q \in M$ with $p \neq \phi_g(q)$ for all $g \in G$, there are neighborhoods $U \ni p$ and $V \ni q$ such that $U \cap \phi_g[V] = \emptyset$ for all $g \in G$.

The two conditions are basically saying the same sort of thing: that the group action pushes points far away from themselves and keeps separated points far away from each other.

Theorem 9.1.8. *If M is a manifold and the action of a discrete group G on M is free and proper, then the quotient space M/G is a manifold. If M is smooth and every ϕ_g is smooth, then M/G is smooth.*

Proof. Let $\pi \colon M \to M/G$ be the projection. Let $y \in M/G$ be arbitrary. Choose any $x \in \pi^{-1}(y)$, and let U be an open subset of M containing x such that $\phi_g[U] \cap U = \emptyset$ for all $g \in G$ with g not the identity. Take any coordinate chart (ψ, V) with $x \in V$. Then ψ maps $U \cap V$ into some open subset of $\mathbb{R}$, which we can assume (by restricting V) is an open ball. Then composing with a homeomorphism, we can assume that $\tilde{\psi}$ maps $U \cap V$ onto all of $\mathbb{R}^n$ homeomorphically. We then set $(W, \tilde{\psi})$ to be the new coordinate chart on M.

Since $\phi_g[U] \cap U = \emptyset$ if g is not the identity, we easily see that $\phi_g[W] \cap W = \emptyset$ if g is not the identity. Now $\pi[W]$ is open in M/G by definition of the quotient

topology, and the map $\tilde{\psi} \circ \pi^{-1}$ takes $\pi[W]$ onto $\mathbb{R}^n$ homeomorphically. This gives us coordinate charts. It remains to check the Hausdorff property and second-countability.

To prove M/G is Hausdorff, pick any two points p and q such that $\pi(p) \neq \pi(q)$. Choose open sets U and V such that $p \in U$ and $q \in V$ and $\phi_g[U] \cap V$ is empty for all $g \in G$; by shrinking U and V we may also assume that $\phi_g[U] \cap U = \phi_g[V] \cap V = \emptyset$ for all g not the identity. Let $\tilde{U} = \cup_{g \in G} \phi_g[U]$ and $\tilde{V} = \cup_{g \in G} \phi_g[V]$, and let $A = \pi[U]$ and $B = \pi[V]$; then A and B are open in M/G since $\pi^{-1}[A] = \tilde{U}$ and $\pi^{-1}[B] = \tilde{V}$. Furthermore $\tilde{U}$ and $\tilde{V}$ are disjoint since otherwise we would have $\phi_g[U] \cap \phi_h[V] \neq \emptyset$ for some $g, h \in G$, and then $\phi_{h^{-1} \cdot g}[U] \cap V \neq \emptyset$. We conclude that A and B are disjoint open sets containing $\pi(p)$ and $\pi(q)$.

Second-countability of M/G is easy: given a countable basis $\{\Omega_n\}$ of M, just take $\pi[\Omega_n]$ and check that it's a countable basis of M/G. □

An example of a group action satisfying the first condition in Definition 9.1.7 but not the second is the action of $\mathbb{Z}$ on $\mathbb{R}^2$ given by

$$\phi_n(x, y) = (x + n, 2^{-n} y).$$

A neighborhood $U = (x - \frac{1}{2}, x + \frac{1}{2})$ satisfies $\phi_n[U] \cap U \neq \emptyset$ iff $n = 0$, but we cannot separate $(0, 0)$ from $(0, 1)$ by translation-invariant open sets: any open neighborhood of $(0, 0)$ will eventually contain some $(0, 2^{-n})$. The quotient space under this action is not Hausdorff; it looks like a cylinder where all the copies of the circle are mashed together. Hence the second condition really is necessary to get the quotient to be Hausdorff.

The action of $\mathbb{Z}^n$ on $\mathbb{R}^n$ given by addition is clearly both free and proper: around one point we consider a ball of radius $1/2$, and around two points we observe that the minimum distance from the images of one point to the other point is positive, and take balls of half that distance. Thus $\mathbb{T}^n$ is a manifold (which we already knew, because we wrote it as a product).

Similarly the action of $\mathbb{Z}_2$ on S^n given by reflection is free and proper. The fact that it's free comes from the fact that reflection preserves only the origin which is not a point on the sphere. Properness comes from the fact that for a single point we can choose an open hemispherical neighborhood (whose reflection does not intersect the original hemisphere), and for two points we just consider two small balls such that none of the four sets including reflections intersect. The quotient space is $\mathbb{P}^n$, the space of lines through the origin in $\mathbb{R}^n$, which is therefore also a manifold.

Definition 9.1.9. If N is a topological manifold and M is a subset of N, we say that M is a *m-dimensional submanifold* if the subspace topology makes it a manifold of dimension m.

If N is a smooth manifold and M is a topological submanifold, we say that M is a *smooth submanifold* if for every $p \in M$ there is a coordinate chart (ϕ, U) of N such that $M \cap U = \phi^{-1}\{x^1, \cdots, x^m, 0, \cdots, 0 \,|\, x^1, \cdots, x^m \in \mathbb{R}\}$.

Example 9.1.10. The typical example is something like the circle $S^1 \subset \mathbb{R}^2$. Let p be a point on the right half of the circle, and consider the polar coordinate chart ϕ defined on the set U given by right open half of the plane given by the formula

$$\phi(x, y) = \big(\arctan(y/x), \sqrt{x^2 + y^2} - 1\big).$$

Then $\phi^{-1}(\theta,0) = \{(\cos\theta, \sin\theta) \mid -\frac{\pi}{2} < \theta < \pi/2\} = U \cap S^1$. Similar charts are easy to find for other points of S^1.

More typically these charts are obtained by the Inverse Function Theorem in the following way. In this case if $F(x,y) = x^2 + y^2$, then $S^1 = F^{-1}(1)$. Consider new coordinates $s = x$ and $t = F(x,y) - 1$. This will give a genuine coordinate chart in a neighborhood of any point $(x,y) \in S^1$ where the Jacobian determinant $s_x t_y - s_y t_x \neq 0$, which happens precisely when $F_y(x,y) \neq 0$; in other words, exactly when we can solve $F(x,y) = 1$ for y in terms of x. In such a coordinate chart, $t = 0$ precisely when $F(x,y) = 1$. Explicitly, we have the coordinate chart $(s,t) = \phi(x,y) = (x, x^2 + y^2 - 1)$, which is invertible on the open upper half-plane and has inverse map $(x,y) = \phi^{-1}(s,t) = (s, \sqrt{1+t-s^2})$. Furthermore $\phi^{-1}(s,0) = \{(s, \sqrt{1-s^2}) \mid -1 < s < 1\}$ is the intersection of the circle with the upper half-plane. ☺

Theorem 9.1.11. *Suppose $F\colon N \to K$ is a smooth function from an n-dimensional smooth manifold M to a k-dimensional smooth manifold K, and that DF has rank k everywhere on $F^{-1}(r)$. Then $F^{-1}(r)$ is a smooth submanifold of N of dimension $n-k$.*

Proof. To prove this, it is sufficient to work locally in coordinate charts, and so we lose nothing by assuming $N = \mathbb{R}^n$ and $K = \mathbb{R}^k$. We are going to use the same technique as in Example 9.1.10.

Assume we have rotated N so that at a particular point $p \in F^{-1}(r)$, the $k \times n$ matrix $DF(p)$ has a nonsingular $k \times k$ matrix on the right. That is, we break up $\mathbb{R}^n$ into $\mathbb{R}^{n-k} \times \mathbb{R}^k$ and write $x \in \mathbb{R}^{n-k}$ and $y \in \mathbb{R}^k$ and $F(x,y)$. We assume that the matrix of partials with respect to the k y-variables is nonsingular at the point p.

Consider the function $G\colon \mathbb{R}^n \to \mathbb{R}^n$ given by

$$G(x,y) = (x, F(x,y) - r).$$

Then the derivative of G is a square matrix that looks like

$$DG(x,y) = \begin{pmatrix} I & 0 \\ F_x & F_y \end{pmatrix},$$

and since F_y is nonsingular, so is DG. Thus by the Inverse Function Theorem 5.2.4 G locally has an inverse function $(x,y) = (s, H(s,t))$; note that $F(s, H(s,t)) = r$ so H is the function obtained by the Implicit Function Theorem 5.2.2. Then G is a smooth coordinate chart and $G^{-1}(s,0)$ is the set of points (x,y) satisfying $F(x,y) = r$ on the open set where G has an inverse, i.e., the set $M \cap U$ where U is the domain of G as a coordinate chart. □

Most of our examples of submanifolds have had the ambient manifold $N = \mathbb{R}^d$, but in some cases we have interesting submanifolds of other manifolds.

Example 9.1.12. The best-known example is the flat torus $\mathbb{T}^2$ in $\mathbb{R}^4$, considered as the set

$$\mathbb{T}^2 = \{(w,x,y,z) \in \mathbb{R}^4 \mid w^2 + x^2 = 1, y^2 + z^2 = 1\}.$$

Obviously this is a subset of the 3-sphere with radius $\sqrt{2}$,

$$S^3_{\sqrt{2}} = \{(w,x,y,z) \in \mathbb{R}^4 \mid w^2 + x^2 + y^2 + z^2 = 2\}.$$

To see this, just consider the function $G\colon \mathbb{R}^4 \to \mathbb{R}$ given by $G(w,x,y,z) = w^2+x^2-y^2-z^2$. The function $\iota\colon S^3 \to \mathbb{R}^4$ is smooth since in the submanifold coordinate chart it just looks like $\iota(u^1,u^2,u^3) = (u^1,u^2,u^3,0)$, and thus $G\circ\iota\colon S^3 \to \mathbb{R}^4$ is a smooth map. It's easy to see that G has rank one on S^3, and thus $G^{-1}(0) = \mathbb{T}^2$ is a submanifold.

The reason this is interesting is that usually one thinks of submanifolds as having "more positive curvature" than the ambient space, based on the example of surfaces embedded in $\mathbb{R}^3$, which must have positive curvature somewhere. In this example the 3-sphere has positive sectional curvature in all directions, but the torus has zero Gaussian curvature. ☺

What's a little strange about this is that a subset of $\mathbb{R}^N$ could be a manifold, and could have a smooth structure, but not be a smooth submanifold. The best example is a square in $\mathbb{R}^2$. It's clearly not a smooth submanifold (at a corner, there's no smooth invertible map from $\mathbb{R}^2$ to $\mathbb{R}^2$ whose level sets are a square). Yet it's homeomorphic to a circle, and a circle has a smooth structure, so in that sense we could think of the square as homeomorphic to a smooth manifold. The issue is that if you lived inside the square with no ability to leave or even look outside, you'd never notice the corners. But if you could see the outside space $\mathbb{R}^2$, you *would* notice them. There's a big difference between *intrinsic* properties that can be detected by people living in the manifold and *extrinsic* properties that are noticed by people with a view of the outside world in which the manifold sits.

Every manifold we've discussed so far can be embedded as a submanifold in some $\mathbb{R}^N$ for a large N. (That's less obvious for the projective spaces $\mathbb{P}^n$ but still true, using the same sort of technique of finding a map from S^n to $\mathbb{R}^N$ which is invariant under reflection and has enough components that it's one-to-one.) This is no coincidence; in fact any manifold (which satisfies our basic assumptions of Hausdorff and second-countable) can be embedded in $\mathbb{R}^N$, so they really aren't terribly strange objects. This is not always the most convenient way to think of a particular manifold, but it's nice to know in general if you can use it.

The hard part of this theorem is proving it for noncompact manifolds, and finding the minimal dimension that will always work (which is $2n$ for an n-dimensional manifold). The complete version is due to Whitney. The basic idea for compact manifolds is to consider finitely many coordinate patches (say m) and map each one into a separate copy of $\mathbb{R}^n$; then the whole thing ends up in $\mathbb{R}^{mn}$. We just have to take care of the overlaps, which we do using a trick called a *partition of unity*. For now though I'll just tell you the result; we'll prove a somewhat easier version of it in Theorem 13.4.3.

Theorem 9.1.13. *Any n-dimensional smooth manifold which is Hausdorff and second-countable is homeomorphic to a closed subset of $\mathbb{R}^{2n}$.*

9.2. **Other examples.** The vast majority of examples come from some combination of the above techniques. But let's look at a few really useful ones that are a bit different.

The next two examples are the Grassmann manifolds and Stiefel manifolds. They are some of the most useful manifolds in applications such as electrical engineering.

Example 9.2.1. The *Grassmann manifold* (or "Grassmannian manifold" or sometimes just "Grassmannian") $Gr(k,n)$ is the set of all k-dimensional subspaces of $\mathbb{R}^n$. Some people denote it by $Gr(n,k)$ or $Gr_k(n)$ or $G_k(n)$.

If $k = 1$, the Grassmann manifold $Gr(1, n)$ is the set of lines through the origin in $\mathbb{R}^n$; hence it's the same thing as $\mathbb{P}^{n-1}$.

If $k = 2$ and $n = 3$, the Grassmann manifold is the set of planes through the origin in $\mathbb{R}^3$, and there is a correspondence between unit normal vectors to these planes and the planes themselves (except that the positive and negative unit normals give the same plane). Using that two-to-one identification, we see that we can identify the planes through the origin with the lines through the origin that they're perpendicular to, and this is actually a bijection. Hence $Gr(2, 3)$ is homeomorphic to $Gr(1, 3)$ and thus to $\mathbb{P}^2$.

After all these pages, projective space is not too hard to understand, so the first interesting Grassmann manifold is the set of 2-planes in $\mathbb{R}^4$. To describe it, we want some kind of coordinate representation for it. Now any plane is the span of two 4-vectors, so we can start with an eight-dimensional space spanned by $\overline{a} = (a_1, a_2, a_3, a_4)$ and $\overline{b} = (b_1, b_2, b_3, b_4)$. Yet clearly the resulting space is not eight-dimensional; for example I could rescale both $\overline{a}$ and $\overline{b}$ so that they're both unit vectors, and that wouldn't change the plane that the vectors span. So at most the space could be six-dimensional.

In fact it's smaller than that, since two different pairs of vectors could easily describe the same plane. Choose a vector perpendicular to the plane (there's a two-dimensional space of these in $\mathbb{R}^4$) and rotate everything around this normal vector; that changes the spanning vectors but not the actual plane. There's a two-dimensional family of perpendicular vectors, and for each vector we have a one-dimensional family of rotations around it, so it seems reasonable we could reduce the number of dimensions of $Gr(2, 4)$ by two more, to four.

To prove this, we write down coordinates. Suppose for example that we take our base point P in $Gr(2, 4)$ to be the plane spanned by vectors $(1, 0, 0, 0)$ and $(0, 1, 0, 0)$. Nearby planes should have nearby vectors spanning them, and this plane has the property that if we consider the projection onto the first two coordinates, we still get a plane (rather than a line or a point). So we choose an open set U containing P to be the set of planes in $\mathbb{R}^4$ such that the projection onto the first two coordinates is also a plane. In terms of vectors (a_1, a_2, a_3, a_4) and (b_1, b_2, b_3, b_4) which span a plane Q in U, this condition says that (a_1, a_2) and (b_1, b_2) span a plane in $\mathbb{R}^2$, and a necessary and sufficient condition for this is that $a_1b_2 - a_2b_1 \neq 0$. If that's the case, then we can rotate the vectors so that the first two components are actually $(1, 0)$ and $(0, 1)$; in other words we span the plane Q by $(1, 0, a_3, a_4)$ and $(0, 1, b_3, b_4)$.

In this way we construct an identification between U and $\mathbb{R}^4$. We could clearly do the same thing near any other plane P: choose an orthonormal basis $\{e_1, e_2, e_3, e_4\}$ of $\mathbb{R}^4$ such that $\{e_1, e_2\}$ spans P; then the spans of $\{e_1 + a_3e_3 + a_4e_4, e_2 + b_3e_3 + b_4e_4\}$ are all distinct planes. In this way we get a coordinate chart around every plane of $Gr(2, 4)$, thus proving that $Gr(2, 4)$ is a 4-dimensional manifold. Clearly the same technique generalizes to give a coordinate chart for any $Gr(k, n)$. ☺

Example 9.2.2. The *Stiefel manifold* $V(k, n)$ is the set of orthonormal sets of k vectors in $\mathbb{R}^n$. Note that an orthonormal set of k vectors is automatically linearly independent, so it spans a k-plane in $\mathbb{R}^n$; hence there is a nice surjective map $F\colon V(k, n) \to Gr(k, n)$.

If $k = 1$ then there is only one vector, and it needs to have unit length. So $V(1, n)$ is the standard round sphere S^{n-1}.

If $k = n$ then an orthonormal set of n vectors can be listed as the columns of an $n \times n$ matrix which is orthogonal. So $V(n,n)$ is homeomorphic to the set of orthogonal $n \times n$ matrices $O(n)$ satisfying $A^T A = I_n$. (We will discuss the orthogonal group $O(n)$ shortly.)

If $k = n-1$ then there are only two unit vectors we can pick to complete the $n-1$ orthonormal vectors to n orthonormal vectors, which means that every element in $V(n-1,n)$ corresponds to *two* elements of $V(n,n)$. Notice that a matrix in $O(n)$ has $(\det A)^2 = 1$, so that either $\det A = 1$ or $\det A = -1$. The matrices with $\det A = 1$ are called the *special orthogonal group* $SO(n)$; we will discuss this shortly as well. If we have $n-1$ orthonormal vectors which we put in the first $n-1$ columns of a matrix, then there is a unique way to choose the sign of the last unit vector orthogonal to all the others so that the determinant of the matrix is one. Hence $V(n-1,n)$ is homeomorphic to $SO(n)$.

More generally we can write a $k \times n$ matrix A consisting of the orthonormal vectors in columns, and the product $A^T A$ is a $k \times k$ matrix I_k. So if $F\colon \mathbb{R}^{nk} \to \mathbb{R}^{k^2}$ is the map $F(A) = A^T A$ from $n \times k$ matrices to $k \times k$ matrices, then the Stiefel manifold is $V(k,n) = F^{-1}(I_k)$, which makes clear that $V(k,n)$ is actually a manifold.

In general for any k we can pick a unit vector e_1 in $\mathbb{R}^n$, and there is an $(n-1)$-dimensional space of these (the unit sphere S^{n-1}). Having chosen e_1 the space of vectors orthogonal to e_1 is homeomorphic to $\mathbb{R}^{n-1}$, and we then have to pick a unit vector e_2 in that space; this is an $(n-2)$-dimensional space homeomorphic to S^{n-2}. Continuing, it's easy to see that the dimension of $V(k,n)$ is

$$\sum_{j=1}^{k} (n-j) = k\big(n - \tfrac{k+1}{2}\big).$$

It is also easy to see how we can construct a coordinate chart near any point of the Stiefel manifold using this idea: just take a coordinate chart on each sphere. ☺

The most popular sort of manifold is a Lie group, since it's got a lot of symmetry. The geometry reduces to algebra in this case, and so computations simplify a lot.

Example 9.2.3. A *Lie group* is a topological manifold which is also a group, and such that the group operations are continuous.

The most basic example is $GL(n)$, the *general linear group* consisting of all invertible $n \times n$ matrices. We think of it as a subset of $\mathbb{R}^{n^2}$. The determinant function $\det\colon \mathbb{R}^{n^2} \to \mathbb{R}$ is continuous (using the explicit formula for it in (3.3.1)), and so the singular matrices (with determinant zero) form a closed set. So the invertible matrices are an open set, and any open subset of a manifold is another manifold. Notice that $GL(n)$ is not connected: it has two components consisting of those matrices with positive determinant and those with negative determinant. The group operation is matrix multiplication, which we know is continuous since it just involves multiplication of components, and therefore $GL(n)$ is a Lie group.

Example 9.2.4. The next example is $SL(n)$, the *special linear group*, consisting of $n \times n$ matrices such that $\det A = 1$. The value 1 is a regular value of the determinant function; for example if $n = 2$ and a matrix is written $\left(\begin{smallmatrix} w & x \\ y & z \end{smallmatrix}\right)$, we have $\det(w,x,y,z) = wz - xy$ so that $D\det = \begin{pmatrix} z & -y & -x & w \end{pmatrix}$. As long as $wz - xy = 1$, at least one of these numbers is nonzero so the rank of $D\det$ is maximal. The same idea works in general. Thus $SL(n)$ is a smooth manifold of dimension $n^2 - 1$. Like all

groups we will discuss, it inherits the group operation of matrix multiplication from $GL(n)$, and it is a subgroup of $GL(n)$ since $\det(AB) = (\det A)(\det B) = 1$. ☺

Example 9.2.5. A third popular example is $O(n)$, the *orthogonal group*. This is the set of $n \times n$ matrices such that $A^T A = I_n$, as discussed in Example 9.2.2. It has two connected components since $A^T A = I_n$ implies that either $\det A = 1$ or $\det A = -1$. We can think of it as defined by an implicit function theorem, using the map $F\colon \mathbb{R}^{n^2} \to \mathbb{R}^{n(n+1)/2}$ given by $F(A) = A^T A$. (Note that the image of this map is symmetric matrices only, which are completely determined by the upper triangular portion including the diagonal; there are $1 + 2 + \cdots + n = n(n+1)/2$ components here.) We can check that I_n is a regular value; for example if $n = 2$ then the map is $F(w, x, y, z) = (w^2 + y^2, wx + yz, x^2 + z^2)$, which has maximal rank on $F^{-1}(1, 0, 1)$. Thus $O(n) = F^{-1}(I_n)$ is a smooth manifold of dimension $n(n-1)/2$. It is a subgroup of $GL(n)$ since $(AB)^T(AB) = B^T(A^T A)B = B^T B = I_n$.

A closely related example is $SO(n)$, the *special orthogonal group*. This is just the intersection of $O(n)$ with $SL(n)$; in other words it is the connected component of $O(n)$ which contains the identity I_n. It also has dimension $n(n-1)/2$. If $n = 3$ then it is a 3-dimensional group which is the group of orientation-preserving rotations of $\mathbb{R}^3$ (or of S^2). In general the word "special" applied to Lie groups is a shortcut for "determinant one." ☺

The orthogonal group is the set of linear transformations that preserve distances, since it preserves the inner product: if $A \in O(n)$, then

$$\langle Ax, Ay \rangle = \langle x, A^T Ay \rangle = \langle x, y \rangle.$$

Hence it consists of the symmetries of an inner product. If some other tensor is important to us, then its symmetry group will also be important. For example, if we have an antisymmetric matrix ω which is nondegenerate (which must be on an even-dimensional space $\mathbb{R}^{2n}$), then the matrices that preserve ω form the *symplectic group* $Sp(n)$. Specifically, we have $A^T \omega A = \omega$ for all $A \in Sp(n)$. As before, this is a smooth manifold which is also a subgroup.

The *unitary group* $U(n)$ is the set of complex matrices A such that $A^* A = I_n$. The Hermitian matrices with $B^* = B$ have real dimension n^2, while all complex matrices have dimension $2n^2$, so that the unitary group has dimension n^2. The determinant of a unitary matrix is a complex number of norm 1, so it could be any point on the circle; the *special unitary group* $SU(n)$ is of course the unitary matrices that have determinant one, so it has dimension $n^2 - 1$. ☺

Although the above is in some sense the "standard list" of Lie groups, there are many others. Some special ones are $\mathbb{R}\backslash\{0\}$, $\mathbb{R}^2\backslash\{(0,0)\}$ and $\mathbb{R}^4\backslash\{(0,0,0,0)\}$ under multiplication by real numbers, complex numbers, and quaternionic numbers respectively. (Recall that the quaternions are numbers of the form $a + bi + cj + dk$ where $a, b, c, d \in \mathbb{R}$ and i, j, k satisfy the multiplication rules $i^2 = j^2 = k^2 = -1$ and $ij = k$, $jk = i$, $ki = j$, $ji = -k$, $kj = -i$, and $ik = -j$.) These are the *only* Euclidean spaces that have a multiplicative group structure which respects the vector space structure.

From the reals, complex numbers, and quaternions, we get via looking at the unit-length elements the group structures also on $S^0 = \{-1, 1\} = \mathbb{Z}_2$, on S^1, and on S^3. No other spheres have group structures, although S^7 comes close. (You can define Cayley numbers which have some of the right properties, but they don't have associativity.)

Finally we can consider complex versions of many of the spaces discussed above. For example the *complex projective space* $\mathbb{CP}^n$ is the set of complex lines through the origin in $\mathbb{C}^{n+1}$, so it has dimension $2n$ (complex dimension n). Similarly we have complex Grassmann manifolds and complex Stiefel manifolds (the latter defined in terms of a complex inner product).

10. Vectors

"All his life has he looked away to the future, to the horizon. Never his mind on where he was, what he was doing."

This Chapter is probably the most important in the entire book: in previous chapters you have essentially been reviewing old material such as linear algebra, multivariable calculus, and topology, while here we have a fundamentally new concept (which looks rather old). Furthermore, every operation on a smooth manifold is defined in terms of what it does to vectors, which means you won't really understand anything that follows unless you have a deep understanding of vectors. Hence read carefully.

10.1. **Tangent vectors, historically.** In the traditional description which one learns in vector calculus, vectors are lists of numbers representing a direction and magnitude; they may have a base point, although we can always translate in an obvious way in order to assume the base point is at the origin. In $\mathbb{R}^n$, points and vectors are described in exactly the same way. None of these things generalize, so we need to figure out what we really want vectors to do and how vectors are different from points.

The motivation for everything we will do in this Chapter comes from the study of surfaces in $\mathbb{R}^3$. There are two basic ways to describe a surface: as a level set of a function $G\colon \mathbb{R}^3 \to \mathbb{R}$, or as the image of a parametrization $F\colon \mathbb{R}^2 \to \mathbb{R}^3$. In the first case, we want DG to have maximal rank (one), and in the second case, we want DF to have maximal rank (two). We saw in Section 7.2 that parametrizations may be difficult globally, but locally either of these maximal rank conditions will ensure that the surface is locally a smooth manifold. (Two hundred years ago, of course, working locally was the best one could hope for anyway.)

A tangent vector is the derivative of a curve that lies completely within the surface M. If $\gamma(t)$ satisfies $\gamma(0) = p$, then $\gamma'(0)$ is a tangent vector to M at p. If the surface is $G^{-1}(r)$ and the curve $\gamma(t)$ is given by $\gamma(t) = \big(x(t), y(t), z(t)\big)$, then we have $G\big(x(t), y(t), z(t)\big) = r$ for all t. Thus by the Chain Rule we have

$$G_x\big(x(t), y(t), z(t)\big)\dot{x}(t) + G_y\big(x(t), y(t), z(t)\big)\dot{y}(t) + G_z\big(x(t), y(t), z(t)\big)\dot{z}(t) = 0.$$

If $\gamma(0) = p$ and $\dot{\gamma}(0) = \langle a, b, c\rangle$, this tells us that $G_x(p)a + G_y(p)b + G_z(p)c = 0$. In vector calculus, you'd write $\operatorname{grad} G(p) = \langle G_x(p), G_y(p), G_z(p)\rangle$ and call it the gradient of G, and say that the condition for v to be a tangent vector is that $\operatorname{grad} G(p) \cdot v = 0$. This explains why we need the rank of G to be one: otherwise $\operatorname{grad} G(p) = 0$ and we will not actually get a tangent plane. In our approach we say instead that the tangent space is the kernel of the map $DG(p)\colon \mathbb{R}^3 \to \mathbb{R}$, which is actually a covector.

Now suppose the surface is $M = F[\Omega]$ for some open $\Omega \subset \mathbb{R}^2$ and a maximal-rank function F. Write $(x, y, z) = F(u, v) = \big(f(u, v), g(u, v), h(u, v)\big)$. If $p = F(u_0, v_0)$, then a curve $\gamma\colon (-\varepsilon, \varepsilon) \to M$ satisfying $\gamma(0) = p$, then we can write $\gamma(t) = F\big(u(t), v(t)\big)$ for some functions $u(t)$ and $v(t)$ with $u(0) = u_0$ and $v(0) = v_0$. Now by the Chain Rule, we have

$$\gamma'(t) = \frac{d}{dt} F\big(u(t), v(t)\big) = F_u\big(u(t), v(t)\big)\dot{u}(t) + F_v\big(u(t), v(t)\big)\dot{v}(t).$$

If $\dot{u}(0) = a$ and $\dot{v}(0) = b$, then we have

$$\gamma'(0) = F_u(u_0, v_0)\, a + F_v(u_0, v_0)\, b.$$

Since a and b are arbitrary numbers, we have expressed any tangent vector at p as a linear combination of two basis vectors F_u and F_v. This is why we needed F to have rank two: so that these vectors would span a two-dimensional space. Of course, a surface may be described by many possible parametrizations, so the vectors F_u and F_v by themselves are not important: only their span is. Classically one might take the normalized cross-product $N = (F_u \times F_v)/|F_u \times F_v|$ in order to obtain a vector that is perpendicular to the tangent plane; however nothing like this works in any other dimension, so we will not care about it.

Example 10.1.1. The simplest example is the 2-sphere S^2. Let $p \in S^2$ be a point $p = (x, y, z)$ with $G(x, y, z) = x^2 + y^2 + z^2 = 1$; then the condition for a vector $v = \langle a, b, c \rangle$ to be in $T_p S^2$ is that

$$DG(x, y, z)v = \begin{pmatrix} 2x & 2y & 2z \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix}.$$

In other words, we just need $xa + yb + zc = 0$.

Alternatively we can use a parametrization $F(u, v) = (\sin u \cos v, \sin u \sin v, \cos u)$ on the open set $0 < u < \pi$, $0 < v < 2\pi$. The vectors F_u and F_v are given by

$$F_u = \begin{pmatrix} \cos u \cos v \\ \cos u \sin v \\ -\sin u \end{pmatrix} \qquad \text{and} \qquad F_v = \begin{pmatrix} -\sin u \sin v \\ \sin u \cos v \\ 0 \end{pmatrix}.$$

Certainly these vectors are both in the kernel of the covector $DG\big(F(u, v)\big) = \begin{pmatrix} 2 \sin u \cos v & 2 \sin u \sin v & 2 \cos u \end{pmatrix}$, as we'd expect. They are linearly independent as long as $\sin u \neq 0$.

Finally we can view portions of the sphere as graphs of functions, via a parametrization like $H(u, v) = (u, v, \sqrt{1 - u^2 - v^2})$ defined on the open unit disc. Then the vectors spanning the tangent space are

$$H_u = \begin{pmatrix} 1 \\ 0 \\ -\frac{u}{\sqrt{1-u^2-v^2}} \end{pmatrix} \qquad \text{and} \qquad H_v = \begin{pmatrix} 0 \\ 1 \\ -\frac{v}{\sqrt{1-u^2-v^2}} \end{pmatrix}.$$

Again these are both orthogonal to $DG\big(H(u, v)\big) = \begin{pmatrix} 2u & 2v & 2\sqrt{1 - u^2 - v^2} \end{pmatrix}$.

Figure 10.1 shows a couple of typical tangent spaces to S^2.

☺

Now the sphere has lots of tangent spaces, all of which are isomorphic to a two-dimensional vector subspace of $\mathbb{R}^3$, but none of which are actually the same. So we certainly could not add vectors at $T_{(0,0,1)} S^2$ to vectors in $T_{(1,0,0)} S^2$ and expect to get anything that makes sense. Nor is there any way to translate vectors from $(0, 0, 1)$ to $(1, 0, 0)$. Now you might object that this is clearly false as you can just rotate all of $\mathbb{R}^3$ around the origin until the point $(0, 0, 1)$ is $(1, 0, 0)$, and then the tangent space at one point can be identified with the other point. But this depends on having lots of isometries (which a general space won't have) and even more importantly it depends on how you do it. If you rotated $(0, 0, 1)$ into $(1, 0, 0)$ and looked at what happened to particular tangent vectors, then instead rotated $(0, 0, 1)$

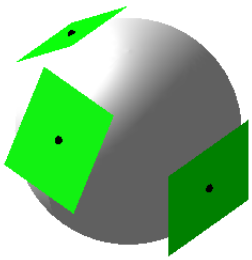

FIGURE 10.1. The usual picture of tangent spaces to points of a surface. The tangent planes are all isomorphic to $\mathbb{R}^2$, but not in a natural or coordinate-independent way.

into $(0,1,0)$ and then into $(1,0,0)$, you'd find that the vector space isomorphisms from one tangent space to the other are different. So for now you should throw away the idea of relating tangent vectors at one point to tangent vectors at another point; in this Chapter we will work with tangent vectors at a single point.

Furthermore there is a philosophical difficulty with the approach of defining tangent spaces through reference to an ambient space like $\mathbb{R}^3$. Particles moving in the surface have positions and velocities (since their trajectories are differentiable curves), but people living inside the surface cannot see the velocity as a vector in $\mathbb{R}^3$. So what are they seeing? A manifold is defined without reference to ambient space, and so vectors should properly also be defined without reference to the ambient space. Finally there is the problem that we are treating vectors tangent to a manifold as fundamentally different from vectors in $\mathbb{R}^2$ or $\mathbb{R}^3$.

10.2. **Tangent vectors, conceptually.** In standard vector calculus, the Cartesian coordinates lead to vectors; on the plane we have e_x and e_y which have unit length and point in the directions of increasing x and y respectively. By the translation invariance, specifying e_x and e_y at a single point (e.g., the origin) will specify the same directions at all points. If we are working in some other coordinate system, the first thing we have to do is figure out how to change the vectors. For example, in polar coordinates, we have unit vectors e_r and e_θ, whose directions change depending on r and θ. Drawing a diagram as shown in Figure 10.2, we can figure out how the vectors e_x and e_y are related to e_r and e_θ. We easily see that $e_r = \cos\theta\, e_x + \sin\theta\, e_y$ and $e_\theta = -\sin\theta\, e_x + \cos\theta\, e_y$.

The problems with this approach, when generalizing to arbitrary coordinate systems, are numerous:

- We start with the Cartesian coordinate vectors and define everything else in terms of them. This obviously violates the principle that everything should be defined independently of coordinates and without any preferred basis.
- We need to draw a diagram. In general the level sets may not be well-understood, the diagrams may be quite complicated, and there may not be any general pattern for the formulas.
- For Cartesian coordinates the unit vectors happen to be derivatives of the level curves: the horizontal line $\gamma_x(t) = (x_o+t, y_o)$ has derivative $\gamma_x'(0) = e_x$ at (x_o, y_o), and similarly for the y-direction. On the other hand, for polar

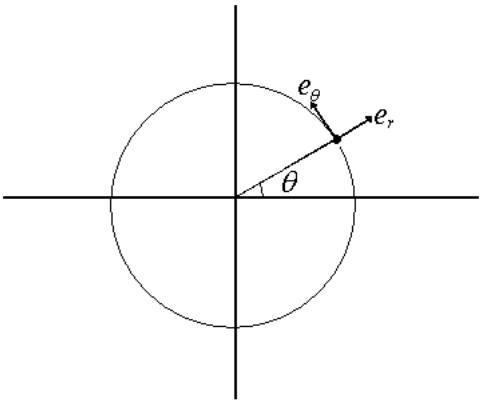

FIGURE 10.2. The standard basis for polar vectors

coordinates this is not so nice: for $\gamma_r(t) = (r_o + t, \theta_o)$ we have $\gamma_r'(0) = e_r$, but for $\gamma_\theta(t) = (r_o, \theta_o + t)$ we have $\gamma_\theta'(0) = r_o e_\theta$ instead of e_θ. In non-Cartesian systems, we thus have to decide whether to work with the actual derivatives of coordinate curves, or with a unit basis.

- The third problem is that a general coordinate system may not have orthogonal level curves. (All the classical coordinate systems defined above do, but this need not be true in general.) If this occurs, it is not at all clear how we should choose the e's.

To get a definition of vectors that does not depend on coordinates, we should first figure out what we want to do with vectors. Intuitively we think of them as tangents to curves: for example a vector $ae_x + be_y$ located at (x_o, y_o) is the tangent vector $\gamma'(0)$ to the curve $\gamma(t) = (x_o + at, y_o + bt)$, and more generally it's also the derivative of *any* curve $\gamma(t) = (\gamma_1(t), \gamma_2(t))$ satisfying the conditions $\gamma_1(0) = x_o$, $\gamma_2(0) = y_o$, $\gamma_1'(0) = a$, and $\gamma_2'(0) = b$. This still doesn't get us coordinate-independence, though: to know whether two curves have the same tangent vector, we are taking their coordinate components and checking whether *those* have the same derivatives. We need something a bit more abstract, but this idea will be helpful.

What we can do instead is the following: consider curves $\gamma\colon (a, b) \to M$ on the plane and functions $f\colon M \to \mathbb{R}$ from the plane to the real numbers. Curves and functions are topological objects: they make sense independently of any particular coordinate system. Then $f \circ \gamma\colon (a, b) \to \mathbb{R}$ is just a real function of a real variable, and if it is smooth, then $(f \circ \gamma)'(t_o)$ is a particular number for any $t_o \in (a, b)$, again independently of any coordinate system.

Example 10.2.1. For a particular example, consider again $\mathbb{R}^2 \cong \mathbb{C}$, as in Chapter 6. Let $f\colon \mathbb{C} \to \mathbb{R}$ be defined by $f(z) = \mathrm{Re}(z^2)$, and let $\gamma\colon \mathbb{R} \to \mathbb{C}$ be defined by $\gamma(t) = (t+1)e^{it}$. Then $(f \circ \gamma)(t) = \mathrm{Re}((t+1)^2 e^{2it}) = (t+1)^2 \cos 2t$, and when $t = 0$ we have $(f \circ \gamma)'(0) = 2$. We can also do this computation in a particular coordinate system: in rectangular coordinates $\mathbf{x}\colon \mathbb{C} \to \mathbb{R}^2$, we have $f \circ \mathbf{x}^{-1}(x, y) = x^2 - y^2$ and

$\mathbf{x}\circ\gamma(t) = \big((t+1)\cos t, (t+1)\sin t\big)$. We can compute using the Chain Rule that

$$\begin{aligned}\frac{d}{dt}\Big|_{t=0}(f\circ\gamma) &= \frac{d}{dt}\Big|_{t=0}(f\circ\mathbf{x}^{-1})\circ(\mathbf{x}\circ\gamma)\\ &= \frac{\partial(f\circ\mathbf{x}^{-1})}{\partial x}\Big|_{\mathbf{x}(\gamma(0))}\frac{d(x\circ\gamma)}{dt}\Big|_{t=0} + \frac{\partial(f\circ\mathbf{x}^{-1})}{\partial y}\Big|_{\mathbf{x}(\gamma(0))}\frac{d(y\circ\gamma)}{dt}\Big|_{t=0}\\ &= 2x\big|_{(1,0)}\big(-(t+1)\sin t+\cos t\big)|_{t=0} - 2y\big|_{(1,0)}\big((t+1)\cos t+\sin t\big)|_{t=0}\\ &= 2.\end{aligned}$$

We can do the same computation in polar coordinates $\mathbf{u}\colon U\subset\mathbb{C}\to\mathbb{R}^2$, where $f\circ\mathbf{u}^{-1}(r,\theta) = r^2\cos 2\theta$ and $\mathbf{u}\circ\gamma(t) = (t+1,t)$, and of course we get the same answer 2, even though $(\mathbf{u}\circ\gamma)'(0)$ is different. You can check the details for practice. ☺

More generally, for a smooth n-dimensional manifold M with any coordinate system $\mathbf{y}\colon U\subset M\to\mathbb{R}^n$, we will have by the Chain Rule that

$$\frac{d}{dt}\Big|_{t=0}(f\circ\gamma)(t) = \sum_{k=1}^{n}\frac{\partial}{\partial y^k}(f\circ\mathbf{y}^{-1})\Big|_{\mathbf{y}\circ\gamma(0))}\cdot\frac{d}{dt}\Big|_{t=0}(y^k\circ\gamma)(t). \tag{10.2.1}$$

This is extremely important! It says that $f\circ\gamma_1$ and $f\circ\gamma_2$ have the same derivative for all functions f if and only if $(y^k\circ\gamma_1)'(0) = (y^k\circ\gamma_2)'(0)$ for all k; that is, if and only if the curves have the same derivative in any coordinate chart. As a result, we can say two curves $\gamma_1,\gamma_2\colon(a,b)\to M$ with $\gamma_1(t_o) = \gamma_2(t_o)$ *have the same derivative at* t_o if and only if, for every smooth function $f\colon M\to\mathbb{R}$, we have $(f\circ\gamma_1)'(t_o) = (f\circ\gamma_2)'(t_o)$. This gets us a a coordinate-independent notion of vectors, and to use this we think of vectors as being objects that are used to differentiate functions in certain directions.

In Figure 10.3 there are several curves with the same derivative at $t=0$.

Remark 10.2.2. It is not yet clear that there are *any* smooth functions $f\colon M\to\mathbb{R}$. Obviously given a coordinate chart (ϕ,U) we may define a function $f\colon U\to\mathbb{R}$ in any way we like, by taking an arbitrary function $\tilde{f}\colon\mathbb{R}^n\to\mathbb{R}$ and setting $f = \tilde{f}\circ\phi$. But we don't know in general that a smooth function defined on an open set can be extended to a smooth function on the entire manifold. Hence technically when we want to compute objects near a point, we actually mean all objects may only be defined in a neighborhood of that point. We will fix this later on when we talk about bump functions.

10.3. **Tangent vectors, formally.** To formalize this discussion, we have to define exactly what we mean by smooth curves and smooth functions on M. The basic notion is that the curves and functions are purely topological objects and defined independently of coordinates, but to do calculus, we need to introduce coordinates. Our point of view is that "honest calculus" only really works for functions from $\mathbb{R}$ to $\mathbb{R}$, and all higher-dimensional calculus (whether on $\mathbb{R}^n$ or on a smooth manifold) only makes sense when reduced to real functions.

Definition 10.3.1. Suppose M is a smooth n-dimensional manifold. A curve $\gamma\colon(a,b)\to M$ is called *smooth* if, for every $t_o\in\gamma$ there is a coordinate chart $(\mathbf{x},U)$ with $\gamma(t_o)\in U$ such that $\mathbf{x}\circ\gamma\colon(a,b)\to\mathbb{R}^n$ has infinitely many continuous derivatives.

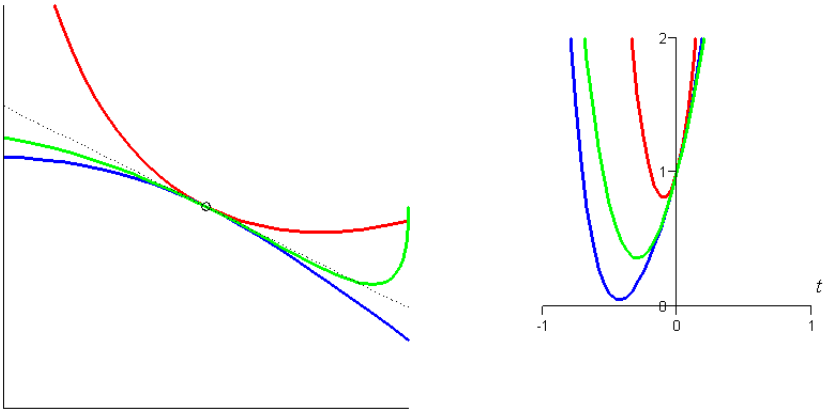

FIGURE 10.3. On the left, we have several parametric curves $\gamma_i\colon \mathbb{R} \to M$ which all appear to have the same derivative at the circled point (represented by the dotted tangent line). On the right, we compose each curve γ_i with a function $f\colon M \to \mathbb{R}$ and plot $f \circ \gamma_i(t)$ as a function of t, then compute the one-variable derivative at $t = 0$. As long as we get the same result no matter which function f we compose with, the curves will have the same derivative.

A function $f\colon M \to \mathbb{R}$ is called *smooth* if, for every $p \in M$ there is a coordinate chart $(\mathbf{x}, U)$ with $p \in U$ such that $f \circ \mathbf{x}^{-1}\colon \mathbf{x}[U] \subset \mathbb{R}^n \to \mathbb{R}$ has infinitely many continuous partial derivatives.

Now we finally come to the definition of tangent vectors, which are the basic objects in differential geometry around which everything else is structured.

Definition 10.3.2 (Tangent vectors)**.** Two smooth curves γ_1 and γ_2 defined on neighborhoods of $t_o \in \mathbb{R}$ are said to have *the same derivative at* t_o if, for every smooth function $f\colon \Omega \subset M \to \mathbb{R}$, we have

$$(f \circ \gamma_1)'(t_o) = (f \circ \gamma_2)'(t_o).$$

Having the same derivative at t_o is an equivalence relation.

The *tangent space at* $p \in M$, denoted by T_pM, is defined as the set of all curves $\gamma\colon (-\varepsilon, \varepsilon) \to M$, for some $\varepsilon > 0$, satisfying $\gamma(0) = p$, modulo the equivalence relation of having the same derivative. A *tangent vector* $v \in T_pM$ is thus an equivalence class $v = [\gamma]$ of locally-defined curves γ through p that all have the same derivative, and we write $v = \gamma'(0) = \frac{d\gamma}{dt}\big|_{t=0}$ for every $\gamma \in v$.

Observe that the only actual derivative computation we ever do here is the derivative of a function $f \circ \gamma_1\colon (a, b) \to \mathbb{R}$; we defined the derivative of a curve $\gamma'(0)$ rather indirectly. By this definition, tangent vectors are actually fairly complicated and abstract objects. But because of this definition, we can think of tangent vectors as being pointwise operators on functions: a tangent vector v at the point p operates on smooth functions f defined in any neighborhood of p by the formula

$$v(f) = \frac{d}{dt} f\big(\gamma(t)\big)\Big|_{t=0}, \text{ where } \gamma \in v. \tag{10.3.1}$$

By definition of a tangent vector as an equivalence class, the number $v(f)$ is independent of $\gamma \in v$ (although of course it depends on the function f).

Remark 10.3.3. The number $v(f)$ is also independent of the domain of f, as long as it is some open set. Hence we think of vectors as being defined on *germs of functions at p*: here a germ of a function at p is a pair (f, Ω) where Ω is an open set containing p and $f\colon \Omega \to \mathbb{R}$ is smooth. (f, Ω_1) and (g, Ω_2) are equivalent if for some open $U \subset \Omega_1 \cap \Omega_2$ with $p \in U$, we have $f|_U = g|_U$. The idea is that it makes sense to think of germs as telling you all the derivatives at p but not much beyond that (however note that the derivatives are not enough to tell you the germ: the function in Example 7.2.2 is not germ-equivalent to the zero function at $x = 0$ in spite of having all the same derivatives). Properly when one is thinking of tangent vectors as derivative operators on functions, the domain of these operators is the space of germs at p. Later on in Chapter 13 we will prove that for any germ there is a representative $f\colon M \to \mathbb{R}$.

First we check that a tangent vector is completely specified by its operation on smooth functions.

Lemma 10.3.4. *Suppose v and w are both tangent vectors in T_pM. If $v(f) = w(f)$ for all smooth functions f defined in some neighborhood of p, then $v = w$.*

Proof. This proof just involves checking our definitions. Since a vector is an equivalence class of curves, we can choose representative curves $\alpha \in v$ and $\beta \in w$. Then for any smooth function f, we have

$$v(f) = \frac{d}{dt} f\big(\alpha(t)\big)\Big|_{t=0} = \frac{d}{dt} f\big(\beta(t)\big)\Big|_{t=0} = w(f).$$

Since the middle two terms are equal for any function f, the curves α and β must have the same derivative at $t = 0$ (by definition). As a result, the equivalence class v must be the same as the equivalence class w. □

The definition of tangent vector makes the following proposition easy; it gives a method that is often more convenient for checking that two vectors are the same.

Proposition 10.3.5. *Suppose $v, w \in T_pM$, and that for some coordinate chart $(\mathbf{x}, U)$ defined in a neighborhood of p, and for some representatives $\alpha \in v$ and $\beta \in w$, that*

$$\frac{d}{dt} x^k\big(\alpha(t)\big)\Big|_{t=0} = \frac{d}{dt} x^k\big(\beta(t)\big)\Big|_{t=0}. \tag{10.3.2}$$

Then $v = w$.

Conversely if $v = w$ then (10.3.2) *is true for* every *chart and* every *pair of representatives.*

Proof. In the coordinate system $(\mathbf{x}, U)$, the Chain Rule (10.2.1) gives

$$v(f) = \frac{d}{dt}(f \circ \alpha)(t)\Big|_{t=0} = \sum_{k=1}^{n} \frac{\partial}{\partial x^k}(f \circ \mathbf{x}^{-1})\Big|_{\mathbf{x}(p)} \frac{d}{dt} x^k\big(\alpha(t)\big)\Big|_{t=0}, \tag{10.3.3}$$

which shows that $v(f)$ depends only on the components $(x^k \circ \alpha)'(0)$ since the components $\frac{\partial}{\partial x^k}\big|_{\mathbf{x}(p)}(f \circ \mathbf{x}^{-1})$ are independent of v. Since the value $v(f)$ does not depend on choice of $\alpha \in v$, the numbers $(x^k \circ \alpha)'(0)$ do not depend on α either. Thus

$v(f) = w(f)$ for every smooth function f if and only if $(x^k \circ \alpha)'(0) = (x^k \circ \beta)'(0)$ for some representatives α of v and β of w. By Lemma 10.3.4, this tells us that $v = w$.

To prove the converse, just notice that for any coordinate chart $(\mathbf{x}, U)$, the individual component functions x^k are smooth functions $x^k \colon U \to \mathbb{R}$. So (10.3.2) is just a special case of Lemma 10.3.4 using definition (10.3.1). □

We now want to actually put a vector space structure on the tangent space. To do this, we need to define multiplication by scalars and addition of vectors. In vector calculus we would just do this by multiplying the actual curves by scalars, and adding the curves together, using the fact that points and vectors are the same thing. Here we are keeping points and vectors separate, so we can't really do this. For example, if α and β are curves representing vectors v and w in T_pM, then $\alpha(0) = p$ and $\beta(0) = p$ but $(\alpha + \beta)(0) = 2p$; thus we can't really add the curves unless we allow ourselves to move the vectors around. Instead it is more useful to think of vectors as operators that differentiate functions.

Definition 10.3.6. Suppose $v_1, v_2, v \in T_pM$. We say that $v = v_1 + v_2$ if, for every smooth function f defined in a neighborhood of p, we have

$$v(f) = v_1(f) + v_2(f).$$

Similarly if $a \in \mathbb{R}$, we say that $v = av_1$ if, for every smooth function f defined in a neighborhood of p, we have

$$v(f) = av_1(f).$$

It is straightforward to check that T_pM satisfies the usual vector space axioms with this definition. Notice that we do not yet know that this vector space is n-dimensional. We will prove this by constructing an explicit basis in the next Section.

10.4. **Tangent vectors in coordinates.** We have seen in Proposition 10.3.5 that a vector v is completely determined by the numbers $(x^k \circ \gamma)'(0)$ for any coordinate system $\mathbf{x}$ for every representative curve γ. As a result, these numbers should be the coefficients of v in some basis. Below we actually construct this basis. What we are aiming for is a generalization of the standard basis e_x, e_y for $\mathbb{R}^2$; to obtain it, we think of e_x as being the derivative of the coordinate curve $x = t + \text{const}$, $y = \text{const}$. The construction below does this. It looks a little awkward, since the coordinate curves are defined naturally in terms of the coordinates themselves: thus we need to apply the coordinate inverse to actually get a curve in M.

Definition 10.4.1. Suppose $(\mathbf{x}, U)$ is a coordinate chart on M, with $p \in U$. Let x^1, x^2, $\ldots x^n$ be the smooth functions representing each of the coordinates, with $c^k = x^k(p)$. Since a coordinate map $\mathbf{x}$ is a homeomorphism, we know that $\mathbf{x}[U]$ is an open subset of $\mathbb{R}^n$ containing $(c^1, c^2, \ldots, c^n)$; hence it also contains an open box $(c^1 - \varepsilon, c^1 + \varepsilon) \times \cdots \times (c^n - \varepsilon, c^n + \varepsilon)$. For each $k \in \{1, \ldots, n\}$, let $\gamma_{x^k} \colon (-\varepsilon, \varepsilon) \to M$ be the curve $\gamma_{x^k}(t) = \mathbf{x}^{-1}(c^1, \ldots, c^k + t, \ldots, c^n)$. For each k, define $\xi_{x^k} \in T_pM$ to be the tangent vector which is the equivalence class of γ_{x^k}. Clearly the ξ_{x^k}'s will depend on the particular coordinate system. (This notation for the basis vectors is just temporary.)

The first thing to observe is that the coordinate vectors ξ_{x^k} do not generally agree with the usual bases used in vector calculus. For example, in polar coordinates, we have the orthonormal basis e_r, e_θ defined in Section 10.2; on the other hand, $\xi_r = e_r$ but $\xi_\theta = re_\theta$, since the θ coordinate curves are given (in rectangular coordinates) by the formula $\gamma_\theta(t) = \big(r_o\cos(\theta_o+t), r_o\sin(\theta_o+t)\big)$, and thus $\xi_\theta = \gamma_\theta'(0) = -r_o\sin\theta_o e_x + r_o\cos\theta_o e_y = r_o e_\theta$. We will prefer using the nonorthonormal bases ξ_{x^k} because the transition formulas between coordinate systems are much easier (as we will see soon).

Now we come to the main Proposition, which expresses all of the abstract notions in concrete terms; in particular it tells us that T_pM is actually an n-dimensional vector space.

Proposition 10.4.2. *In the coordinate chart $(\mathbf{x}, U)$, the vectors ξ_{x^k} for $1 \le k \le n$ form a basis for T_pM.*

Proof. First we show that $\{\xi_{x^k}\}$ span, i.e., that any $v \in T_pM$ can be written as $v = \sum_{k=1}^n a^k \xi_{x^k}$ for some numbers $\{a^k\}$. By the definition of vector addition and scalar multiplication, we just need to show that for any smooth function f defined in a neighborhood of p, we have

$$v(f) = \sum_{k=1}^n a^k \xi_{x^k}(f). \tag{10.4.1}$$

First we compute $\xi_{x^k}(f)$:

$$\xi_{x^k}(f) = \sum_{j=1}^n \frac{\partial}{\partial x^j}(f\circ\mathbf{x}^{-1})\Big|_{\mathbf{x}(\gamma_{x^k}(0))} \frac{d}{dt}x^j(\gamma_{x^k}(t))\Big|_{t=0}.$$

Now use fact that $x^j(\gamma_{x^k}(t)) = x^j(p)$ if $j \neq k$ and $x^k(\gamma_{x^k}(t)) = x^k(p) + t$, by our definition of γ_{x^k} in Definition 10.4.1, to obtain $\dfrac{d}{dt}x^j(\gamma_{x^k}(t))\Big|_{t=0} = \delta^j_k$, so that

$$\xi_{x^k}(f) = \frac{\partial}{\partial x^k}(f\circ\mathbf{x}^{-1})\Big|_{\mathbf{x}(p)}. \tag{10.4.2}$$

Thus, equation (10.4.1) follows from the fact that if α is any curve representing v, then by formula (10.3.3),

$$v(f) = \sum_{k=1}^n (x^k\circ\alpha)'(0)\frac{\partial}{\partial x^k}(f\circ\mathbf{x}^{-1})\Big|_{\mathbf{x}(p)} = \sum_{k=1}^n (x^k\circ\alpha)'(0)\xi_{x^k}(f).$$

Thus we can take $a^k = (x^k\circ\alpha)'(0)$ for any representative curve $\alpha \in v$ to obtain (10.4.1).

Finally we need to show that $\{\xi_{x^k}\}$ are linearly independent. So suppose we have $\sum_{k=1}^n a^k\xi_{x^k} = 0$ for some numbers $\{a^k\}$. Now we apply this vector to the function x^j, for some j; by (10.4.2), we have

$$\sum_{k=1}^n a^k \xi_{x^k}(x^j) = \sum_{k=1}^n a^k\delta^k_j = a^j = 0.$$

Since j was arbitrary, we see that all $\{a^k\}$ are zero, so that $\{\xi_{x^k}\}$ are linearly independent. Thus we have an n-element basis coming from any coordinate system. □

Now we come to the question of how to change coordinates. Because of our definition of the basis vectors ξ_{x^k}, the formula is fairly simple.

Proposition 10.4.3. *Suppose $(\mathbf{x}, U)$ and $(\mathbf{y}, U)$ are two coordinate systems on the same open set $U \subset M$ containing a point p. Then the basis vectors ξ_{x^k} and ξ_{y^k} are related by the formulas*

$$\xi_{x^k} = \sum_{j=1}^{n} \frac{\partial y^j}{\partial x^k}\Big|_{\mathbf{x}(p)} \xi_{y^j} \qquad \text{and} \qquad \xi_{y^k} = \sum_{j=1}^{n} \frac{\partial x^j}{\partial y^k}\Big|_{\mathbf{y}(p)} \xi_{x^j}. \tag{10.4.3}$$

Here $\dfrac{\partial y^j}{\partial x^k}$ stands for $\dfrac{\partial}{\partial x^k}(y^j \circ \mathbf{x}^{-1})$, the partials of the transition function.

In addition, if a vector v is expressed in two ways as

$$v = \sum_{k=1}^{n} a^k \xi_{x^k} = \sum_{k=1}^{n} b^k \xi_{y^k},$$

then the components a^k and b^k are related by the formulas

$$a^k = \sum_{j=1}^{n} \frac{\partial x^k}{\partial y^j}\Big|_{\mathbf{y}(p)} b^j \qquad \text{and} \qquad b^k = \sum_{j=1}^{n} \frac{\partial y^k}{\partial x^j}\Big|_{\mathbf{x}(p)} a^j. \tag{10.4.4}$$

Proof. By Lemma 10.3.4, vectors are completely determined by their operation on smooth functions. As a result, we just have to compute, for any smooth f,

$$\begin{aligned}
\xi_{x^k}(f) &= \frac{\partial}{\partial x^k}\Big|_{\mathbf{x}(p)} (f \circ \mathbf{x}^{-1}) \\
&= \frac{\partial}{\partial x^k}\Big|_{\mathbf{x}(p)} \Big(f \circ \mathbf{y}^{-1}) \circ (\mathbf{y} \circ \mathbf{x}^{-1})\Big) \\
&= \sum_{j=1}^{n} \frac{\partial}{\partial y^j}(f \circ \mathbf{y}^{-1})\Big|_{\mathbf{y}(p)} \frac{\partial y^j}{\partial x^k}\Big|_{\mathbf{x}(p)} \\
&= \sum_{j=1}^{n} \xi_{y^j}(f) \frac{\partial y^j}{\partial x^k}\Big|_{\mathbf{x}(p)}.
\end{aligned}$$

The first part of equation (10.4.3) follows, since the function f was arbitrary, and the second is of course proved by the same technique in reverse.

The change-of-components formula follows from the change-of-basis formula exactly as in a general vector space, in equation (3.1.2) at the beginning of Section 3.1. Notice that components transform in the opposite way as basis vectors just as we saw then. □

Proposition 10.4.3 formed the basis of what is now called "classical differential geometry," as invented by Ricci and Levi-Civita in the late 1800s. One defined vectors by their components in some coordinate system, then checked that the components in different systems satisfied the transition formulas (10.4.4). Some mathematicians still do this, but most differential geometers now prefer the derivative-operator approach, since it helps explain *why* the transition formulas are valid. This more modern approach was devised in the mid-1900s, once the notions of topological curves and functions, along with the notion of an abstract equivalence class, were understood. The main conceptual difference is that in the classical approach, one defines everything by coordinates but then checks that the objects obtained are actually

independent of coordinates; in the modern approach, one defines objects abstractly by their operation on more basic objects (such as functions), then checks that the operations are independent of the choice of basic object. The two approaches differ more philosophically than mathematically.

It would be difficult to remember the formulas in Proposition 10.4.3 with this notation. Henceforth we are going to abandon the ξ_{x^k} notation for another, which will help us remember the formulas in the same way the Leibniz notation for calculus was invented to make the formulas seem obvious. The motivation for this comes from the formula (10.4.2).

Definition 10.4.4. From now on, we will write the vector basis ξ_{x^k} at $p \in M$ as $\xi_{x^k} = \dfrac{\partial}{\partial x^k}\Big|_p$. Thus general vectors in T_pM will be written as

$$v = \sum_{k=1}^{n} a^k \frac{\partial}{\partial x^k}\Big|_p,$$

and the corresponding operation on functions will be

$$v(f) = \sum_{k=1}^{n} a^k \frac{\partial}{\partial x^k}\Big|_p (f) = \sum_{k=1}^{n} a^k \frac{\partial}{\partial x^k}(f \circ \mathbf{x}^{-1})\Big|_{\mathbf{x}(p)}.$$

The derivative of a curve $\gamma\colon (-\varepsilon, \varepsilon) \to M$ can be written as

$$\frac{d}{dt}\gamma(t)\Big|_{t=0} = \sum_{k=1}^{n} \frac{d}{dt} x^k\big(\gamma(t)\big) \frac{\partial}{\partial x^k}\Big|_{\gamma(0)} \tag{10.4.5}$$

Observe that this definition serves as a mnemonic for the formulas in Proposition 10.4.3. The following Corollary gives the coordinate change formula in the form we will find most useful. Notice that once we decide to write vectors as partial derivatives, the coordinate change formula is an obvious consequence of the Chain Rule.

Corollary 10.4.5. *If $(\mathbf{x}, U)$ and $(\mathbf{y}, V)$ are two coordinate charts with $p \in U \cap V$, then the two vector bases at p are related by*

$$\frac{\partial}{\partial x^k}\Big|_p = \sum_{j=1}^{n} \frac{\partial y^j}{\partial x^k}\Big|_{\mathbf{x}(p)} \frac{\partial}{\partial y^j}\Big|_p. \tag{10.4.6}$$

There are several features of this formula that will be very important to us:

- The index k appears on the "bottom" in both equations. In general we will see that formulas representing coordinate-invariant objects must always have this property.
- The index j being summed over appears once on top and once on the bottom. This is natural here because it's a consequence of the Chain Rule; in general we will place the indices in such a way that all sums will have an index which appears once on "top" and once on the "bottom." The formula $v = \sum_{k=1}^{n} a^k \frac{\partial}{\partial x^k}$ is another example of this, and explains why we are using superscripts instead of subscripts. The idea is that it helps you remember formulas; any formula that involves summing over two lower indices must not be coordinate-invariant.

- The basis-change formula implies the component-change formula. We have
$$v = \sum_{k=1}^{n} a^k \frac{\partial}{\partial x^k} = \sum_{k=1}^{n} b^k \frac{\partial}{\partial y^k} = \sum_{k=1}^{n}\sum_{j=1}^{n} b^k \frac{\partial x^j}{\partial y^k}\frac{\partial}{\partial x^j}.$$
Components change in the *opposite* way that the basis vectors do, which needs to happen to get coordinate invariance. We have consistency because
$$\begin{aligned}\sum_{k=1}^{n} b^k \frac{\partial}{\partial y^k} &= \sum_{k=1}^{n}\sum_{j=1}^{n} a^j \frac{\partial y^k}{\partial x^j}\sum_{i=1}^{n}\frac{\partial x^i}{\partial y^k}\frac{\partial}{\partial x^i} \\ &= \sum_{j=1}^{n} a^j \left(\sum_{k=1}^{n}\sum_{i=1}^{n}\frac{\partial y^k}{\partial x^j}\frac{\partial x^i}{\partial y^k}\frac{\partial}{\partial x^i}\right).\end{aligned}$$
Now notice that, by the Chain Rule, we have
$$\sum_{k=1}^{n}\frac{\partial y^k}{\partial x^j}\frac{\partial x^i}{\partial y^k} = \frac{\partial x^i}{\partial x^j} = \delta^i_j.$$
As a result the formula above reduces to
$$\sum_{k=1}^{n} b^k \frac{\partial}{\partial y^k} = \sum_{j=1}^{n} a^j \sum_{i=1}^{n}\delta^i_j \frac{\partial}{\partial x^i} = \sum_{j=1}^{n} a^j \frac{\partial}{\partial x^j}.$$
Again, notice how in the above computations all indices summed over appear twice: once above, and once below.

Because of these properties, we will always write vectors in the partial-derivative notation, even when we are not planning to actually differentiate any functions. We should keep in mind, however, that a vector actually exists *independently* of any coordinate system: its formula in a particular system is just a single element of an equivalence class.

As a familiar application, we transform rectangular vectors into polar vectors, using the transition functions $(x, y) = (r\cos\theta, r\sin\theta)$:

$$\frac{\partial}{\partial r} = \frac{\partial x}{\partial r}\frac{\partial}{\partial x} + \frac{\partial y}{\partial r}\frac{\partial}{\partial y} \tag{10.4.7}$$

$$\frac{\partial}{\partial \theta} = \frac{\partial x}{\partial \theta}\frac{\partial}{\partial x} + \frac{\partial y}{\partial \theta}\frac{\partial}{\partial y} \tag{10.4.8}$$

In terms of our abandoned notation ξ_r and ξ_θ, this gives the formulas $\xi_r = \cos\theta\,\xi_x + \sin\theta\,\xi_y$ and $\xi_\theta = -r\sin\theta\,\xi_x + r\cos\theta\,\xi_y$. Since $\xi_x = e_x$ and $\xi_y = e_y$ in the standard vector calculus basis, and these are orthonormal in the Euclidean norm, we have $|\xi_r| = 1$ and $|\xi_\theta| = r$, so that we can get orthonormal polar vectors $e_r = \xi_r$ and $e_\theta = r^{-1}\xi_\theta$. (Observe that we get the same formulas as we started the section with; the difference is that we didn't need any diagrams to figure out the formulas.) In general we don't have an inner product on T_pM, and even if we did our coordinate vectors would not necessarily be orthogonal, so it doesn't make sense to ask for an orthonormal basis associated to a coordinate system. Thus the difficulty in finding transition formulas for vector calculus in different coordinate charts is entirely due to the fact that we thought orthonormal bases were the correct way to think about things.

11. Derivatives

"If one is to understand the great mystery, one must study all its aspects."

11.1. **Derivatives as operators on tangent spaces.** This new notion of vector means that we need to redefine the notion of derivative of maps. Suppose we have some smooth map $F\colon M \to N$, where M and N are smooth manifolds. What do we mean by the derivative of F?

In standard vector calculus, we think of the derivative of a multivariable function as being a matrix: for example, when we have $F\colon \mathbb{R}^3 \to \mathbb{R}^2$ given by $(u,v) = \big(f(x,y,z), g(x,y,z)\big)$, then the derivative of $F = (f,g)$ is the 2×3 matrix

$$DF = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} & \frac{\partial u}{\partial z} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} & \frac{\partial v}{\partial z} \end{pmatrix}.$$

In conjunction with our new thinking about tangent vectors as abstract elements of certain linear spaces, we want to think about derivatives as linear operators from one such space to another. First recall the definition of smooth maps from one manifold to another in Definition 9.1.1, which sets up our context.

We assume in the definition of smoothness that F has infinitely many partial derivatives in coordinates even though for right now we only want a single derivative map. We will be able to obtain the higher derivatives later on as naturally more complicated objects, and in so doing we will see why for example acceleration is more complicated than velocity. Even in multivariable calculus, it is already clear that the correct notion of derivative is as a linear map from one Euclidean space to another (rather than the set of partial derivatives). The correct analogue for maps of manifolds is a linear map from one tangent space to another. For each $p \in M$, we are looking for a map $F_*\colon T_pM \to T_{F(p)}N$. When the point p may be changing, as in the next Chapter, we denote this map by $(F_*)_p$. Let's figure out what it has to look like.

Since vectors are basically equivalence classes of curves, we just need to consider how a smooth map F changes one class of curves to another. Clearly, if $\gamma\colon (-\varepsilon,\varepsilon) \to M$ is a smooth curve with $\gamma(0) = p$, then $F\circ\gamma\colon (-\varepsilon,\varepsilon) \to N$ is a smooth curve with $(F\circ\gamma)(0) = F(p)$. For this to make sense, we need to check that if α and β are curves in M through p that have the same derivative at $t=0$, then $F\circ\alpha$ and $F\circ\beta$ are curves in N through $F(p)$ that have the same derivative. The following Proposition accomplishes this.

Proposition 11.1.1. *Let M and N be smooth manifolds, and suppose that $F\colon M \to N$ is smooth. Let α and β be two curves from $(-\varepsilon,\varepsilon)$ to M, with $\alpha(0) = \beta(0) = p$, which have the same derivative (in the sense of Definition 10.3.2). Then the curves $F\circ\alpha$ and $F\circ\beta$ also have the same derivative. Thus there is a map $F_*\colon T_pM \to T_{F(p)}N$ defined so that F_*v is the equivalence class of $F\circ\gamma$ whenever γ is a representative of v.*

Equivalently we may specify F_*v by its action on smooth functions h defined in a neighborhood of $F(p)$: we have

$$(F_*v)(h) = v(h\circ F). \tag{11.1.1}$$

Hence in particular F_* is linear.

Proof. Suppose α and β have the same derivative at $t = 0$ in the sense of Definition 10.3.2; we want to show that $F \circ \alpha$ and $F \circ \beta$ do as well. By definition, we just need to check that for any smooth function $h\colon \Omega \subset N \to \mathbb{R}$ with $F(p) \in \Omega$, we have

$$\frac{d}{dt}(h \circ F \circ \alpha)\Big|_{t=0} = \frac{d}{dt}(h \circ F \circ \beta)\Big|_{t=0}. \tag{11.1.2}$$

Now $f = h \circ F$ is a smooth function by the Chain Rule 5.1.10: using the definition of smoothness we can write $f \circ \mathbf{x}^{-1} = (h \circ \mathbf{y}^{-1}) \circ (\mathbf{y} \circ F \circ \mathbf{x}^{-1})$ which is the composition of smooth functions on Euclidean spaces. We have $f\colon F^{-1}[\Omega] \subset M \to \mathbb{R}$ and $p \in F^{-1}[\Omega]$, so (11.1.2) follows from Definition 10.3.2.

The formula (11.1.1) comes, as all our formulas have, from the Chain Rule. Let h be any smooth function defined in a neighborhood of $F(p)$. Choose any curve γ representing v, and let $\delta = F \circ \gamma$; then $\delta \in F_* v$ by definition. Then, again by chasing through the definitions, we have

$$(F_* v)(h) = \frac{d}{dt}\Big|_{t=0}(h \circ \delta)(t) = \frac{d}{dt}\Big|_{t=0}(h \circ F \circ \gamma)(t) = v(h \circ F).$$

Thus we have for any $v, w \in T_pM$ and $a, b \in \mathbb{R}$ that

$$(F_*(av+bw))(h) = (av+bw)(h\circ F) = av(h\circ F)+bw(h\circ F) = a(F_*v)(h)+b(F_*w)(h).$$

Since this is true for every function h, we conclude $F_*(av + bw) = aF_*v + bF_*w$ by Lemma 10.3.4, so that F_* is actually a linear operator. □

If you keep track of domains and ranges of the various objects (for example by drawing a diagram), formula (11.1.1) is the only possible formula that could make sense. F_*v is a vector at $F(p)$ in N, so the only smooth functions it could operate on are real-valued functions h defined on N. If we want to relate it to v, a vector at p in M, we need to get a real-valued function from M, which must therefore be $h \circ F$.

Example 11.1.2. Here is an explicit example. Suppose we are thinking of M and N (which happen to be the same space) as $\mathbb{C}$. Let $F\colon M \to N$ be the continuous function $F(z) = -iz^2/2$. Put a Cartesian coordinate system $\mathbf{x} = (x, y)$ on the domain and a Cartesian coordinate system $\mathbf{u} = (u, v)$ on the range. Then first of all, $\mathbf{u} \circ F \circ \mathbf{x}^{-1}\colon \mathbb{R}^2 \to \mathbb{R}^2$ is given by $(u, v) = \big(xy, \frac{1}{2}(y^2 - x^2)\big)$, and this is obviously smooth.

Consider a fixed point $z_o \in M$ with $w_o = F(z_o) \in N$, given in coordinates respectively by $\mathbf{x}(z_o) = (x_o, y_o)$ and $\mathbf{u}(w_o) = (u_o, v_o) = (x_o y_o, \frac{1}{2}(y_o^2 - x_o^2))$. An arbitrary vector $V \in T_{z_o}M$ can be expressed in the coordinate basis as $V = a\,\dfrac{\partial}{\partial x}\Big|_{z_o} + b\,\dfrac{\partial}{\partial y}\Big|_{z_o}$ for some $a, b \in \mathbb{R}$. Take $\gamma(t) = z_o + (a + ib)t$; then in coordinates we have $\mathbf{x} \circ \gamma(t) = (x_o + at, y_o + bt)$ so that $\gamma'(0) = v$.

Now $F \circ \gamma(t) = -i/2\big(z_o + (a + ib)t\big)^2$, which is given in coordinates by

$$\big(u(t), v(t)\big) = \mathbf{u}\circ F\circ\gamma(t) = \big(x_o y_o+ay_o t+bx_o t+abt^2, \tfrac{1}{2}y_o^2-\tfrac{1}{2}x_o^2+by_o t-ax_o t-\tfrac{1}{2}a^2t^2+\tfrac{1}{2}b^2\big);$$

thus the derivative at $t = 0$ is

$$(F\circ\gamma)'(0) = u'(0)\,\frac{\partial}{\partial u}\Big|_{w_o} + v'(0)\,\frac{\partial}{\partial v}\Big|_{(u_o,v_o)} = (ay_o+bx_o)\,\frac{\partial}{\partial u}\Big|_{w_o} + (by_o-ax_o)\,\frac{\partial}{\partial v}\Big|_{w_o}.$$

We have thus computed that

$$F_*\left(a\,\frac{\partial}{\partial x}\Big|_{z_o} + b\,\frac{\partial}{\partial y}\Big|_{z_o}\right) = (ay_o + bx_o)\,\frac{\partial}{\partial u}\Big|_{F(z_o)} + (by_o - ax_o)\,\frac{\partial}{\partial v}\Big|_{F(z_o)}.$$

As expected from Proposition 11.1.1, the map F_* is linear, and the rectangular coordinate bases we have

(11.1.3)
$$F_*\Big(\frac{\partial}{\partial x}\Big|_{z_o}\Big) = y_o\,\frac{\partial}{\partial u}\Big|_{F(z_o)} - x_o\,\frac{\partial}{\partial v}\Big|_{F(z_o)} \qquad \text{and} \qquad F_*\Big(\frac{\partial}{\partial y}\Big|_{z_o}\Big) = x_o\,\frac{\partial}{\partial u}\Big|_{F(z_o)} + y_o\,\frac{\partial}{\partial v}\Big|_{F(z_o)}.$$

Not coincidentally, in matrix form this is precisely what we'd get by viewing $F\colon \mathbb{R}^2 \to \mathbb{R}^2$ as the map $(u,v) = \big(xy, \frac{1}{2}(y^2 - x^2)\big)$ and computing

$$DF(x_o, y_o) = \begin{pmatrix} u_x(x_o,y_o) & v_x(x_o,y_o) \\ u_y(x_o,y_o) & v_y(x_o,y_o) \end{pmatrix}.$$

Also notice that even though the domain and range are the same space, we use different coordinates to represent them; it's what allows us to use the matrix $\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$ to compute F_*. If we used the same coordinate labels (x,y), it would be easy to get confused between both the partial derivatives and especially the locations of the vectors.

Now suppose we use a genuinely *different* coordinate system for the domain and range. Let $\mathbf{x} = (x,y)$ be rectangular coordinates on the domain, and let $\mathbf{s} = (\sigma, \tau)$ be parabolic coordinates on the range as in (6.2.4), which are related to the rectangular coordinates (u,v) on the range by $(u,v) = \big(\sigma\tau, \frac{1}{2}(\tau^2 - \sigma^2)\big)$. Then the map in coordinates, $\mathbf{s} \circ F \circ \mathbf{x}^{-1}$, is given by

$$\begin{aligned}(\sigma,\tau) &= \mathbf{s} \circ F \circ \mathbf{x}^{-1}(x,y) \\ &= \mathbf{s} \circ \mathbf{u}^{-1} \circ \mathbf{u} \circ F \circ \mathbf{x}^{-1}(x,y) \\ &= \mathbf{s} \circ \mathbf{u}^{-1}\big(xy, \frac{1}{2}(y^2 - x^2)\big) \\ &= (x,y).\end{aligned}$$

Thus in these coordinates, F looks like the identity map. Using the same technique as above, we can compute that

$$F_*\Big(\frac{\partial}{\partial x}\Big|_{z_o}\Big) = \frac{\partial}{\partial \sigma}\Big|_{F(z_o)} \qquad \text{and} \qquad F_*\Big(\frac{\partial}{\partial y}\Big|_{z_o}\Big) = \frac{\partial}{\partial \tau}\Big|_{F(z_o)}. \tag{11.1.4}$$

☺

Equation (11.1.4) is basically just an expression of the fact that a linear transformation from one vector space to another has only one invariant—the rank—and if we can change the basis of both the domain and range, we can make a maximal-rank linear transformation look like the identiy as in the discussion after Definition 3.2.2. In our context, where the linear transformation comes from the derivative of a smooth map and our basis vectors come from coordinate charts, we can think of this as the ability to choose coordinates on the domain and range (separately) so that any invertible map F actually looks locally like

$$\mathbf{s} \circ F \circ \mathbf{x}^{-1}(x^1, \dots, x^n) = (y^1, \dots, y^n). \tag{11.1.5}$$

In fact the Inverse Function Theorem essentially tells us that if we have chosen coordinates $\mathbf{x}$ and $\mathbf{u}$ on the domain and range in some way, and that if $\mathbf{u} \circ F \circ \mathbf{x}^{-1}$ has an invertible derivative somewhere, then $\mathbf{s} = \mathbf{x} \circ F^{-1} = (\mathbf{u} \circ F \circ \mathbf{x}^{-1})^{-1} \circ \mathbf{u}$ is also a coordinate chart on N in which F takes the trivial form (11.1.5); this is precisely what we did in the Example above.

This is the first instance of a general phenomenon: a map F from one manifold to another which is smooth and has smooth inverse can, for many purposes, be treated just as a coordinate change on a single manifold. We did something like this when discussing linear transformations: in trying to prove that the determinant was basis-independent in Proposition 3.3.4, we first proved $\det AB = \det A \det B$ and then used it in the case where A was a coordinate-change matrix and B was a matrix expressing a linear transformation (genuinely different types of objects that happen to both look like matrices), and we did it again in order to make sense of the "nonzero determinant" condition for a linear transformation from one vector space to another to be invariant. The consequence for manifolds is that objects which are coordinate-invariant also end up being invariant under smooth maps and well-behaved in the category theory sense. This will be most obvious when we work with differential forms, and is hugely important for things like deRham cohomology.

Returning to the situation of Example 11.1.2, let's view F as a map from M to itself and consider the special case of a fixed point $F(z_o) = z_o$.

Example 11.1.3. If $F(z) = -iz^2/2$, then F has two fixed points at $z_o = 0$ and $z_o = 2i$. In either case, F_* is a linear map from $T_{z_o}M$ to itself. Notice that even if F is a smooth map from M to itself, it's still convenient for actual computations to separate the domain and range and use different names for the coordinates, as we did in Example 11.1.2. For example if I have coordinates (x, y) on both the domain and range, I will frequently use (x, y) as the domain coordinates and (u, v) as the range coordinates, and only when I'm done with the computation will I then put (u, v) back in terms of (x, y). This is especially true for the type of computation I'm doing here at a fixed point of the map.

When $z_o = 0$ we have $(x_o, y_o) = (0, 0)$, and the map $F_*\colon T_0M \to T_0M$ is identically zero by formula (11.1.3). Since it is zero in the rectangular coordinate basis, it is zero in *every* coordinate basis: the map F_* is a linear map defined independently of any coordinates, and so it either sends all vectors to zero or not regardless of what basis we are working in. Notice that since F_* is not invertible at the origin, we cannot get any coordinate chart at the origin which makes F look like the identity as in (11.1.5).

On the other hand when $z_o = 2i$ we have $x_o = 0$ and $y_o = 2$, so that $F_*\colon T_{2i}M \to T_{2i}M$ is given according to (11.1.3) by

$$F_*\left(\frac{\partial}{\partial x}\Big|_{2i}\right) = 2\,\frac{\partial}{\partial x}\Big|_{2i} \qquad \text{and} \qquad F_*\left(\frac{\partial}{\partial y}\Big|_{2i}\right) = 2\,\frac{\partial}{\partial y}\Big|_{2i}.$$

In other words, F_* is twice the identity, and hence it will be twice the identity in any other coordinate chart as well: here the number 2 actually means something. ☺

We now come to a very simple but extremely useful result: the Chain Rule. As an indication of the utility of the definition F_*, we find that the Chain Rule on manifolds is actually simpler than the Chain Rule 5.1.10 for multivariable functions (although as always the multivariable calculus is really behind our elegant definitions).

Theorem 11.1.4. *Suppose L, M, and N, are smooth manifolds of dimensions ℓ, m, and n respectively. Let $F\colon L \to M$ and $G\colon M \to N$ be smooth maps, and let $H\colon L \to N$ be the smooth map $H = G \circ F$. Let $p \in L$, and define $q = F(p) \in M$*

and $r = G(q) \in N$. Then the linear maps $F_*\colon T_pL \to T_qM$ and $G_*\colon T_qM \to T_rN$ and $H_*\colon T_pL \to T_rN$ are related by $H_* = G_* \circ F_*$.

Proof. Let $v \in T_pL$ be an arbitrary vector, and let $h\colon \Omega \subset N \to \mathbb{R}$ be an arbitrary function defined in a neighborhood Ω of $r \in N$. Then for any representative curve $\gamma\colon (-\varepsilon, \varepsilon) \to L$ with $\gamma(0) = p$ and $\gamma'(0) = v$, we have by Definition 11.1.1 that $(H_*v)(h) = \dfrac{d}{dt}(h \circ H \circ \gamma)(t)\Big|_{t=0}$. On the other hand we have

$$h \circ H \circ \gamma(t) = (h \circ G) \circ (F \circ \gamma)(t).$$

If $w = F_*v \in T_qM$ then $\beta = F \circ \gamma$ is a representative curve of w, and by we have

$$(H_*v)(h) = \frac{d}{dt}\Big((h \circ G) \circ (F \circ \gamma)\Big)(t)\Big|_{t=0} = \frac{d}{dt}(h \circ G \circ \beta)(t)\Big|_{t=0} = (G_*w)(h).$$

Since h was an arbitrary function, we conclude $H_*v = G_*w$ whenever $w = F_*v$. And since v was arbitrary we conclude $H_* = G_* \circ F_*$. □

What's nice about the formula $(G \circ F)_* = G_* \circ F_*$ is that the star operation keeps track of base points in the correct way automatically. Even the usual one-dimensional version of the Chain Rule $(G \circ F)'(t) = G'(F(t))F'(t)$ is more complicated to write. Ironically, the differential geometry approach is more useful precisely because one is allowed to do less with it; the fact that points are different from vectors prevents a lot of confusion, and the fact that tangent spaces at different points cannot be naturally identified helps us remember where all our operations must be taking place.

11.2. **A coordinate approach to derivatives.** We have defined F_* in a coordinate-independent way, but as in Example 11.1.2, it is usually convenient to compute in given coordinate bases on T_pM and $T_{F(p)}N$. Here we demonstrate the general formula. We also make our first attempt to relate the modern approach to the classical approach.

Proposition 11.2.1. *Let M and N be smooth manifolds of respective dimensions m and n. Suppose we have a coordinate chart $(\mathbf{x}, U)$ on M and $(\mathbf{u}, W)$ on N and a smooth map $F\colon M \to N$ with $p \in U$ and $F(p) \in W$. Then the linear map $F_*\colon T_pM \to T_{F(p)}N$ is given in the coordinate bases by*

$$F_*\left(\frac{\partial}{\partial x^j}\Big|_p\right) = \sum_{k=1}^{n} \frac{\partial}{\partial x^j}\big(u^k \circ F \circ \mathbf{x}^{-1}\big)\Big|_{\mathbf{x}(p)} \frac{\partial}{\partial u^k}\Big|_{F(p)}, \qquad 1 \le j \le m. \tag{11.2.1}$$

Proof. Let $(a^1, \ldots, a^m) = \mathbf{x}(p)$. For each $j \in \{1, \ldots, m\}$, consider the "coordinate curve" γ_k given by

$$\gamma_j(t) = \mathbf{x}^{-1}(a^1, \ldots, a^j + t, \ldots, a^m)$$

and defined on some small neighborhood $(-\varepsilon, \varepsilon)$ of $t = 0$. Then $\gamma_j(0) = p$, and since $\mathbf{x} \circ \gamma_j(t) = (a^1, \ldots, a^j + t, \ldots, a^m)$ we know that γ_j is in the equivalence class $\dfrac{\partial}{\partial x^j}\Big|_p$ by Definition 10.4.1 (after the renaming by Definition 10.4.4) of the coordinate basis vectors. For each $k \in \{1, \ldots, n\}$, define functions $g_j^k\colon (-\varepsilon, \varepsilon) \to \mathbb{R}$ by $g_j^k = u^k \circ F \circ \gamma_j$ where $\mathbf{u} = (u^1, \ldots, u^n)$ are the components of the coordinate chart on N. Then the curve $F \circ \gamma_j\colon (-\varepsilon, \varepsilon) \to N$ in coordinates takes the form

$$\mathbf{u} \circ F \circ \gamma_j(t) = \big(g_j^1(t), \ldots, g_j^n(t)\big).$$

Thus by formula (10.4.5), we have

$$F_*\left(\frac{\partial}{\partial x^j}\Big|_p\right) = (F\circ\gamma_j)'(0) = \sum_{k=1}^{n} \frac{dg_j^k}{dt}\Big|_{t=0} \frac{\partial}{\partial u^k}\Big|_{F(p)}.$$

Clearly we have

$$\frac{dg_j^k}{dt}\Big|_{t=0} = \frac{\partial}{\partial x^j}(u^k\circ F\circ \mathbf{x}^{-1})\Big|_{(a^1,\dots,a^n)},$$

which gives formula (11.2.1). □

Of course if $F\colon \mathbb{R}^m\to\mathbb{R}^n$, we would just write $\mathbf{u}=F(\mathbf{x})$, and the derivative map F_* would just be DF as the matrix with components $\frac{\partial u^j}{\partial x^k}$. Obviously this looks simpler once we are working in Euclidean coordinates, but it makes coordinate changes very difficult.

Example 11.2.2. For example suppose we have a simple map from $\mathbb{R}^2$ to itself which is naturally expressed in polar coordinates as $F(r,\theta)=(r^4,2\theta)$, and I want to compute $DF(e_r)$ and $DF(e_\theta)$ in the orthonormal polar basis. The classical approach already has some difficulty here, and people unfamiliar with differential geometry are often tempted to do things like convert everything to Cartesian coordinates, take the derivatives, and then convert back (often getting the wrong answer along the way).

In the differential geometric approach, the computation is trivial: we would write

$$F_*\left(\frac{\partial}{\partial r}\Big|_p\right) = 4r^3\,\frac{\partial}{\partial r}\Big|_{F(p)} \qquad\text{and}\qquad F_*\left(\frac{\partial}{\partial \theta}\Big|_p\right) = 2\,\frac{\partial}{\partial \theta}\Big|_{F(p)}.$$

Of course in the classical approach the "correct vectors" are e_r and e_θ, which satisfy $e_r|_p = \frac{\partial}{\partial r}\big|_p$ and $e_\theta|_p = \frac{1}{r(p)}\,\frac{\partial}{\partial\theta}\big|_p$ where $r(p)$ is the radius of point p, and thus in this basis

$$F_*(e_r|_p) = 4r^3 e_r|_{F(p)} \qquad\text{and}\qquad F_*(e_\theta|_p) = 2\,\frac{r^4}{r} e_\theta|_{F(p)},$$

or in the usual vector calculus notation where you forget where your vectors live, $F_*(e_r)=4r^3e_r$ and $F_*(e_\theta)=2r^3e_\theta$, and now that factor of r^3 in front of e_θ is mysterious.

These computations are manageable since polar coordinates are common enough that people just copy them from the back of a classical mechanics textbook, but clearly working in a nonstandard coordinate system in the classical approach is going to be extremely difficult and for no good reason. This is an example of a situation where understanding things from the geometric point of view is extremely helpful even if you're not doing geometry. ☺

We have defined F_* in a coordinate-independent way, and the coordinate formula (11.2.1) is a consequence. Thus if we had two different coordinate charts $\mathbf{x}$ and $\mathbf{y}$ on M, and two different charts $\mathbf{u}$ and $\mathbf{v}$ on N, then we would certainly have

$$F_*\left(\frac{\partial}{\partial y^j}\Big|_p\right) = \sum_{k=1}^{n} \frac{\partial}{\partial y^j}\big(v^k\circ F\circ \mathbf{y}^{-1}\big)\Big|_{\mathbf{y}(p)}\,\frac{\partial}{\partial v^k}\Big|_{F(p)}. \tag{11.2.2}$$

Now classically a vector would be *defined* in terms of its components in some particular coordinate chart, but it would only really represent an invariant object if we knew that its components in a different coordinate chart satisfied the formulas

in Corollary 10.4.5. As such, the map F_* would have been *defined* by the formula (11.2.1), and one would have needed to *prove* formula (11.2.2) from it using the transition formulas (10.4.6). In the next Proposition, we will show that this actually works. The proof is just the Chain Rule a few times, and if you're willing to believe it, you can skip the proof entirely and move on to the next Section. Later on we will need this technique when we start trying to differentiate vector fields or related objects, but you can wait until then if you want.

Proposition 11.2.3. *The linear map*

$$F_*\left(\sum_{j=1}^n a^j \frac{\partial}{\partial x^j}\right) = \sum_{j=1}^n \sum_{k=1}^n a^j \frac{\partial}{\partial x^j}(u^k \circ F \circ \mathbf{x}^{-1}) \frac{\partial}{\partial u^k} \tag{11.2.3}$$

is coordinate-invariant; in other words, if we define the operation F_* by (11.2.3) *in* $\mathbf{x}$- *and* $\mathbf{u}$-*coordinates on M and N, then in any other coordinate systems* $\mathbf{y}$ *on M and* $\mathbf{v}$ *on N, we will have*

$$F_*\left(\sum_{j=1}^n b^j \frac{\partial}{\partial y^j}\right) = \sum_{j=1}^n \sum_{k=1}^n b^j \frac{\partial}{\partial y^j}(v^k \circ F \circ \mathbf{y}^{-1}) \frac{\partial}{\partial v^k}.$$

Proof. We just apply Proposition (10.4.3) and the Chain Rule. So let us start with (11.2.3) and change coordinates from $\mathbf{x}$ to $\mathbf{y}$ on M (keeping coordinates $\mathbf{u}$ on N). Then for any index j,

$$\begin{aligned}
F_*\left(\frac{\partial}{\partial y^j}\right) &= F_*\left(\sum_{i=1}^n \frac{\partial x^i}{\partial y^j}\frac{\partial}{\partial x^i}\right) \\
&= \sum_{i=1}^n \frac{\partial x^i}{\partial y^j} F_*\left(\frac{\partial}{\partial x^i}\right) \qquad \text{(since } F_* \text{ is a linear map)} \\
&= \sum_{i=1}^n \frac{\partial x^i}{\partial y^j} \sum_{k=1}^n \frac{\partial}{\partial x^i}(u^k \circ F \circ \mathbf{x}^{-1}) \frac{\partial}{\partial u^k} \\
&= \sum_{i=1}^n \frac{\partial x^i}{\partial y^j} \sum_{k=1}^n \frac{\partial}{\partial x^i}\Big((u^k \circ F \circ \mathbf{y}^{-1}) \circ (\mathbf{y} \circ \mathbf{x}^{-1})\Big) \frac{\partial}{\partial u^k} \\
&= \sum_{k=1}^n \frac{\partial}{\partial y^j}(u^k \circ F \circ \mathbf{y}^{-1}) \frac{\partial}{\partial u^k},
\end{aligned}$$

using the Chain Rule $\dfrac{\partial}{\partial y^j}(g(y)) = \displaystyle\sum_{i=1}^n \frac{\partial}{\partial x^i}\big(g(\mathbf{y} \circ \mathbf{x}^{-1}(x))\big)\frac{\partial x^i}{\partial y^j}$.

Now we show that F_* does not depend on coordinates on the range:

$$\begin{aligned} F_*\left(\frac{\partial}{\partial y^j}\right) &= \sum_{k=1}^{n} \frac{\partial}{\partial y^j}(u^k \circ F \circ \mathbf{y}^{-1})\, \frac{\partial}{\partial u^k} \\ &= \sum_{k=1}^{n}\sum_{l=1}^{n} \frac{\partial}{\partial y^j}(u^k \circ F \circ \mathbf{y}^{-1})\, \frac{\partial v^l}{\partial u^k}\, \frac{\partial}{\partial v^l} \\ &= \sum_{l=1}^{n} \frac{\partial}{\partial y^j}(v^l \circ \mathbf{u}^{-1}) \circ (\mathbf{u} \circ F \circ \mathbf{y}^{-1})\, \frac{\partial}{\partial v^l} \qquad \text{(Chain Rule again)} \\ &= \sum_{l=1}^{n} \frac{\partial}{\partial y^j}(v^l \circ F \circ \mathbf{y}^{-1})\, \frac{\partial}{\partial v^l}, \end{aligned}$$

which is again the same formula as (11.2.3) with the **u**-coordinates replaced with **v**-coordinates. □

11.3. **The classical tangent space.** Recall that for surfaces in $\mathbb{R}^3$, we described in Section 10.1 the tangent space as a particular two-dimensional plane, a particular subspace of $\mathbb{R}^3$. Our actual definition of a tangent space is a rather abstract space which happens (in this case) to also be two-dimensional. Of course all two-dimensional tangent spaces are isomorphic to each other, but it's not obvious that there is any *natural* or basis-independent isomorphism. In this Section we will use the maps F_* to define the isomorphism.

First of all, note that $\mathbb{R}^n$ is a manifold, and so our new notion of tangent vector should coincide with the old notion in this case. Considering $M \cong \mathbb{R}^n$ as the abstract Euclidean space and $\mathbf{x}\colon M \to \mathbb{R}^n$ as the Cartesian coordinate chart, the space T_pM is spanned by the coordinate vectors $\left.\frac{\partial}{\partial x^k}\right|_p$, and the correspondence

$$\begin{pmatrix} a^1 \\ \vdots \\ a^n \end{pmatrix} \longleftrightarrow a^1 \left.\frac{\partial}{\partial x^1}\right|_p + \cdots + a^n \left.\frac{\partial}{\partial x^n}\right|_p$$

gives a natural isomorphism.

Now suppose that we have a surface in $\mathbb{R}^3$ defined locally as the inverse image $G^{-1}\{r\}$ for a function $F\colon \mathbb{R}^3 \to \mathbb{R}$ where r is a regular value. Let $M = G^{-1}\{r\}$; we know that M is a smooth submanifold by Theorem 9.1.11. Let $\iota\colon M \to \mathbb{R}^3$ be the inclusion. By definition of submanifold, there is a coordinate chart $(\mathbf{x}, U)$ around any point of M on $\mathbb{R}^3$ such that $U \cap M = \phi^{-1}[\{x^1, x^2, 0\}]$. The restriction $\tilde{\phi} = \phi|_M = \phi \circ \iota$ is thus a coordinate chart defined on $U \cap M$. In these coordinates we have the formula

$$\phi \circ \iota \circ \tilde{\phi}^{-1}(x^1, x^2) = (x^1, x^2, 0), \tag{11.3.1}$$

which is obviously a C^∞ function from $\mathbb{R}^2$ to $\mathbb{R}^3$. This shows that ι is smooth in the sense of Definition 9.1.1. As a consequence we must have that $G \circ \mathbf{x}^{-1}(x^1, x^2, x^3) = x^3 + r$.

Now $G \circ \iota\colon M \to \mathbb{R}$ is constant since $M = G^{-1}\{r\}$. Thus for any curve $\gamma\colon (-\varepsilon, \varepsilon) \to M$, we know that $(G \circ \iota \circ \gamma)(t)$ is constant so that $(G \circ \iota)_*(v) = 0$ for every v in every tangent space T_pM. In particular $(G \circ \iota)_*$ is identically zero. By the Chain Rule for manifolds, Theorem 11.1.4, this implies that $G_* \circ \iota_* = 0$. Now

formula (11.3.1) shows that ι_* has maximal rank (in particular that the image of $\iota_*\colon T_pM \to T_p\mathbb{R}^3$ is a two-dimensional subspace of $T_p\mathbb{R}^3$). Furthermore the map $G_*\colon T_p\mathbb{R}^3 \to T_r\mathbb{R} \cong \mathbb{R}$ has maximal rank by assumption that $r \in \mathbb{R}$ is a regular value of G, and hence the kernel of G_* is two-dimensional. The only way the image of ι_* can be contained in the kernel of G_* is if they are actually equal. Hence we have

$$\iota_*[T_pM] = \ker(G_*|_{T_pM}),$$

and this shows that the historical description of T_pM was actually the image $\iota_*[T_pM]$.

Similarly suppose $F\colon \Omega \subset \mathbb{R}^2 \to \mathbb{R}^3$ locally defines a manifold M. Again let $\iota\colon M \to \mathbb{R}^3$ denote the inclusion, and define $\tilde{F}\colon \Omega \subset \mathbb{R}^2 \to M$ to be the restriction of F to its image M. Then obviously $\iota \circ \tilde{F} = F$. In submanifold coordinates we can check that ι and $\tilde{F}$ are smooth in the sense of manifolds. By the Chain Rule for manifolds, Theorem 11.1.4, we have $\iota_* \circ \tilde{F}_* = F_*$. Now by the usual assumptions, the immersion F_* has rank two as a map from $T_a\mathbb{R}^2 \cong \mathbb{R}^2$ to $T_{\iota(p)}\mathbb{R}^3 \cong \mathbb{R}^3$, if $p = F(a)$. Now $\tilde{F}_*[T_a\mathbb{R}^2]$ is a linear subspace of the two-dimensional space T_pM, and since $\iota_*\big[\tilde{F}_*[T_a\mathbb{R}^2]\big] = F_*[T_a\mathbb{R}^2]$ is two-dimensional, we conclude that $\tilde{F}_*[T_a\mathbb{R}^2] = T_pM$. Thus we have $\iota_*[T_pM] = F_*[T_a\mathbb{R}^2]$, which again shows that the image under the inclusion ι_* of T_pM is what mathematicians historically thought of as the tangent space.

12. The tangent bundle

"Traveling through hyperspace ain't like dusting crops, boy!"

In the past two Chapters we have thought of tangent vectors as being defined at a single point $p \in M$. We have a vector space structure on each tangent space T_pM, but we must think of the space T_pM and T_qM as different and not related in any natural way. However we still want to fit these tangent spaces together and try to get another smooth manifold. For example, if $\gamma\colon \mathbb{R} \to M$ is a smooth curve, then $\gamma'(t)$ is a tangent vector in $T_{\gamma(t)}M$ for each t. Since γ is C^∞ in coordinate charts, we want to think of γ' as being at least a continuous and preferably a smooth curve in some other space. Classically we would think of a particle in $\mathbb{R}^n$ as having a position and a velocity, and keep track of both x and v; the pair (x, v) then lives in the *phase space* $\mathbb{R}^n \times \mathbb{R}^n$. This is natural since Newton's equations of motion then become a first-order differential equation on $\mathbb{R}^n \times \mathbb{R}^n$.

12.1. **Examples.** The *tangent bundle* TM is the disjoint union of all the tangent planes:

$$TM = \bigcup_{p \in M} T_pM.$$

However to do anything useful on it, we need to put a topology on it. The easiest way to do this is to define the manifold structure on it. Let's do some examples first to figure out what's important.

Example 12.1.1. Observe that if M is homeomorphic to $\mathbb{R}^n$, then we expect TM to be homeomorphic to $\mathbb{R}^{2n}$. We just take a global coordinate chart $\phi = \mathbf{x}$ on M, get at each $p \in M$ a basis of vectors $\frac{\partial}{\partial x^k}\big|_p$, and then say that our coordinates on TM are given by the following rule: if $v \in T_pM$ is expressed as $v = \sum_{i=1}^{n} a^i \frac{\partial}{\partial x^i}\Big|_p$ where $\mathbf{x}(p) = (q^1, \cdots, q^n)$, then the coordinate chart Φ on TM will be given by[23]

$$\Phi(v) = (q^1, \cdots, q^n, a^1, \cdots, a^n).$$

Since ϕ is a globally-defined chart, every p and every $v \in T_pM$ has a unique representation in this way, and conversely given the coordinates $(q^1, \cdots, q^n, a^1, \cdots, a^n)$, we set $p = \mathbf{x}^{-1}(q^1, \ldots, q^n)$ and $v \in T_pM$ to be $v = \sum_i a^i \frac{\partial}{\partial x^i}\big|_p$.

Let's see what happens in a different coordinate system such as polar coordinates on $M \cong \mathbb{R}^2$. Here the natural vector basis to use is the coordinate basis $\{\frac{\partial}{\partial r}, \frac{\partial}{\partial \theta}\}$. For any vector v in any tangent space T_pM, we can write $\phi(p) = (x, y)$ and $v = a\,\frac{\partial}{\partial x}\big|_p + b\,\frac{\partial}{\partial y}\big|_p$ for some numbers $\Phi(v) = (x, y, a, b)$ which give us the Cartesian coordinates of $v \in TM$. The same point p and vector $v \in T_pM$ may be written as $\psi(p) = (r, \theta)$ and $v = c\,\frac{\partial}{\partial r}\big|_p + d\,\frac{\partial}{\partial \theta}\big|_p$, to obtain the polar coordinates $\Phi(v) = (r, \theta, c, d)$ of $v \in TM$.

Of course we know the transition map $\phi \circ \psi^{-1}$: we have $(x, y) = \phi \circ \psi^{-1}(r, \theta) = (r\cos\theta, r\sin\theta)$, and we can solve for r and θ in the standard way. But what is the

[23]Notice that Φ is a function of $v \in TM$: every v is in some T_pM and no vector can be in two different tangent spaces, which means we always know what p is as soon as we know what v is. Hence we should *not* write $\Phi(p, v)$ because that makes it look like TM is actually $M \times \mathbb{R}^n$.

transition map $\Phi \circ \Psi^{-1}$? To get it, we need to use the coordinate vector transition formula (10.4.6) to obtain

$$\frac{\partial}{\partial r}\Big|_p = \cos\theta\,\frac{\partial}{\partial x}\Big|_p + \sin\theta\,\frac{\partial}{\partial y}\Big|_p \qquad \text{and} \qquad \frac{\partial}{\partial \theta}\Big|_p = -r\sin\theta\,\frac{\partial}{\partial x}\Big|_p + r\cos\theta\,\frac{\partial}{\partial y}\Big|_p$$

in order to compare the two expressions of v. We have

$$\begin{aligned} v = c\,\frac{\partial}{\partial r}\Big|_p + d\,\frac{\partial}{\partial \theta}\Big|_p &= \big(c\cos\theta - dr\sin\theta\big)\,\frac{\partial}{\partial x}\Big|_p + \big(c\sin\theta + dr\cos\theta\big)\,\frac{\partial}{\partial y}\Big|_p \\ &= a\,\frac{\partial}{\partial x}\Big|_p + b\,\frac{\partial}{\partial y}\Big|_p, \end{aligned}$$

which implies $a = c\cos\theta - dr\sin\theta$ and $b = c\sin\theta + dr\cos\theta$. Putting it all together, we get the transition formula

$$(x, y, a, b) = \Phi \circ \Psi^{-1}(r, \theta, c, d) = \big(r\cos\theta, r\sin\theta, c\cos\theta - dr\sin\theta, c\sin\theta + dr\cos\theta\big).$$

Obviously this is C^∞, but more important is that it is *linear* in the velocity components (c, d), though of course quite nonlinear in the position components (r, θ). ☺

The fact that $\mathbb{R}^n$ can be covered by a single coordinate chart means that the tangent bundle $T\mathbb{R}^n$ can as well, and this implies that $T\mathbb{R}^n$ must be trivial. Still we see that coordinate changes twist the tangent spaces in various ways. Now let's see what happens in the first nontrivial case, when $M = \mathbb{S}^2$.

Example 12.1.2. We cover S^2 with north-pole and south-pole stereographic coordinate charts as in Example 8.2.15, given by the formulas

$$(u, v) = \phi(x, y, z) = \left(\frac{x}{1-z}, \frac{y}{1-z}\right) \quad \text{and} \quad (s, t) = \psi(x, y, z) = \left(\frac{x}{1+z}, \frac{y}{1+z}\right),$$

where the inverses are given by

$$\begin{aligned} (x, y, z) = \phi^{-1}(u, v) &= \left(\frac{2u}{1+u^2+v^2}, \frac{2v}{1+u^2+v^2}, \frac{u^2+v^2-1}{u^2+v^2+1}\right) \\ = \psi^{-1}(s, t) &= \left(\frac{2s}{1+s^2+t^2}, \frac{2t}{1+s^2+t^2}, \frac{1-s^2-t^2}{1+s^2+t^2}\right). \end{aligned}$$

Here ϕ covers all but $(0, 0, 1)$ and ψ covers all but $(0, 0, -1)$. The transition map is

$$(s, t) = \psi \circ \phi^{-1}(u, v) = \left(\frac{u}{u^2+v^2}, \frac{v}{u^2+v^2}\right)$$

which is its own inverse, so that $u = s/(s^2+t^2)$ and $v = t/(s^2+t^2)$.

Using (10.4.6), we find that

$$\begin{aligned} \frac{\partial}{\partial s}\Big|_p &= \frac{v^2-u^2}{(u^2+v^2)^2}\,\frac{\partial}{\partial u}\Big|_p - \frac{2uv}{(u^2+v^2)^2}\,\frac{\partial}{\partial v}\Big|_p \qquad \text{and} \\ \frac{\partial}{\partial t}\Big|_p &= -\frac{2uv}{(u^2+v^2)^2}\,\frac{\partial}{\partial u}\Big|_p + \frac{u^2-v^2}{(u^2+v^2)^2}\,\frac{\partial}{\partial v}\Big|_p. \end{aligned}$$

Hence if a vector $V \in T_pM$ is written in two ways as $V = a\,\frac{\partial}{\partial s}\big|_p + b\,\frac{\partial}{\partial t}\big|_p$ and as $V = c\,\frac{\partial}{\partial u}\big|_p + d\,\frac{\partial}{\partial v}\big|_p$, so that $\Psi(V) = (s, t, a, b)$ and $\Phi(V) = (u, v, c, d)$, then

$$c = \frac{a(v^2-u^2) - 2buv}{(u^2+v^2)^2} \qquad \text{and} \qquad d = \frac{-2auv + b(u^2-v^2)}{(u^2+v^2)^2}.$$

Writing this in terms of s and t, we obtain the transition formula

$$\begin{aligned}(u,v,c,d) &= \Phi\circ\Psi^{-1}(s,t,a,b)\\ &= \left(\frac{s}{s^2+t^2}, \frac{t}{s^2+t^2}, a(t^2-s^2)-2bst, -2ast+b(s^2-t^2)\right).\end{aligned}$$

Again the first two components are just the coordinate transition map, while the last two components are linear in the vector components (a,b). Clearly the transition map $\Phi\circ\Psi^{-1}$ is a diffeomorphism on its domain.

The chart Φ is defined on $\overline{U}_1 = TS^2\backslash T_{(0,0,1)}S^2$ while the chart Ψ is defined on $\overline{U}_2 = TS^2\backslash T_{(0,0,-1)}S^2$. Now define the topology on TS^2 as $\Omega\subset TS^2$ is open if and only if $\Phi[\Omega\cap U_1]$ and $\Psi[\Omega\cap U_2]$ are both open. This topology makes both Φ and Ψ homeomorphisms (since the transition map $\Phi\circ\Psi^{-1}$ is a homeomorphism), and thus makes TS^2 a smooth manifold. ☺

The example above gives us two coordinate charts which cover the entire space TS^2, and we have thus defined a smooth manifold structure on TS^2. However it's a bit mysterious; for example it's not entirely obvious that TS^2 is not homeomorphic to $S^2\times\mathbb{R}^2$. (See Example 12.2.5.) Let's work in a more concrete situation.

Example 12.1.3. Recall that $T\mathbb{R}^3$ is basically $\mathbb{R}^3\times\mathbb{R}^3$, using the Cartesian coordinate chart. For the surface S^2 in $\mathbb{R}^3$, we have computed the tangent spaces T_pS^2 both informally as in Section 10.1 and then rigorously as in Section 11.3. We saw that if $p=(x,y,z)\in S^2$, then a vector $\tilde{v} = a\,\dfrac{\partial}{\partial x}\Big|_p + b\,\dfrac{\partial}{\partial y}\Big|_p + c\,\dfrac{\partial}{\partial z}\Big|_p$ can be identified with $v\in T_pS^2$ if $ax+by+cz=0$; since $x^2+y^2+z^2=1$, this condition always gives a two-dimensional subspace of $\mathbb{R}^3$.

Now consider the map $F\colon\mathbb{R}^6\to\mathbb{R}^2$ given by $F(x,y,z,a,b,c) = \big(x^2+y^2+z^2, ax+by+cz\big)$, and let $M = F^{-1}\{(1,0)\}$. I claim that $TS^2\cong M$, and the first step to showing this is to see that M actually is a manifold. Computing

$$DF(x,y,z,a,b,c) = \begin{pmatrix} 2x & 2y & 2z & 0 & 0 & 0\\ a & b & c & x & y & z\end{pmatrix},$$

it is easy to see that this has rank two everywhere on M, which means that M is a submanifold of dimension four. Now we want to set up the correspondence between TS^2 (as constructed in Example 12.1.2) and our concrete space M.

The easiest way to do this is to define a map $J\colon TS^2\to M$ as follows: let $\iota\colon S^2\to\mathbb{R}^3$ be the inclusion (which is smooth since S^2 is a submanifold). As discussed in Section 11.3, we get at each $p\in S^2$ a map $(\iota_*)_p\colon T_pS^2\to T_{\iota(p)}\mathbb{R}^3$, and the image $(\iota_*)_p[T_pS^2]$ is the kernel of the map $F_*\colon T_{\iota(p)}\mathbb{R}^3\to\mathbb{R}$ which is given in this case by

$$(F_*)_{(x,y,z)}\left(a\,\frac{\partial}{\partial x}\Big|_{(x,y,z)} + b\,\frac{\partial}{\partial y}\Big|_{(x,y,z)} + c\,\frac{\partial}{\partial z}\Big|_{(x,y,z)}\right) = 2(ax+by+cz).$$

Now if $\pi\colon TS^2\to S^2$ denotes the projection which gives the base point of a vector,

$$\pi(v) = p \quad \text{if } v\in T_pS^2, \tag{12.1.1}$$

then we define $\tilde{J}(v) = \big(\iota\big(\pi(v)\big), (\iota_*)_{\pi(v)}(v)\big)\in\mathbb{R}^6$; then since $(x,y,z)=\iota(p)$ satisfies $x^2+y^2+z^2=1$ and $a\,\frac{\partial}{\partial x}+b\,\frac{\partial}{\partial y}+c\,\frac{\partial}{\partial z}$ satisfies $ax+by+cz=0$, we know $\tilde{J}$ actually maps TS^2 into M, and we define $J\colon TS^2\to M$ to be the restriction to M of $\tilde{J}$. Clearly for every $(x,y,z,a,b,c)\in M$ we can let $p=(x,y,z)\in S^2$ and find a

$v \in T_pM$ such that $(\iota_*)_p(v) = a\,\frac{\partial}{\partial x} + b\,\frac{\partial}{\partial y} + c\,\frac{\partial}{\partial z}$ (because $(\iota_*)_p$ is surjective onto the kernel of F_*), so that $J(v) = (x, y, z, a, b, c)$. Hence J is actually a bijection.

To prove that J is actually smooth we can compute it in coordinates. Let's just do north-pole coordinates; south-pole coordinates are basically the same. Recall that if $V = c\,\frac{\partial}{\partial u}\big|_p + d\,\frac{\partial}{\partial v}\big|_p$ where $\phi(p) = (u, v)$, then $\Phi(V) = (u, v, c, d)$. Furthermore since

$$(x, y, z) = \iota \circ \phi^{-1}(u, v) = \left(\frac{2u}{u^2+v^2+1}, \frac{2v}{u^2+v^2+1}, \frac{u^2+v^2-1}{u^2+v^2+1} \right),$$

we compute from the coordinate formula (11.2.1) that

$$\begin{aligned}
\iota_* \left(\frac{\partial}{\partial u} \right) &= \frac{\partial x}{\partial u}\,\frac{\partial}{\partial x} + \frac{\partial y}{\partial u}\,\frac{\partial}{\partial y} + \frac{\partial z}{\partial u}\,\frac{\partial}{\partial z} \\
&= \frac{2(v^2-u^2+1)}{(u^2+v^2+1)^2}\,\frac{\partial}{\partial x} - \frac{4uv}{(u^2+v^2+1)^2}\,\frac{\partial}{\partial y} + \frac{4u}{(u^2+v^2+1)^2}\,\frac{\partial}{\partial z} \\
\iota_* \left(\frac{\partial}{\partial v} \right) &= -\frac{4uv}{(u^2+v^2+1)^2}\,\frac{\partial}{\partial x} + \frac{2(u^2-v^2+1)}{(u^2+v^2+1)^2}\,\frac{\partial}{\partial y} + \frac{4v}{(u^2+v^2+1)^2}\,\frac{\partial}{\partial z}.
\end{aligned}$$

Thus we have

$$\begin{aligned}
\tilde{J} \circ \Phi^{-1}(u, v, c, d) &= \tilde{J} \left(c\,\frac{\partial}{\partial u}\Big|_{\phi^{-1}(u,v)} + d\,\frac{\partial}{\partial v}\Big|_{\phi^{-1}(u,v)} \right) \\
&= \Big(\frac{2u}{u^2+v^2+1}, \frac{2v}{u^2+v^2+1}, \frac{u^2+v^2-1}{u^2+v^2+1}, \frac{2c(v^2-u^2+1) - 4duv}{(u^2+v^2+1)^2}, \\
&\qquad \frac{4uvc + 2d(u^2-v^2+1)}{(u^2+v^2+1)^2}, \frac{4cu+4dv}{(u^2+v^2+1)^2} \Big).
\end{aligned}$$

This shows that the map $\tilde{J}$ is smooth, and since M is a smooth submanifold of $\mathbb{R}^6$ with $\tilde{J}[TS^2] \subset M$, we must have that J is smooth as well. ☺

We have seen in the examples above how the coordinate charts on M generate coordinate charts on TM automatically. Let's generalize this to get the manifold structure on TM; in a moment we will point out the special features of this manifold structure that make it a bundle.

Definition 12.1.4. If M is a smooth n-dimensional manifold, the *tangent bundle* TM is defined to be the union of all tangent spaces $TM = \cup_{p\in M} T_pM$. Given any coordinate chart (ϕ, U) on M, a coordinate chart Φ on TM is defined on $TU = \cup_{p\in U} T_pM$ by

$$\Phi \left(\sum_{k=1}^{n} a^k\,\frac{\partial}{\partial x^k}\Big|_p \right) = \big(\phi(p), a^1, \ldots, a^n\big) \in \mathbb{R}^{2n}. \tag{12.1.2}$$

These coordinate charts cover TM and are C^∞-compatible. The topology on TM is defined by the condition that $\Omega \subset TM$ is open if and only if $\Phi[\Omega \cap TU]$ is open in $\mathbb{R}^{2n}$, and in this topology the charts Φ are all homeomorphisms onto their images.

This definition includes several claims, and to prove them, we need to verify that the transition maps are C^∞ and that the charts are homeomorphisms in the specified topology. First of all, suppose (ϕ, U) and (ψ, V) are two charts on M; then the transition map $\phi \circ \psi^{-1}$ is C^∞. Using the vector change-of-basis formula

(10.4.6), we compute that if $\phi(p) = (x^1, \dots, x^n)$ and $\psi(p) = (y^1, \dots, y^n)$, then a vector

$$v = \sum_{k=1}^{n} a^k \left.\frac{\partial}{\partial x^k}\right|_p = \sum_{k=1}^{n} b^k \frac{\partial}{\partial y^k}$$

must have its components related by

$$a^k = \sum_{j=1}^{n} \frac{\partial x^k}{\partial y^j}\, b^j,$$

where $\frac{\partial x^k}{\partial y^j}$ denotes the partial derivative of the k^{th} component of $\phi \circ \psi^{-1}$ with respect to its j^{th} variable. Thus the transition map is

$$(x^1, \dots, x^n, a^1, \dots, a^n) = \Phi \circ \Psi^{-1}(y^1, \dots, y^n, b^1, \dots, b^n)$$

whose components are

$$x^k = \phi^k \circ \psi^{-1}(y^1, \dots, y^n) \qquad \text{and} \qquad a^k = \sum_{j=1}^{n} b^j \frac{\partial \phi^k \circ \psi^{-1}}{\partial y^j}(y^1, \dots, y^n).$$

Since $\phi \circ \psi^{-1}$ is C^∞, so are its partial derivatives, and thus $\Phi \circ \Psi^{-1}$ is C^∞. Again we note that while $\mathbf{x}$ is almost never linear as a function of $\mathbf{y}$, the vector components $\mathbf{a}$ are always linear as a function of $\mathbf{b}$ (for fixed $\mathbf{y}$).

To check that the charts (Φ, TU) are homeomorphisms, suppose W is an open set in $\mathbb{R}^{2n}$. Then $\Phi^{-1}[W] \subset TM$ is open if and only if $\Psi[TV \cap \Phi^{-1}[W]] \subset \mathbb{R}^{2n}$ is open for every chart (Ψ, TV). But we know $\Psi[TV \cap \Phi^{-1}[W]] = (\Psi \circ \Phi^{-1})(\Phi[TU \cap TV] \cap W])$; furthermore $\Phi[TU \cap TV] = \phi[U \cap V] \times \mathbb{R}^n$ is open, and since $\Psi \circ \Phi^{-1}$ is a homeomorphism, we conclude $\Psi[TV \cap \Phi^{-1}[W]]$ is open for every chart (Ψ, TV). The fact that Φ is an open map is even easier: if $\Omega \subset TU$ is open, then $\Phi[TU \cap \Omega] = \Phi[\Omega]$ is open in $\mathbb{R}^{2n}$ by definition.

Now we have defined a smooth manifold structure on TM, and the projection map $\pi\colon TM \to M$ is smooth in this structure, but TM is clearly more than just a manifold. For every $p \in M$ the set $\pi^{-1}\{p\} = T_pM$ is a vector space, and the coordinate charts respect this vector space structure (since the transition maps are all linear in the last n variables). We are going to want to generalize this pretty soon, since as soon as we have a vector space T_pM we have lots of other vector spaces constructed in terms of it as in Chapter 4 (such as the dual space, the space of k-forms, the space of $(0,2)$ symmetric tensors, etc.). We'll want to fit together all these vector spaces just as we fit together all the tangent spaces, and we'd expect to see the same structure. Let us now describe the properties of TM that we expect any vector bundle structure to have:

- There is a smooth projection $\pi\colon TM \to M$.
- For every $p \in M$, the set $T_pM = \pi^{-1}\{p\}$ is a vector space isomorphic to $\mathbb{R}^n$.
- The coordinate charts (Φ, TU) define an diffeomorphism from $TU = \pi^{-1}[U]$ to $\phi[U] \times \mathbb{R}^n$.
- For each chart (Φ, TU) and $p \in \pi[TU]$, we have $\Phi[T_pM] = \{x\} \times \mathbb{R}^n$, and $\Phi|_{T_pM}$ is a vector space isomorphism. (That is, if you hold the base point constant, you get an invertible linear transformation.)

A space which satisfies properties like these is called a *smooth vector bundle*; the tangent bundle is the prototypical example.

Definition 12.1.5. A *smooth vector bundle* consists of a smooth manifold M (the base space), a smooth manifold E (the total space), and a vector space V (the fiber space) with a map $\pi\colon E \to M$ (the projection) such that

(1) π is smooth and surjective;
(2) For each $p \in M$ the space $E_p = \pi^{-1}\{p\}$ (the fiber over p) is a vector space isomorphic to V.
(3) For each $p \in M$ there is an open $\Omega \subset M$ and a diffeomorphism $\Psi\colon \pi^{-1}[\Omega] \to \Omega \times V$ (a local trivialization) such that:
   (a) $\Psi[T_qM] = \{q\} \times V$ for all $q \in \Omega$
   (b) $\Psi|_{T_qM}$ is a linear isomorphism onto $\{q\} \times V$.

The tangent bundle is of course a special case of a vector bundle: in that case $E = TM$ and $V = \mathbb{R}^n$, and the local trivializations (Ψ, Ω) come from coordinate charts $(\mathbf{x} = \phi, U)$ by $\Omega = U$ and $\Psi(v) = \big(\pi(v), a^1, \dots, a^n\big)$ if $v = \sum_k a^k \frac{\partial}{\partial x^k}\big|_{\pi(v)}$. Note that this is slightly different from the coordinate charts Φ defined in the Examples above; the last n components of Φ and Ψ are the same, but the first component of Φ is $\phi(p) \in \mathbb{R}^n$ while the first component of Ψ is p itself. The reason for the difference is that in Definition 12.1.5 we allow for the possibility that the local trivialization Ψ may be defined globally on all of M even if M itself is not homeomorphic to $\mathbb{R}^n$. See Example 12.2.2 in the next Section.

Definition 12.1.6. Suppose E and $\tilde{E}$ are two smooth vector bundles over M, with projections $\pi\colon E \to M$ and $\tilde{\pi}\colon \tilde{E} \to M$ and fiber spaces V and $\tilde{V}$. We say that E and $\tilde{E}$ are *bundle-isomorphic* if there is a diffeomorphism $\Upsilon\colon E \to \tilde{E}$ such that

(1) $\tilde{\pi} \circ \Upsilon = \pi$.
(2) For each $p \in M$, the restriction $\Upsilon|_{E_p}$ is a linear isomorphism onto $\tilde{E}_p$.

Since $E_p = \pi^{-1}\{p\}$ and $\tilde{E}_p = \tilde{\pi}^{-1}\{p\}$, the first condition is what allows the second condition to make sense.

Definition 12.1.5 could be weakened to *topological vector bundle* by replacing "smooth" everywhere with "continuous," although all vector bundles we care in differential geometry are smooth. Note that a tangent bundle is a smooth vector bundle, although a topological manifold which does not have a smooth structure may not have a tangent bundle (differentiating the coordinate transition maps was essential to get our local trivializations on TM).

Finally one frequently understands the global structure of a tangent bundle by looking at vector fields defined on M.

Definition 12.1.7. If E is a vector bundle over M with projection π and fiber space V, a *section of the bundle* is a smooth map $X\colon M \to E$ such that $\pi \circ X$ is the identity; in other words $X(p) \in E_p$ for all $p \in M$. If $E = TM$ with $E_p = T_pM$, a section of the bundle is called a *vector field*.

12.2. **Special cases.** The first question one can ask about a vector bundle is whether it is trivial or not.

Example 12.2.1. If M is any smooth manifold and V is any vector space, then $E = M \times V$ is a vector bundle over M. I just take π to be projection on the first factor, and a single global trivialization (Φ, M) defined by $\Phi(p, v) = (p, v)$. This is called a *trivial bundle*. More generally any bundle which is bundle-equivalent to a trivial bundle is also called a trivial bundle, and the bundle-equivalence map

is called a "trivialization," which justifies the terminology "local trivialization" for the maps in Definition 12.1.5.

Suppose we know that TM is a trivial bundle with projection π, and let $\Upsilon\colon TM \to M \times \mathbb{R}^n$ be a trivialization. Let x be an arbitrary nonzero vector in $\mathbb{R}^n$. Define a map $X\colon M \to TM$ by the formula $X(p) = \Upsilon^{-1}(p, x)$. Since Υ^{-1} maps each $\{p\} \times \mathbb{R}^n$ to T_pM isomorphically, we know that $X(p) \in T_pM$ for each p and that $X(p)$ is never zero. Furthermore since Υ must be a diffeomorphism we conclude that X is smooth. Thus X is a vector field on M as in Definition 12.1.7.

For example when $M = \mathbb{R}^n$ we have many nowhere zero vector fields such as

$$X(p) = \sum_{k=1}^{n} a^k \left.\frac{\partial}{\partial x^k}\right|_p,$$

for any choice of constants $(a^1, \dots, a^n)$, expressed in the global Cartesian chart. In particular the vector fields E_k defined by $E_k(p) = \left.\frac{\partial}{\partial x^k}\right|_p$ for $1 \le k \le n$ form a basis of every T_pM. ☺

We can use the same idea in the opposite direction to show that a vector bundle is trivial.

Example 12.2.2. Let $M = S^1$ considered as a submanifold of $\mathbb{R}^2$, and let $\iota\colon S^1 \to \mathbb{R}^2$ be the inclusion. We will construct a nowhere zero vector field on S^1. Begin with the vector field $\tilde{X}\colon \mathbb{R}^2 \to T\mathbb{R}^2$ given by

$$\tilde{X}(x, y) = -y \left.\frac{\partial}{\partial x}\right|_{(x,y)} + x \left.\frac{\partial}{\partial y}\right|_{(x,y)}. \tag{12.2.1}$$

Certainly this is smooth and is zero if and only if $(x, y) = (0, 0)$. Let $\overline{X}\colon S^1 \to T\mathbb{R}^2$ be $\overline{X} = \tilde{X} \circ \iota$; since ι is smooth, so is $\overline{X}$.

Recall that the circle is defined as $F^{-1}\{1\}$ where $F(x, y) = x^2 + y^2$. Thus as in Section 11.3, we know $\iota_*[T_pS^1] = \ker F_* \subset T_{\iota(p)}\mathbb{R}^2$, which means a vector $a \left.\frac{\partial}{\partial x}\right|_{(x,y)} + b \left.\frac{\partial}{\partial y}\right|_{(x,y)}$ is in $\iota_*[T_pS^1]$ if and only if $\iota(p) = (x, y)$ and $ax + by = 0$. We therefore check immediately that $\overline{X}(p) \in \iota_*[T_pS^1]$ for every $p \in S^1$; since ι_* is an isomorphism we conclude that for each $p \in S^1$ there is a unique vector $X(p) \in T_pS^1$ such that $\iota_*(X(p)) = \overline{X}(p)$.

This gives a map $X\colon S^1 \to TS^1$ with $\pi \circ X$ equal to the identity, but we need to check that X is actually smooth in order to conclude that X is a vector field on S^1. By definition we have to check its expression in coordinates, so consider a stereographic coordinate chart (ϕ, U) as in Example 7.1.7, given explicitly by

$$(x, y) = \left(\frac{2t}{t^2+1}, \frac{t^2-1}{t^2+1}\right), \qquad t = \frac{x}{1-y}. \tag{12.2.2}$$

In this chart we have

$$\iota_*\left(\left.\frac{\partial}{\partial t}\right|_p\right) = \frac{2(1-t^2)}{(t^2+1)^2} \left.\frac{\partial}{\partial x}\right|_{\iota(p)} + \frac{4t}{(t^2+1)^2} \left.\frac{\partial}{\partial y}\right|_{\iota(p)}, \tag{12.2.3}$$

while equations (12.2.1) and (12.2.2) combine to give

$$\overline{X}(p) = -\frac{(t^2-1)}{t^2+1} \left.\frac{\partial}{\partial x}\right|_{\iota(p)} + \frac{2t}{t^2+1} \left.\frac{\partial}{\partial y}\right|_{\iota(p)}.$$

Matching against (12.2.3) and using $\iota_* \circ X = \overline{X}$, we find that

$$X(p) = \frac{2}{t^2+1} \frac{\partial}{\partial t}\Big|_p \qquad \text{where } t = \phi(p),$$

so that its expression in coordinates is

$$\Phi \circ X \circ \phi^{-1}(t) = \left(t, \frac{2}{t^2+1}\right)$$

which is certainly C^∞ as a function on $\mathbb{R}$. The same computation works in the other stereographic chart.

We have defined a vector field $X\colon S^1 \to TS^1$ which is nowhere zero; since T_pS^1 is one-dimensional for every p, we see that $X(p)$ spans T_pS^1 for every p. Define a map $\Upsilon\colon S^1 \times \mathbb{R} \to TS^1$ by $\Upsilon(p,a) = aX(p)$. Given any vector $v \in TS^1$ with $p = \pi(v)$, we know $v = aX(p)$ for some unique a and thus $v = \Upsilon(p,a)$; hence Υ is a bijection. To prove that Υ is a bundle-isomorphism, we write it in coordinates as

$$\Phi \circ \Upsilon \circ \tilde{\phi}^{-1}(t,a) = aX(\phi^{-1}(t)) = \left(t, \frac{2a}{t^2+1}\right)$$

where $\tilde{\phi} = \phi \times \mathrm{id}$ is a coordinate chart on $S^1 \times \mathbb{R}$. Clearly this map is smooth and has smooth inverse $(s,b) \mapsto (s, (s^2+1)b/2)$, so that Υ is a diffeomorphism. The other parts of Definition 12.1.6 are easy to check: Υ maps $\{p\} \times \mathbb{R}$ isomorphically onto T_pS^1 for each p. ☺

In Example 12.2.2, we get a trivialization $\Upsilon^{-1}\colon TS^1 \to S^1 \times \mathbb{R}$ which does *not* come from a coordinate chart on S^1. This is the reason that our Definition 12.1.5 did not require that the open sets $\Omega \subset M$ be homeomorphic to $\mathbb{R}^n$. The technique of Example 12.2.2 works the same way in general: if we have a collection $X_1, \ldots, X_n$ of smooth vector fields on a manifold M such that at every $p \in M$ the vectors $\{X_1(p), \ldots, X_n(p)\}$ form a basis of T_pM, then the tangent bundle is trivial. This is a very useful result which we write as a theorem.

Theorem 12.2.3. *A vector bundle E over M with projection π and fiber space V of dimension K is trivial if and only if there is a collection of smooth sections $X_1, \ldots, X_K$ such that at every $p \in M$, the vectors $\{X_1(p), \ldots, X_K(p)\}$ are a basis of E_p.*

Proof. If E is trivial and $\Upsilon^{-1}\colon E \to M \times V$ is a trivialization, then given any basis $\{x_1, \ldots, x_K\}$ of V, we define $X_k(p) = \Upsilon(p, x_k)$ for any p and $1 \le k \le K$; then each X_k is smooth, and since for fixed p the map $v \mapsto \Upsilon(p,v)$ is a vector space isomorphism, the vectors $X_k(p)$ are a basis if and only if the vectors x_k are a basis.

Conversely if $\{X_1, \ldots, X_K\}$ is a set of vector fields which form a basis at every point, define a map $\Upsilon\colon M \times V \to E$ by constructing a basis $\{x_1, \ldots, x_K\}$ and setting

$$\Upsilon\left(p, \sum_{k=1}^K a^k x_k\right) = \sum_{k=1}^K a^k X_k(p)$$

for any collection of coefficients $\{a^1, \ldots, a^K\}$. We know Υ is smooth, that it maps $\{p\} \times V$ isomorphically onto E_p, and that it is thus a bijection globally. The fact that its inverse is also smooth follows from using the trivializations. Hence it is a bundle isomorphism. □

The simplest vector bundle which is not trivial is the Möbius bundle over S^1, which we will construct in the next example. Of course, this bundle is distinct from TS^1, which is trivial. The simplest tangent bundle which is nontrivial is TS^2, which we will discuss in a moment.

Example 12.2.4. The only vector bundles you can easily visualize are one-dimensional vector bundles over S^1, since they are two-dimensional manifolds. There are essentially two such bundles. The trivial bundle over S^1 with one-dimensional fibers looks like $S^1 \times \mathbb{R}$, and you can visualize it as a cylinder. You can visualize TS^1 in the same way by imagining it as a tangent line attached to the circle at each point: rotating the tangent line into a direction perpendicular to the plane of the circle gives the cylinder, and this is the trivialization.

The nontrivial bundle with one-dimensional fibers is homeomorphic to the Möbius band; compare Example 7.2.3. Let E be the quotient space of $\mathbb{R} \times \mathbb{R}$ modulo the equivalence relation $(x, y) \equiv (2n\pi + x, (-1)^n y)$, which is a free and proper discrete group action on $\mathbb{R}^2$. A fundamental domain is the set $[0, 2\pi] \times \mathbb{R}$, and we can think of E as the quotient modulo the identification $(0, v) \equiv (2\pi, -v)$. There are two trivializations: let $\Omega_1 = E\backslash\{0\} \times \mathbb{R} \cong (0, 2\pi) \times \mathbb{R}$ with the obvious trivialization Φ_1, and let $\Omega_2 = E\backslash\{\pi\} \times \mathbb{R}$ which is equivalent to $(-\pi, \pi) \times \mathbb{R}$ with its obvious trivialization.

To prove this bundle is nontrivial, we prove there is no nonzero section. Certainly on Ω_1 there are plenty of nonzero sections, and every one can be expressed in the trivialization Φ_1 as $X(t) \simeq t \mapsto (t, f(t))$ where $f(t)$ is a positive function on $(0, 2\pi)$. In the other trivialization, this section would be defined on $(-\pi, 0) \cup (0, \pi)$ and would take the form

$$X(t) \simeq \begin{cases} f(t) & 0 < t < \pi, \\ -f(t + 2\pi) & -\pi < t < 0, \end{cases}$$

using the transition formulas found in Example 7.2.3. The only way for this to represent a restriction of a vector field on S^1 is if we can define X at $t = 0$, where it would have to be zero (since f is always positive). Hence there is no way to construct a nowhere-zero vector field on the Möbius bundle, and it is nontrivial. ☺

Now let us prove that TS^2 is nontrivial; in fact we can prove the stronger result that there is *no* nowhere-zero smooth vector field on S^2. This result is called the "hairy ball theorem" (I did not name it). The best proofs use tools of algebraic topology and can be easily generalized; however there is a cute proof due to Milnor[24] which I will present here.

Example 12.2.5. Using the bundle-isomorphism we constructed in Example 12.1.3 between TS^2 and the set

$$\{(x, y, z, u, v, w) \in \mathbb{R}^6 \,|\, x^2 + y^2 + z^2 = 1 \quad \text{and} \quad xu + yv + zw = 0\},$$

it is sufficient to show that any smooth map $X = (u, v, w) \colon S^2 \to \mathbb{R}^3$ satisfying

$$xu(x, y, z) + yv(x, y, z) + zw(x, y, z) = 0 \quad \text{whenever } x^2 + y^2 + z^2 = 1 \tag{12.2.4}$$

must be $(0, 0, 0)$ somewhere on the sphere.

[24]Analytic proofs of the "hairy ball theorem" and the Brouwer fixed point theorem, Amer. Math. Monthly, July 1978, pp. 521–524

Assume that $u^2+v^2+w^2 \neq 0$ everywhere on S^2; then the norm of X is a smooth function which we can divide by, and so we lose no generality by assuming that X is a *unit* vector field with $u^2+v^2+w^2=1$. Define $F\colon \mathbb{R}\times S^2 \to \mathbb{R}^3$ by

$$F(t,x,y,z) = \big(x - tu(x,y,z), y - tv(x,y,z), z - tw(x,y,z)\big).$$

Then the orthogonality condition (12.2.4) together with the fact that (x,y,z) is on the sphere and X is unit implies that

$$\|F(t,x,y,z)\|^2 = 1 + t^2.$$

Hence F maps the unit sphere into the sphere of radius $\sqrt{1+t^2}$. I want to prove that it is in fact a bijection, at least for t close to 0. For a fixed (small) $t = t_o$, compute

$$DF(t,x,y,z) = \begin{pmatrix} 1-tu_x & -tu_y & -tu_z \\ -tv_x & 1-tv_y & -tv_z \\ -tw_x & -tw_y & 1-tw_z \end{pmatrix}.$$

This is of the form $DF(t_o,x,y,z) = I - t_o M(x,y,z)$ and since determinant is a continuous function, for any particular (x,y,z) there is a positive $T(x,y,z)$ such that for $|t_o| < T(x,y,z)$ the matrix $DF(t_o,x,y,z)$ is invertible. In fact we can estimate $T(x,y,z)$ explicitly in terms of M and conclude that T is continuous on S^2, which means it attains its (positive) minimum somewhere. Hence for sufficiently small values of t_o the map $p \mapsto F(t_o,p)$ is a diffeomorphism from the unit sphere onto the sphere of radius $\sqrt{1+t^2}$.

Now extend F to a map $G\colon \mathbb{R}\times\mathbb{R}^3 \to \mathbb{R}^3$ by

$$G(t,p) = |p| F(t, p/|p|).$$

Then since at each fixed time F maps the unit sphere onto the sphere of radius $\sqrt{1+t^2}$, we know that the image of the unit ball under G is the ball of radius $\sqrt{1+t^2}$, and that means by the change-of-variables formula Theorem (5.3.2), we have

$$\begin{aligned} \int_{B_1(0)} \det DG(x,y,z)\,dx\,dy\,dz &= \int_{G[B_1(0)]} dx\,dy\,dz \\ &= \operatorname{vol}\{G[B_1(0)]\} = \operatorname{vol}\{B_{\sqrt{1+t^2}}(0)\} = \frac{4\pi}{3}(1+t^2)^{3/2}. \end{aligned} \tag{12.2.5}$$

On the other hand, $\det DG(x,y,z)$ is clearly a *polynomial* in t and that means the left side of (12.2.5) is also a polynomial in t. This is a contradiction, and thus (u,v,w) must be zero somewhere on the sphere. ☺

There is actually a quite simple thing going on here. In $\mathbb{R}^2$ we can define a rotation vector field $X = -y\frac{\partial}{\partial x} + x\frac{\partial}{\partial y}$ which is zero only at the origin and descends to a nowhere-zero vector field on S^1 as in Example 12.2.2. On the other hand every rotation vector field in $\mathbb{R}^3$ is basically equivalent to the rotation that fixes the z-axis, which is $X = -y\frac{\partial}{\partial x} + x\frac{\partial}{\partial y}$ (with zero components in the $\frac{\partial}{\partial z}$ direction); hence it fixes the north and south poles of the 2-sphere. But on $\mathbb{R}^4$ we can set up the vector field $X = -x\frac{\partial}{\partial w} + w\frac{\partial}{\partial x} - z\frac{\partial}{\partial y} + y\frac{\partial}{\partial z}$, which lies in the tangent space of the 3-sphere everywhere. Since $w^2+x^2+y^2+z^2=1$ on the 3-sphere, this vector field descends to a nowhere-zero vector field on S^3. Hence there is a very significant distinction between odd-dimensional spheres and even-dimensional spheres in terms of the tangent bundle.

12.3. **The push-forward map.** Now recall from that if $F\colon M \to N$ is a smooth map, we know from Chapter 11 that for each p there is a derivative map $(F_*)_p\colon T_pM \to T_{F(p)}N$, defined by the condition that

$$((F_*)_p v)(h) = v(h \circ F)$$

for any smooth function h defined in a neighborhood of $F(p)$. We can put these all together to get a global derivative map on the tangent bundles:

$$F_*\colon TM \to TN \tag{12.3.1}$$

given in the obvious way, where F_* maps any vector v based at p to the vector $(F_*)_p(v)$ based at $F(p)$.

This is a special case of a *bundle map*, which is a generalization of a bundle-isomorphism: we still want to commute with the projections and obtain linear maps, but we no longer demand that the linear maps be isomorphisms. We will mostly be concerned with bundle maps that are induced by maps of manifolds in this text.

The notion (12.3.1) makes it easier to talk about global properties of maps. For example the nicest maps $F\colon M \to N$ are those where the rank of F_* is constant.

Definition 12.3.1. Suppose $F\colon M \to N$ is a smooth map.

F is called an *immersion* if the rank of F_* is everywhere equal to the dimension of M. If F is both an immersion and a homeomorphism onto its image, then F is called an *embedding*.

F is called a *submersion* if the rank of F_* is everywhere equal to the dimension of N.

Clearly F can only hope to be an immersion if $\dim(N) \geq \dim(M)$, and can only hope to be a submersion if $\dim(M) \geq \dim(N)$. Finding an immersion or submersion between two manifolds is a great way to relate their topological properties. For example a nonsingular curve $\gamma\colon \mathbb{R} \to M$ with $\gamma'(t) \neq 0$ for every t is an immersion. And if F is a submersion then all its inverse images are smooth manifolds.

We have already computed the bundle maps F_* in coordinates: we computed the maps at each individual tangent space using Proposition 11.2.1, and we defined coordinates on the tangent bundle by Definition 12.1.4. We just have to combine these to get the coordinate expression of F_*.

Proposition 12.3.2. *Let $F\colon M \to N$ be a smooth map, and let $F_*\colon TM \to TN$ denote the map defined on each T_pM for $p \in M$ by Proposition 11.1.1. Then F_* is a smooth map from TM to TN.*

Proof. We just have to compute in charts. Given coordinate charts (ϕ, U) on M and (ψ, V) on N with corresponding charts (Φ, TU) on TM and (Ψ, TV) on TN, we have

$$(u^1, \ldots, u^n, b^1, \ldots, b^n) = \Psi \circ F_* \circ \Phi^{-1}(x^1, \ldots, x^m, a^1, \ldots, a^m),$$

where

$$u^k = \psi^k \circ F \circ \phi^{-1}(x^1, \ldots, x^m) \qquad \text{and} \qquad b^k = \sum_{j=1}^{m} a^j \, \frac{\partial \psi^k \circ F \circ \phi^{-1}}{\partial x^j}(x^1, \ldots, x^m). \tag{12.3.2}$$

Clearly if F is smooth then $\psi \circ F \circ \phi^{-1}$ is smooth by definition, and thus the functions (12.3.2) are smooth. Hence F_* is smooth by definition. □

The reason this Proposition is important is that maps $F\colon M \to N$ give induced maps $F_*\colon TM \to TN$ with algebraic structure, and thus we get induced linear maps from one algebraic-topological space to another.

13. Bumps, partitions of unity, and the Whitney embedding

"I beg your pardon, but what do you mean, 'naked'? My parts are showing? Oh, my goodness!"

13.1. **Motivation.** We now want to solve the problem we alluded to in Remark 10.2.2 and Remark 10.3.3. Namely, we want to work with functions $f\colon M \to \mathbb{R}$ which are equal to an arbitrary germ near a point, and specifically we want to be able to define a smooth function by any smooth formula in a coordinate chart and know that it extends to the manifold in a C^∞ way. In Chapter 12 we assumed that our test functions were defined on some open set containing the point p: it's easy to find lots of smooth functions on $\mathbb{R}^n$ just by writing down formulas, but it's hard to patch together functions defined by different formulas in different coordinate charts together, to get a function on the entire manifold. This section is where we solve this problem.

To illustrate that this is not a trivial sort of problem with an obvious solution, here's an example where it *doesn't* work.

Example 13.1.1. Suppose that instead of requiring all maps to be C^∞ in coordinates, we had required that they be real-analytic. In other words every transition map $\phi \circ \psi^{-1}$ not only had infinitely many derivatives at each point, but in fact converged in some open neighborhood to the corresponding Taylor series. This is a very natural thing to do in algebraic geometry when one defines the manifolds as zero-sets of polynomials (which are of course real-analytic functions), and also when studying complex manifolds (where the transition functions should be holomorphic). We already have a problem on $\mathbb{R}$: the function $f\colon (-1,1) \to \mathbb{R}$ given by $f(x) = 1/(2-x)$ is real-analytic on $(-1,1)$ since

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!}\, x^n = \sum_{n=0}^{\infty} \frac{x^n}{2^{n+1}} = \frac{1}{2-x} = f(x) \quad \text{everywhere on } (-1,1),$$

but there is no way to extend f to a real-analytic function $\tilde{f}\colon \mathbb{R} \to \mathbb{R}$. The reason is that for any point x_o with $|x_o| < 2$ we can write

$$f(x) = \sum_{n=0}^{\infty} \frac{1}{2 - x_o - (x - x_o)} = \sum_{n=0}^{\infty} \frac{(x-x_o)^n}{(2-x_o)^{n+1}},$$

and by uniqueness of power series, this must be the power series for any extension $\tilde{f}$ at x_o. In other words, any real-analytic function which agrees with the given f on a small open set must actually agree with it everywhere on the common domain, and thus is forced to blow up at $x = 2$. Hence it's actually quite easy for the extension problem to be unsolvable if we aren't working in the right context. ☺

Here's a familiar example from physics and partial differential equations, where we define a smooth function in a coordinate chart and then need to extend it to the entire manifold.

Example 13.1.2. A common technique is to use separation of variables to solve the equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0.$$

The coordinates (x, y) are fine if you're just concerned about it in a rectangle, but nature tends to prefer circles over rectangles, and so polar coordinates (r, θ) are often preferable. In polar coordinates the equation becomes

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} = 0.$$

To do separation of variables, you assume a solution of the form

$$u(r, \theta) = R(r)\Theta(\theta).$$

If there is such a solution, then you must have

$$\Theta''(\theta) = -\lambda\Theta(\theta) \qquad \text{and} \qquad r\frac{d}{dr}\left(r\frac{dR}{dr}\right) = \lambda R(r).$$

If $\lambda \geq 0$ then[25] you get the solution $\Theta(\theta) = A\sin(\sqrt{\lambda}\theta) + B\cos(\sqrt{\lambda}\theta)$, and $R(r) = Cr^{\sqrt{\lambda}} + Dr^{-\sqrt{\lambda}}$. These are solutions on the region $(r, \theta) \in (0, \infty) \times (-\pi, \pi)$, with no problem. However you don't want solutions on that nonphysical half-infinite strip, you want solutions on the actual Euclidean plane. Recall the diagram in Figure 13.1 relating these spaces: on the left we have the nonphysical space of coordinate curves, which fills up everything but the left half-line in the plane.

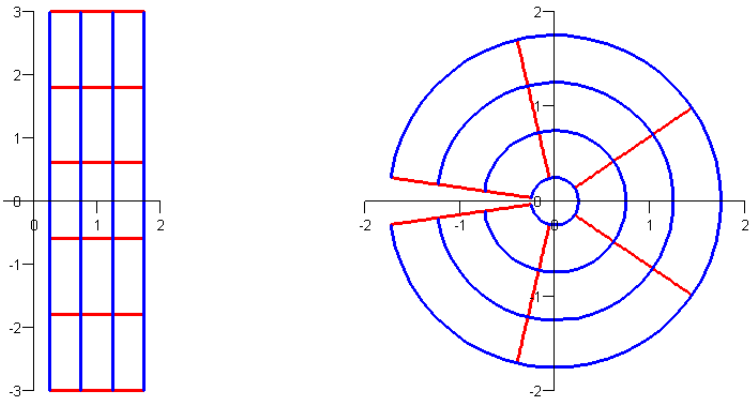

FIGURE 13.1. Coordinate curves in the $r\theta$-plane on the left, and their image under (6.2.1) in the xy-plane on the right.

The formulas above solve the equations on a coordinate chart of the manifold (which is the plane minus the leftward ray from the origin), which imposes no restriction, but to get a solution on the entire manifold, you need to extend your function from the coordinate chart to the entire manifold. So first of all, if your solution $u(r, \theta)$ is actually a smooth function on the manifold $M = \mathbb{R}^2$, then in particular it will respect that $u(r, \theta) = u(r, \theta + 2\pi)$, and this forces $\lambda = n^2$ for some nonnegative integer n. Having done this, you now have a function Θ which is defined not just on $(-\pi, \pi)$ but actually on the manifold S^1. Remember, you want to get a smooth function on the entire manifold, and the definition of a smooth function is one such that, for every coordinate chart (ϕ, U), the map $f \circ \phi^{-1}$ is smooth. You started out by noticing that your function was smooth on the polar

[25] If $\lambda < 0$ you can do the same thing, but nothing at all will work when you try to extend it.

chart coming from excluding the leftward ray from the origin. Now satisfying the periodicity corresponds to looking at the function in another polar chart, this one coming from excluding the rightward ray from the origin. So in forcing periodicity, you've essentially shown that your function is smooth in two coordinate charts which together cover the plane minus the origin. But you haven't proved the function is smooth in a coordinate chart that contains the origin.

From your differential equation solution you get that $R(r) = Cr^n + Dr^{-n}$. Now your solution is defined on $\mathbb{R}^2$ minus the origin, and if you want the origin included, you have to again restrict the solution because r^{-n} is not a continuous function on the plane. Having done this you now get the solutions $u(r,\theta) = Ar^n \sin n\theta + Br^n \cos n\theta$, and you're *still* not done because you need to worry about this being smooth near the origin (remember, $r = 0$ is a very bad coordinate singularity). So you change into Euclidean coordinates by using $z = re^{i\theta} = x + iy$ and noticing that

$$r^n \cos n\theta = \mathrm{Re}(r^n e^{in\theta}) = \mathrm{Re}(x+iy)^n = \sum_{k=0}^{\lfloor n/2 \rfloor} (-1)^k \binom{n, 2k}{x}^{n-2k} y^{2k}$$

$$r^n \sin n\theta = \mathrm{Im}(r^n e^{in\theta}) = \mathrm{Im}(x+iy)^n = \sum_{k=0}^{\lfloor (n-1)/2 \rfloor} (-1)^k \binom{n, 2k+1}{x}^{n-2k-1} y^{2k+1},$$

which clearly *are* smooth in a neighborhood of the origin. We thus find that the most general solution is of the form

$$u(x,y) = \sum_{n=0}^{\infty} a_n \mathrm{Re}(x+iy)^n + b_n \mathrm{Re}(x-iy)^n$$

where the coefficients a_n and b_n can be determined by a boundary condition on some circle. ☺

All of what was done above is necessary because you wanted to define a smooth function on the entire manifold, and you had some idea what it was in a coordinate chart, but the fact that a function is smooth on a coordinate chart does not at all imply it can be extended to be a smooth function on the entire manifold. So you've probably dealt with the problem of extending functions from a coordinate chart to an entire manifold before, but probably without thinking of it in those terms.

Let's take a look at another example on an actual nontrivial manifold.

Example 13.1.3. If I want to define a smooth function on S^2, I could write something down in spherical coordinates like $f(\theta,\phi) = \sin\phi$, but this won't work. It is clearly a continuous and in fact smooth function on an open set in $\mathbb{R}^2$, defined by the coordinate range $0 < \theta < \pi$, $-\pi < \phi < \pi$, but that only covers a portion of the sphere. More precisely, if $\varphi\colon U \subset S^2 \to (0,\pi)\times(-\pi,\pi) \subset \mathbb{R}^2$ is the coordinate chart, then $F = f \circ \varphi$ defines a smooth function on $U \subset S^2$ but not on all of S^2. Why not? Well the coordinate chart covers all of the sphere except for the north and south poles, along with the great semicircle $y = 0$, $x \leq 0$ which joins them. So the question is whether there's a way to define the function at those points.

The easiest way to understand this is to look at F in another coordinate chart. Let's pick one that includes the north pole, since that's where we suspect there may be a problem. For example a parametrization of the open top hemisphere V is given by $(x,y,z) = (u,v,\sqrt{1-u^2-v^2})$ for (u,v) in the open unit disc, so that the coordinate chart is $\psi\colon V \to \mathbb{R}^2$. In these coordinates we have $u = \sin\theta\cos\phi$ and

$v = \sin\theta \sin\phi$, so that $\sin\theta = \sqrt{u^2+v^2}$ and $\sin\phi = \dfrac{v}{\sqrt{u^2+v^2}}$. The intersection of the coordinate charts is

$$U \cap V = \{(x,y,z) \in \mathbb{R}^3 \,|\, x^2+y^2+z^2 = 1, z > 0, \text{either } y \neq 0 \text{ or } y = 0 \text{ and } x > 0.\}$$

Now $F \circ \psi^{-1}(u,v) = \frac{v}{\sqrt{u^2+v^2}}$, and so if F *were* a smooth function on all of S^2, then $\lim_{(u,v)\to 0} F \circ \psi^{-1}(u,v)$ would have to exist. But it doesn't since $F(u,0) = 0$ for any $u > 0$ while $F(0,v) = 1$ for any $v > 0$.

The same sort of thing happens with functions like $F \circ \varphi^{-1}(\theta,\phi) = \sin\theta$, which in the top hemisphere chart looks like $F \circ \psi^{-1}(u,v) = \sqrt{u^2+v^2}$, so that it's continuous but not differentiable at the poles.

On the other hand a function like $F \circ \varphi^{-1}(\theta,\phi) = \cos\theta$ is actually C^∞. The easiest way to see this is to note that F is actually the restriction of a smooth function $\tilde{F}\colon \mathbb{R}^3 \to \mathbb{R}$ given by $\tilde{F}(x,y,z) = z$. If $\iota\colon S^2 \to \mathbb{R}^3$ is the embedding, then $F = \tilde{F} \circ \iota$, and since $\tilde{F}$ is obviously C^∞ and ι is C^∞, we know F is C^∞ as a function on S^2. In general it is easy to define smooth functions on submanifolds of Euclidean space by restricting smooth functions on the whole space, but this is not ideal for us. For one thing, it's not intrinsic to the manifold (it depends on an ambient space). But more importantly, it's hard to guarantee we could get *all* the functions we want by restricting some smooth function on the ambient space. ☺

What we want then is a technique to extend a function (or vector field, or other object) that is defined on a single coordinate chart to the entire manifold in such a way that it agrees on some open set with our formula. This will enable us to obtain results like the following.

Theorem 13.1.4. *Let M be a smooth manifold, let $p \in M$, and let (ϕ, U) be any coordinate chart on M with $p \in U$. Let $F\colon \mathbb{R}^n \to \mathbb{R}$ be an arbitrary smooth function, so that $f = F \circ \phi$ is a smooth function on U. Then there is a smooth function $\tilde{f}\colon M \to \mathbb{R}$ such that for some open $V \ni p$ we have $f|_V = \tilde{f}|_V$.*

There is an analogous result for vector fields on M, which is nice because despite everything we did in Chapter 12, the only vector field that we know to exist on a general manifold is the identically-zero vector field. Furthermore actually building vector fields was a bit of a pain each time we did it. And recall that by Example 12.1.2, there cannot be any smooth vector field on S^2 satisfying the rather mild-seeming condition that it is nowhere zero. So it's not obvious that there are lots of vector fields on a general smooth manifold, and what we do in the future will depend on the fact that there are. And more generally we will be able to use the same technique to show that there are many sections of any vector bundle. The tool for handling this is "bump functions," which we will discuss in Section 13.2.

One problem with what the Extension Theorem 13.1.4 is that although it lets us specify functions arbitrarily on a possibly small open set, we lose control of the function on the rest of the manifold. For example if the function $F \circ \phi$ is positive on the open set U, we don't know it will be positive on the rest of M. This becomes important when we want to define an object on a manifold, and we know what it should be in a coordinate chart, and we know that it's invariant when we change the coordinate chart, but we need it to extend to the entire manifold. For example suppose I have somehow defined a smooth function $f\colon S^2 \to \mathbb{R}$, and I want to integrate it. Integration is not an invariant thing in differential geometry (if it were, the volume of the sphere would be a constant independent of the shape of the

sphere, and clearly I can do lots of deformations of the sphere which preserve the smooth-manifold structure but change the volume). However I can decide that I will compute the area of a region of the sphere by using the vector calculus technique (i.e., taking seriously the embedding into $\mathbb{R}^3$). But vector calculus only tells me how to calculate in coordinates; how do I know that the integral over the entire sphere is actually well-defined? The same issue arises when discussing Riemannian metrics. The technical tool for getting around this is called a *partition of unity*, which is closely related to bump functions and which we will discuss in Section 13.3.

Finally as our first real application of these tools, we will prove the Whitney embedding theorem, that any compact smooth manifold is actually a submanifold of some $\mathbb{R}^N$, in Section 13.4. We will also discuss a problem from dynamical systems, the Takens embedding theorem, which actually has very practical application for understanding time series. Both of these theorems are consequences of the fact that there are many smooth functions on a smooth manifold and that generically it is easy to make them linearly independent.

13.2. **Bump functions.** The starting point for piecing things together on a manifold is the basic fact that there are C^∞ functions from $\mathbb{R}$ to $\mathbb{R}$ which are identically zero outside of a finite interval (say $[0,1]$ to be specific) and nonzero inside. This is surprising when you first see it, especially if you think of functions as being defined in terms of Taylor series, because it would seem that you could compute the Taylor series at base point $a = 0$, and you get $f^{(k)}(0) = 0$ for all k, then the Taylor series should be $T(x) \equiv 0$, which converges for all x, and therefore the function must be zero for all x.

The problem with this reasoning is that the Taylor approximation theorem doesn't actually say that. Recall the Taylor Remainder Formula:

$$f(x) = f(a) + f'(a)(x-a) + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + \frac{1}{n!}\int_a^x f^{(n+1)}(t)(x-t)^n\,dt;$$

there's no real reason that integral has to go to zero as $n \to \infty$. So in general a function doesn't really need to be close to its Taylor series.

Example 13.2.1. Consider the function $f\colon \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} e^{-1/x^2} & x \neq 0, \\ 0 & x = 0. \end{cases} \tag{13.2.1}$$

Recall that we discussed this function in Example 7.2.2. Its graph is shown in Figure 13.2.

Clearly $\lim_{x\to 0} f(x) = 0$, so that f is continuous. In fact it's obvious that you could use the chain rule repeatedly to show that f is C^∞ on $(0,\infty)$ and also on $(-\infty,0)$, so the only thing we need to do to show f is C^∞ everywhere is to show that f has derivatives of all orders at $x = 0$.

The first step is to compute $f'(0)$ directly from the definition, and observe that

$$f'(0) = \lim_{h\to 0}\frac{f(h)}{h} = \lim_{h\to 0}\frac{e^{-1/h^2}}{h} = \lim_{y\to\infty}\frac{e^{-y^2}}{1/y} = \lim_{y\to\infty}\frac{y}{e^{y^2}} = \lim_{y\to\infty}\frac{1}{2ye^{y^2}} = 0,$$

where we used the substitution $y = \frac{1}{h}$ and L'Hopital's rule. So f' exists, and then we need to show that it's a continuous function. We have

$$\lim_{x\to 0} f'(x) = \lim_{x\to 0}\frac{2e^{-1/x^2}}{x^3} = \lim_{y\to\infty}\frac{2y^3}{e^{y^2}} = 0.$$

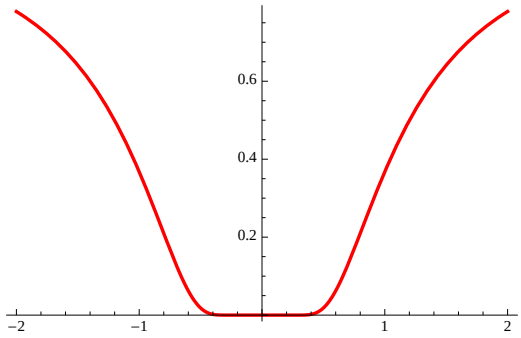

FIGURE 13.2. The graph of $f(x) = e^{-1/x^2}$. Despite the apparent singularity in the formula, it is smooth at $x = 0$, and all derivatives are equal to zero.

So f' is continuous, and the next step is to use the definition to show that $f''(0)$ exists.

We can do all the other steps at once if we notice that inductively

$$f^{(n)}(x) = \sum_{k=1}^{n} \frac{a_{k,n}}{x^{2k+n}} e^{-1/x^2}$$

for some coefficients $a_{k,n}$ if $x \neq 0$, since assuming that leads to

$$f^{(n+1)}(x) = \sum_{k=1}^{n} \frac{-(2k+n)a_{k,n}}{x^{2k+n+1}} e^{-1/x^2} + \sum_{k=2}^{n+1} \frac{2a_{k-1,n}}{x^{2k+n+1}} e^{-1/x^2}.$$

Clearly if we replace x with $1/y$ we get

$$\lim_{x\to 0} f^{(n)}(x) = \sum_{k=1}^{n} a_{k,n} \lim_{y\to\infty} \frac{y^{2k+n}}{e^{y^2}} = 0$$

and also

$$\lim_{h\to 0} \frac{f^{(n)}(h)}{h} = \sum_{k=1}^{n} a_{k,n} \lim_{y\to\infty} \frac{y^{2k+n+1}}{e^{y^2}} = 0$$

so that $f^{(n+1)}(0) = 0$. Inductively then, $f^{(n)}$ is differentiable everywhere and $f^{(n+1)}$ is continuous. So f is C^∞. ☺

Now if we define

$$g(x) = \begin{cases} e^{-1/x^2} & x > 0, \\ 0 & x \leq 0, \end{cases}$$

then we get a C^∞ function which is identically zero on half the line. Furthermore if we define

$$h(x) = g(x)g(1-x),$$

we get a C^∞ function which is positive on $(0,1)$ and zero everywhere else. This h is a common sort of bump function. Its graph is shown in Figure 13.3.

This is just the beginning. Clearly the function

$$j(x) = \frac{\int_0^x h(\sigma)\,d\sigma}{\int_0^1 h(\sigma)\,d\sigma}$$

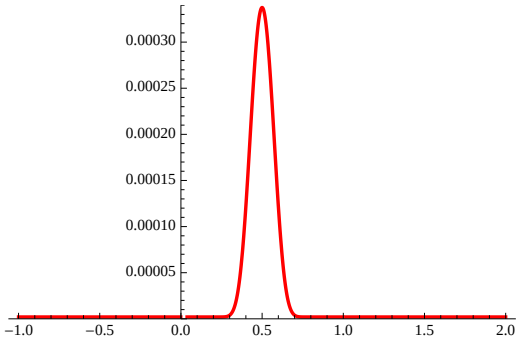

FIGURE 13.3. The function defined by $h(x) = e^{-1/x^2-1/(x-1)^2}$ for $0 < x < 1$ and $h(x) = 0$ otherwise. It is C^∞ and nonzero if and only if $0 < x < 1$.

is C^∞ (its denominator is a positive constant) and satisfies $j(x) = 0$ for $x \leq 0$, $j(x) = 1$ for $x \geq 1$, and $j(x)$ is strictly increasing on $(0, 1)$. Figure 13.4 shows the graph.

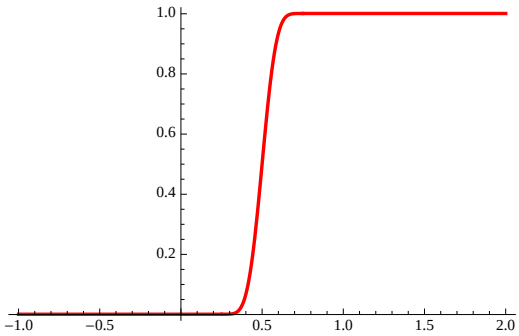

FIGURE 13.4. The function $j(x)$ defined by integrating the function $h(x)$ from Figure 13.3, then normalizing. It is zero when $x \leq 0$, one when $x \geq 1$, and strictly between zero and one on $(0, 1)$.

Next it is easy to see how to get, for any numbers $a < b < c < d$, a function k such that $k(x) = 0$ for $x \leq a$ or $x \geq d$, $k(x) = 1$ for $b \leq x \leq c$, k is strictly increasing on (a, b), and k is strictly decreasing on (c, d). We just take

$$k(x) = j\left(\frac{x-a}{b-a}\right) j\left(\frac{d-x}{d-c}\right).$$

Such a function is shown in Figure 13.5, for $a = 0$, $b = 1$, $c = 2$, and $d = 3$. Such a k is called a *cutoff function*.

Theorem 13.2.2. *If $f\colon \mathbb{R} \to \mathbb{R}$ is any C^∞ function, with $p \in \mathbb{R}$ any point and $0 < \delta < \varepsilon$ any real numbers, then there is a C^∞ function $\tilde{f}\colon \mathbb{R} \to \mathbb{R}$ such that*

$$\tilde{f}(x) = \begin{cases} f(x) & |x-p| < \delta, \\ 0 & |x-p| > \varepsilon. \end{cases}$$

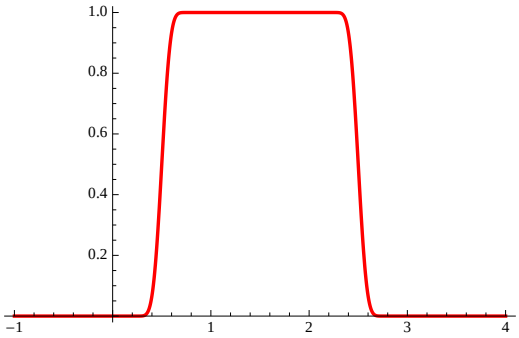

FIGURE 13.5. The C^∞ function $k(x) = j(x)j(3-x)$, which is identically zero outside $(0,3)$ and identically equal to one inside $[1,2]$, while strictly increasing on $(0,1)$ and strictly decreasing on $(2,3)$.

Proof. Clearly all we do is choose $a = p - \varepsilon$, $b = p - \delta$, $c = p + \delta$, and $d = p + \varepsilon$, construct the function k as above, and then set $\tilde{f} = kf$. □

Then in a small neighborhood of p, $\tilde{f}$ is indistinguishable from f, but $\tilde{f}$ is indistinguishable from the zero function outside a slightly larger neighborhood of p. It is obvious how to extend this from open intervals in the line to open cubes in $\mathbb{R}^n$; we just take the product

$$\tilde{f}(x^1, \cdots, x^n) = f(x^1, \cdots, x^n)k(x^1) \cdots k(x^n).$$

We could also do it with balls instead of cubes by taking

$$\tilde{f}(x^1, \cdots, x^n) = f(x^1, \cdots, x^n)k\left(\sqrt{(x^1)^2 + \cdots + (x^n)^2}\right).$$

This is an extremely common trick in differential geometry (and even moreso in fields like functional analysis) for "localizing" a function. We have some function, and we want to do a bunch of computations with it near some point, but we don't want to worry about its global behavior. Well we could just assume it's identically zero outside some small open set, but we can still do all the computations we want in a smaller open set near the desired point.

This shows us how to prove Theorem 13.1.4. Suppose we have some smooth function f defined in an open coordinate neighborhood U of a point p, where $\phi\colon U \to \mathbb{R}^n$ is the coordinate chart. Typically f would be some function of the coordinate functions. We want to think of f as defined on all of M, so we would just take an open cube $K_1 \subset \mathbb{R}^n$ and a larger open cube $K_2 \subset \mathbb{R}^n$, both of which contain the point $0 = \mathbf{x}(p)$. Define g on $\phi[U] = \mathbb{R}^n$ so that $g = f \circ \phi$ inside the small cube K_1 and $g = 0$ outside the large cube K_2. Then $\tilde{f} = g \circ \phi$ defines a function on U which agrees with f inside $V = \phi^{-1}[K_1]$ and is zero outside $\tilde{U} = \phi^{-1}[K_2]$. We might as well define $\tilde{f}$ to be zero everywhere else on the manifold, because then $\tilde{f}$ will clearly still be C^∞ on the entire manifold.

Hence if we're only interested in the behavior of a function on a small open set near the point, we can make it equal to whatever we want on that open set and zero outside a slightly larger open set. We almost never care what happens in between the small open set and the large open set, although if we ever needed to know, we

could say that it just involves multiplying the function by a positive number less than one. (So for example if we started with a positive function on the small open set, we'd know we'd have a nonnegative function on the entire manifold; if we had a function bounded between -1 and 1 on the small open set, the function would be bounded between -1 and 1 on the entire manifold, etc.)

13.3. **Partition of unity.** The problem with the previous construction is that you end up with functions and other objects that are only nonzero on a small open set. For various reasons, you may want objects that are nowhere zero or at least mostly nonzero. One example is in proving the existence of a smooth inner product which is positive-definite everywhere on the manifold (a Riemannian metric), which is used to prove for example that the cotangent bundle is bundle-isomorphic to the tangent bundle. Another example is in constructing a notion of integration of functions on a compact manifold.

The tool to do this is to have a family of coordinate charts (φ_k, U_k), and a positive real-valued function $\xi_k \colon M \to \mathbb{R}$ such that the support[26] of ξ_k is in U_k for each k. We want to require that for every p there is a ξ_k such that $\xi_k(p) > 0$, and also that for each p there are only finitely many ξ_k such that $\xi_k(p) \neq 0$. These requirements ensure that for every point p, the sum $\sum_k \xi_k(p)$ is positive and finite, which means we can normalize to assume that $\sum_k \xi_k(p) = 1$ for every p. Hence the name "partition of unity": we divide up the number "one" into finitely many positive weights at each point p, the weights telling you how much of each coordinate chart you're using to define your object.

Example 13.3.1. The circle S^1 can of course be covered by two coordinate charts (north-pole and south-pole stereographic coordinates), as in Example 7.1.7. Call them $\phi \colon U \to \mathbb{R}$ and $\psi \colon V \to \mathbb{R}$ as before.

Let $k \colon \mathbb{R} \to \mathbb{R}$ be a C^∞ bump function such that $k(u) > 0$ iff $|u| < 2$ and $k(u) = 0$ if $|u| \geq 2$. Define $\zeta_i \colon S^1 \to \mathbb{R}$ by $\zeta_1(p) = k(\phi(p))$ if $p \in U$ and $\zeta_1(s) = 0$ at the south pole, and similarly $\zeta_2(p) = k(\psi(p))$ if $p \in V$ and $\zeta_2(n) = 0$ at the north pole. Using the formula $\phi(x, y) = x/(y + 1)$ for the stereographic coordinates, we see that ϕ maps the upper closed semicircle to $[-1, 1]$ and thus ζ_1 is strictly positive on the upper closed semicircle. Similarly ζ_2 is strictly positive on the lower closed semicircle. Therefore $\zeta_1 + \zeta_2$ is positive everywhere on S^1. Define $\xi_1 = \zeta_1/(\zeta_1 + \zeta_2)$ and $\xi_2 = \zeta_2/(\zeta_1 + \zeta_2)$. Then (ξ_1, U) and (ξ_2, V) give a partition of unity since at every point either ξ_1 or ξ_2 is positive, and the sum of the two is always equal to one. The two functions are graphed in Figure 13.6.

☺

As one example of how to use a partition of unity, let's suppose we want to define a smooth positive-definite inner product on M (i.e., a Riemannian metric). This is a function taking a point p in M to a symmetric positive-definite tensor $g(p)$ of type $(2, 0)$ on T_pM, such that in any coordinate chart $(\mathbf{x}, U)$ the function $p \mapsto g(p)(\frac{\partial}{\partial x^i}\big|_p, \frac{\partial}{\partial x^j}\big|_p)$ is smooth on U. How do we know such a g exists? If we have a partition of unity, we can define for each k an inner product g_k on U_k: just set $g_k(p)(\frac{\partial}{\partial x^i}\big|_p, \frac{\partial}{\partial x^j}\big|_p = \delta_{ij}$, in other words use the Euclidean inner product on each

[26]The *support* of a continuous function is the closure of the set where the function is nonzero. Supports are usually interesting only when the function is expected to be identically zero on a large set.

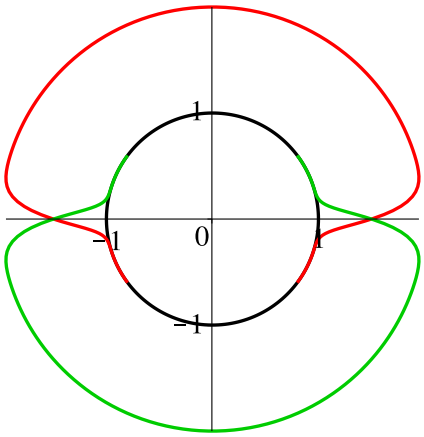

FIGURE 13.6. The partition of unity constructed in Example 13.3.1; the functions are shown in red and green, and adding the values of them gives one everywhere.

coordinate chart. The g_k will of course look different in another coordinate chart, but it will still be smooth and positive-definite on the overlap. Then define

$$g(p) = \sum_k \xi_k(p) g_k(p).$$

Then since each ξ_k is positive, we know $g(p)(u,u) = \sum_k \xi_k(p) g_k(p)(u,u) > 0$, so that g is positive-definite at each point p. Since the ξ_k are smooth, we know g is smooth as well. And thus every manifold that has a partition of unity has a smooth positive-definite inner product.

It's easy to use the same technique as in Example 13.3.1 to construct a partition of unity on any compact manifold.

Theorem 13.3.2. *If M is a compact n-dimensional manifold, then there is a C^∞ partition of unity on M, such that each of the nonnegative functions ξ_k has support in a coordinate chart.*

Proof. Just observe that we can cover M with finitely many charts (φ_k, U_k) for $1 \leq k \leq N$. Now we want to find some nice cubes in each U_k that are large enough so that the cubes also cover M. This can clearly be done: for each k and $m > 0$ let C_{km} be the inverse image of the cube $(-m, m)^n$ under φ_k; then the family C_{km} forms an open cover of M, and since M is compact we only need finitely many of these open sets homeomorphic to cubes to cover M. Clearly by taking the union of finitely many cubes associated to each k, we can replace them all with a single cube $C_k \equiv C_{km}$ for each k.

For each cube $\varphi_k[C_k]$, construct a C^∞ function τ_k on $\mathbb{R}^n$ which is identically one on $\varphi_k[C_{km}]$ and identically zero outside of the slightly larger cube $\varphi_k[C_{k,m+1}]$. Then $\psi_k = \tau_k \circ \varphi_k$ is a smooth function on U_k which is zero outside of $C_{k,m+1}$, and so we can extend it to a C^∞ function on all of M by just defining it to be zero outside U_k.

For each point p, there is some C_k such that $p \in C_k$ since these cubes cover M, and by definition we have $\psi_k(p) > 0$. Also there are only finitely many ψ_k since we

only needed finitely many C_k to cover M. So we can just define

$$\xi_k(p) = \frac{\psi_k(p)}{\sum_i \psi_i(p)},$$

and we see that the functions ξ_k add up to one at each point. □

More generally any Hausdorff and second-countable manifold has a partition of unity, but this requires a bit more work. (It uses local compactness and the basic ideas of the proof in the compact case.)

13.4. **Whitney embedding.** The nicest application of a partition of unity is in proving that every compact smooth n-dimensional manifold is homeomorphic to a smooth submanifold of some Euclidean space $\mathbb{R}^N$. This proof is taken from Spivak's *Comprehensive introduction to differential geometry, volume 1.*

Theorem 13.4.1. *If M is a compact C^∞ manifold, then there is a smooth embedding $f\colon M \to \mathbb{R}^d$ for some positive integer d.*

Proof. Choose finitely many coordinate charts, with some finite number K of cube subsets C_k for $1 \leq k \leq K$ covering M, exactly as in the proof of Theorem 13.3.2, with coordinate charts (φ_k, U_k). Then we have functions ψ_k such that $\psi_k(p) = 1$ whenever $p \in C_k$.

Now define $d = nK + K$ and define $f\colon M \to \mathbb{R}^d$ by

$$f = (\psi_1, \cdots, \psi_K, \psi_1\varphi_1, \cdots, \psi_K\varphi_K).$$

First prove that f is an immersion by showing it has rank n at each point p. To do this, find a C_k containing p; then $\psi_k\varphi_k = \varphi_k$ on C_k, so $f \circ \varphi_k^{-1}$ looks like just the identity map $(x^1, \cdots, x^n)$ (along with a bunch of other terms in the other rows). Hence $D(f \circ \varphi_k^{-1})$ has an $n \times n$ submatrix which is the identity, and so in particular Df has rank n. We have thus proved that f is an immersion.

To show that f is an embedding, we need to know that f is a homeomorphism onto its image. But the domain of f is compact, and any one-to-one continuous map from a compact space to a Hausdorff space is always a homeomorphism onto its image by general topology. So we are done once we show that f is one-to-one.

To do this, suppose that $f(p) = f(q)$. Find a cube C_k such that $p \in C_k$; then $\psi_k(p) = 1$, and therefore $\psi_k(q) = 1$ also (using the first K coordinates of f). But the only points where ψ_k equals one are points in C_k, so that q is in the same cube as p. Now using the later coordinates of f, we see that $\psi_k(p)\varphi_k(p) = \psi_k(q)\varphi_k(q)$, and since $\psi_k(p) = \psi_k(q) = 1$, we actually have $\varphi_k(p) = \varphi_k(q)$. But φ_k is a coordinate chart on C_k, which means in particular that it's one-to-one, and hence $p = q$. □

The dimension of the Euclidean space d is much larger than the dimension of the actual manifold n, in this proof. Actually one doesn't need nearly so much. Generally there is a minimal dimension $d = e(n)$ such that any n-dimensional manifold is homeomorphic to a submanifold of $\mathbb{R}^d$. Clearly the circle embeds in $\mathbb{R}^2$ but not $\mathbb{R}^1$, so that $e(1) = 2$. We know that the projective plane cannot be embedded in $\mathbb{R}^3$ (because any compact two-dimensional manifold in $\mathbb{R}^3$ is orientable), but there is an embedding into $\mathbb{R}^4$. All other two-dimensional manifolds also embed into $\mathbb{R}^4$ (since we know them all), so that $e(2) = 4$. C.T.C. Wall[27] proved that $e(3) = 5$.

[27] "All 3-manifolds imbed in 5-space," Bull. AMS vol. 71

In general $e(n) \leq 2n-1$ unless $n = 2^r$ for some r (and as an explicit counterexample, the real projective spaces $\mathbb{P}^{2^r}$ never embed in $\mathbb{R}^{2^{r+1}-1}$). The optimal dimension is not known in general, but it is related to Stiefel-Whitney classes in algebraic topology; for much more on the subject see Osborn[28].

Here I will demonstrate the "easy" version of the Whitney embedding theorem, which states that $e(n) \leq 2n+1$. The basic idea is to first use Theorem 13.4.1 to reduce to the case where you have a submanifold of *some* Euclidean space, and then show that if d is large, you can find a direction which is not parallel to any chord or tangent vector of M. Then the map from M to the hyperplane $\mathbb{R}^{d-1}$ is still a one-to-one immersion, and hence it is an embedding. The proof is based on an exercise in Spivak's Volume 1.

Roughly speaking, the point of the next theorem is that on an n-dimensional space, $2n+1$ "generic" functions will be independent. An intuitive way to see this is looking at the ways draw a curve in $\mathbb{R}^2$. It's quite easy to have the curve end up crossing itself unless you draw it carefully. However when drawing a curve in $\mathbb{R}^3$ it's very rare that the curve will cross itself, and if it does you can always perturb it slightly to remove the crossing. (No small perturbation of a figure-eight in $\mathbb{R}^2$ will give you a simple closed curve, but a small perturbation of a figure-eight in $\mathbb{R}^3$ will.) The same idea gets used in other contexts as well.

In the proof we will need a version of Sard's Theorem, which is the basic explanation of where this dimension count comes from. It says that for any smooth function from $\mathbb{R}^n$ to itself, the set of critical values is "small." (The set of critical points may be large.) For example if $f(x,y) = (x,0)$, the critical points are all of $\mathbb{R}^2$, but the critical values are just the horizontal line $\mathbb{R}$, which is only a small portion of $\mathbb{R}^2$.

Theorem 13.4.2. *Suppose M is a smooth connected n-dimensional manifold and M is connected. If $f\colon M \to \mathbb{R}^n$ is a smooth function, then the critical values of f form a set of Lebesgue measure zero in $\mathbb{R}^n$.*

Proof. The basic idea is that if we have a set B where the rank of f is less than maximal, then on that set we have that $\det Df(x) = 0$. Now this set B may be complicated, but it's the countable intersection of the open sets $U_k = \{x \,|\, |\det Df(x)| < \frac{1}{k}\}$, which are fairly simple. Take a finite-volume cube C; then the volume of the image $f[U_k \cap C]$ is (by the change of variables formula Theorem 5.3.2)

$$\operatorname{vol}(f[U_k \cap C]) = \int_{U_k \cap C} |\det Df(x)|\, dx \leq \frac{1}{k} \operatorname{vol}(U_k \cap C).$$

Take the intersection over all k and we conclude that the volume of $f[B \cap C]$ is zero whenever C is a cube of finite volume. But all of $\mathbb{R}^n$ is a countable union of cubes of finite volume, so we get a countable union of sets of measure zero, which still has measure zero. □

We can extend Sard's Theorem to smooth maps from one n-dimensional manifold M to another n-dimensional manifold N; we say that a set E in N has measure zero if, for every coordinate chart (φ, U) on N, the set $\varphi[E]$ has Lebesgue measure zero in $\mathbb{R}^n$. This gives a consistent definition since N is second-countable, and hence there are only countably many coordinate charts needed.

[28]Vector bundles: Volume 1, Foundations and Stiefel-Whitney classes

The result is that if we have a map $f\colon M \to N$, it is very easy to find a point $q \in N$ which is a regular value (that is, either $f^{-1}(q)$ is empty, or $f^{-1}(q)$ is an $(n-1)$-dimensional submanifold of M), since the set of all other q has measure zero in any reasonable way of measuring. By the way, this is the reason that q is still called a regular value of f even if $f^{-1}(q)$ is empty: because this definition makes the set of critical values small and makes the statement of Sard's theorem easier.

The more general version of Sard's theorem is that if M has dimension m and N has dimension n, and $f\colon M \to N$ is a smooth function, then the image of the set of points where the rank of Df is less than n has measure zero in N. If $m < n$ then every point of M has rank less than n, so it says that $f[M]$ has measure zero in N. If $m > n$ then for almost every $q \in N$ where $f^{-1}(q)$ is nonempty, the restriction $f\big|_{f^{-1}(q)}$ is a submersion. This more general version is proved in the same sort of way, though it's a bit harder. We will skip it here.

Now we show how to reduce the dimension d of the ambient space $\mathbb{R}^d$ obtained from Theorem 13.4.1.

Theorem 13.4.3. *Suppose M is an n-dimensional smooth compact submanifold of $\mathbb{R}^d$, where $d > 2n+1$ and $\iota\colon M \to \mathbb{R}^d$ is the embedding. Then there is a unit vector v in $\mathbb{R}^d$ such that for every pair of points $p, q \in M$, the vector $\iota(q) - \iota(p)$ is not parallel to v, and also for every point p the tangent plane $\iota_*[T_pM]$ does not contain v. Hence if $v^\perp$ is the hyperplane through the origin perpendicular to v and $\pi_{v^\perp}$ is the projection of $\mathbb{R}^d$ onto $v^\perp$, then $\pi_{v^\perp} \circ \iota\colon M \to v^\perp \cong \mathbb{R}^{d-1}$ is an embedding.*

Proof. In $M \times M$, the diagonal set is the set $D = \{(p,p) \,|\, p \in M\}$. Let $\iota\colon M \to \mathbb{R}^d$ be the embedding. We want to show there is a vector $v \in \mathbb{R}^d$ so that $\pi_{v^\perp} \circ \iota\colon M \to \mathbb{R}^{d-1}$ is still an embedding, which will happen if it is an immersion which is one-to-one. So we want a vector v such that $\iota(p) - \iota(q)$ is never parallel to v for $p \neq q$ (for then $\pi_{v^\perp}(\iota(p)) \neq \pi_{v^\perp}(\iota(q))$ for $p \neq q$) and such that $\iota_*(u)$ is not parallel to v for any $u \in TM$ (for then $\pi_{v^\perp} \circ \iota_*$ has a trivial kernel).

Let $U = (M \times M) \backslash D$; since D is a closed subset of $M \times M$, we know that U is an open subset of $M \times M$ and hence a manifold of dimension $2n$. Also since ι is an embedding, the map $\iota(p) - \iota(q)$ is never the zero vector if $(p,q) \in U$. Hence we can define $G\colon U \to S^{d-1}$ by

$$G(p,q) = \frac{\iota(p) - \iota(q)}{\|\iota(p) - \iota(q)\|},$$

and this map is smooth. So G is a smooth map from a manifold of dimension $2n$ to a manifold S^{d-1} of dimension $d-1$. Now $d - 1 > 2n$ by assumption, and that implies that the image $G[U]$ has measure zero in S^{d-1}, by Sard's Theorem.

Hence there are lots of vectors $v \in S^{d-1}$ which are not parallel to $\iota(p) - \iota(q)$ for any distinct p and q. So if $v \in S^{d-1} \backslash G[U]$, then $\pi_{v^\perp} \circ \iota$ is one-to-one.

Now TM is also a manifold of dimension $2n$. Let $V \subset TM$ be

$$V = \{u \,|\, u \in T_pM \text{ for some } p \in M, u \neq 0\}.$$

Then V is an open subset of TM, so it is also a $2n$-dimensional smooth manifold. Again we consider the map $H\colon V \to S^{d-1}$ defined by

$$H(v) = \frac{\iota_*(u)}{\|\iota_*(u)\|},$$

which is smooth on V. So the image $H[V]$ has measure zero in S^{d-1} by Sard's Theorem.

Now choose a vector $v \in S^{d-1}$ such that $v \notin G[U] \cup H[V]$. Let $\pi\colon \mathbb{R}^d \to v^\perp \cong \mathbb{R}^{d-1}$ be orthogonal projection

$$\pi(u) = u - (u \cdot v)v.$$

Consider the map $\pi \circ \iota\colon M \to v^\perp$. Since $\iota(p) - \iota(q)$ is never parallel to v if $p \neq q$, the orthogonal projection $\pi_{v^\perp}$ never takes $\iota(p)$ and $\iota(q)$ to the same point, so that $\pi_{v^\perp} \circ \iota$ is one-to-one. Furthermore since $\iota_*[T_pM]$ never contains a vector parallel to v, we know $\pi_{v^\perp} \circ \iota_*$ always has a trivial kernel, and thus $\pi_{v^\perp} \circ \iota$ is an immersion.

We conclude that $\pi \circ \iota$ is an embedding, since M is compact and $\pi \circ \iota$ is both an immersion and one-to-one. □

Corollary 13.4.4. *Every compact n-dimensional smooth manifold M can be embedded in $\mathbb{R}^{2n+1}$.*

Proof. Embed M into some high-dimensional Euclidean space $\mathbb{R}^d$ using Theorem 13.4.1. If $d > 2n+1$ then use the technique of Theorem 13.4.3 to find a subspace $\mathbb{R}^{d-1}$ into which M can still be embedded. Then keep going; this process stops working once you get down to $\mathbb{R}^{2n+1}$. □

Thus if you understand the proof of Theorem 13.4.3, you see that we only ever used Sard's theorem. If you wanted to do better (as Whitney originally did), you can go beyond this to actually prove that the maps G and H cannot cover all of S^{2n}, but then you have to use some more delicate topological arguments (since you're dealing with a map from one $2n$-dimensional manifold to another $2n$-dimensional manifold), rather than the brute force instrument that is Sard's theorem.

Since we never specified anything particular about how the embedding into high-dimensional space had to work, it really doesn't matter what we use to get the embedding, as long as we have something to start with. And what is an embedding, really? It's just a collection of $2n+1$ real-valued functions $f_1, \cdots, f_{2n+1}\colon M \to \mathbb{R}$ whose derivatives span an n-dimensional space and don't all agree at any point. One can then try to think of this directly: if one just picks $2n+1$ smooth functions on an n-dimensional smooth manifold, do they give you an embedding? Intuitively if you were picking functions "randomly," with probability one you'd pick some which led to an embedding. You can make this precise (for example, by proving the set of functions which don't work is of first Baire category in the right topology), but the idea is good enough for now.

The power of this idea is that you can use it for other things as well. Takens[29] proved a famous theorem in dynamical systems that basically says you can use a generic time series to embed an attractor into a Euclidean space. The way it works is, suppose you have a smooth map $\phi\colon M \to M$ (which is a discrete dynamical system; you're interested in things like the behavior of the sequence $(p, \phi(p), \phi(\phi(p)), \phi(\phi(\phi(p))), \ldots)$, and you take an observable $h\colon M \to \mathbb{R}$, then as long as you're not extremely unlucky, the map

$$p \mapsto (h(p), h(\phi(p)), h(\phi(\phi(p))), \cdots, h(\phi^{\circ 2n}(p)))$$

[29] "Detecting strange attractors in turbulence," Springer Lecture Notes vol. 898

will be an embedding of the configuration space M into $\mathbb{R}^{2n+1}$; this essentially just comes from Sard's theorem. The sequence $h(p), h(\phi(p)), h(\phi(\phi(p))), \cdots$ is called a "time series."

Why might you care about this? Well, typically you have some dynamical system, and you don't know everything that's going on. Maybe you have a bunch of data on stock prices of a company. Now the stock price of a particular company at any given time depends on a bunch of different factors, and you have no idea what they are. But maybe there are only a finite number of factors, and if so you can hope that just plotting enough of these stock prices (with one dimension representing stock price at time 8:00 am, another dimension representing 8:05 am, etc.) will give you some picture of the space of all stock prices. Of course to use this you need the stock prices of *all* companies, and possibly a bunch of other data (the volume of treasury bill sales, etc.) to really capture all of the space M. Obviously you can't hope to do that, so this isn't really useful by itself.

On the other hand, frequently dynamical systems can exhibit chaos. One form of chaos is a "strange attractor," and a heuristic idea of a strange attractor is that you take one point p in the set, follow the trajectory $p, \phi(p), \phi(\phi(p)), \ldots$, and watch all these points fill up some set. In non-chaotic systems they may fill up a nice curve, or all approach a point, or go out to infinity. In a chaotic system they may fill up a surface. The first and most famous example is the Lorenz attractor, which is what you get from solutions of the differential equation

$$\begin{aligned} \tfrac{dx}{dt} &= 10(y - x) \\ \tfrac{dy}{dt} &= x(28 - z) - y \\ \tfrac{dz}{dt} &= xy - \tfrac{8z}{3}. \end{aligned}$$

Pick some initial condition (x_0, y_0, z_0) and plot the trajectory in $\mathbb{R}^3$; you end up with the image in Figure 13.7, where a curve seems to fill up some portion of a two-dimensional surface. (Actually the Lorenz attractor is a fractal, so it's not smooth, and its Hausdorff dimension is slightly more than two.) The idea is that if you picked *any* point $p = (x_0, y_0, z_0)$ that was actually *in* that set, its trajectory should fill up the same space.

Now if that's the case, then I don't need to take a bunch of short time series with a bunch of different starting points p. I can take a long time series with *one* starting point p, and choose some N, choose some h, and plot the points that show up periodically. Let's say $N = 3$ for simplicity, and I had a time series $(h_1, h_2, h_3, h_4, \cdots)$ which I assume comes from $h(\phi^{\circ k}(p))$ for some function h, dynamical system ϕ, and point p. Then I'd just plot the points

$$(h_1, h_2, h_3), (h_4, h_5, h_6), (h_7, h_8, h_9), \cdots$$

until I ran out of patience. I don't expect this to tell me what the entire space M is, but I do expect it to show me a picture of the strange attractor if the dimension of the strange attractor is 1. (I'd expect to see the points all clustering into a nice curve.) If I had a 2-dimensional strange attractor, then this process might just give me a jumble of points with no real structure, because my dimension N is too low to get an embedding.

If I thought the dimension of the strange attractor were higher, I'd just increase N, until I started detecting some kind of nice structure. The nice thing about this is that I really only need one time series (e.g., the price of a stock at all times),

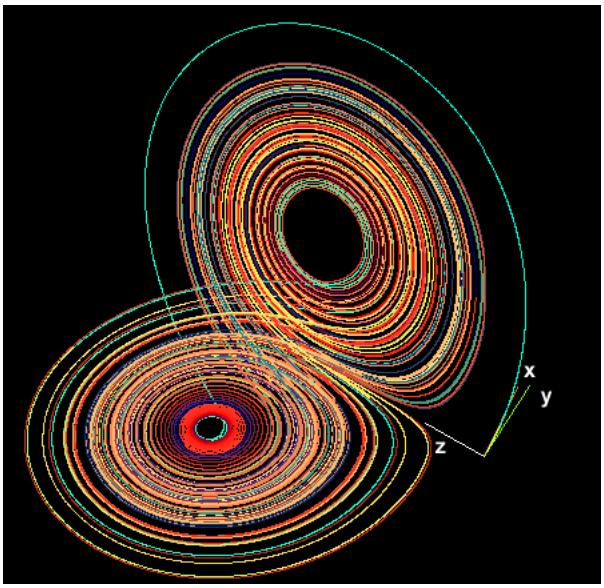

FIGURE 13.7. A single trajectory of the Lorenz differential equations appears to fill up a fractal of dimension slightly more than two in space.

and I can find some hidden structure in the prices that other people can't see, then take advantage of it and make lots of money. Of course, if it were this easy we'd all do it, and so there are complications, but that's another story.

14. Vector fields and differential equations

"Calculate every possible destination along their last known trajectory."

14.1. **Vector fields as derivations.** We defined vector fields in Definition 12.1.7, but we didn't do much with them then: our main concern was whether they existed and whether they had zeroes. But in differential geometry they are very useful for two purposes: as differential equations which generate a flow of diffeomorphisms, and as operators which differentiate smooth functions.

- As we will see, the proper way to think of an ordinary differential equation is in terms of vector fields. This is already intuitive in basic calculus: we can visualize a first-order differential equation $y'(x) = F\big(x, y(x)\big)$ as being a family of arrows on the plane, whose solution is obtained by starting at some initial value and walking in the direction of the arrows. You've probably seen pictures of such slope fields in a differential equations class. One is shown in Figure 14.1.

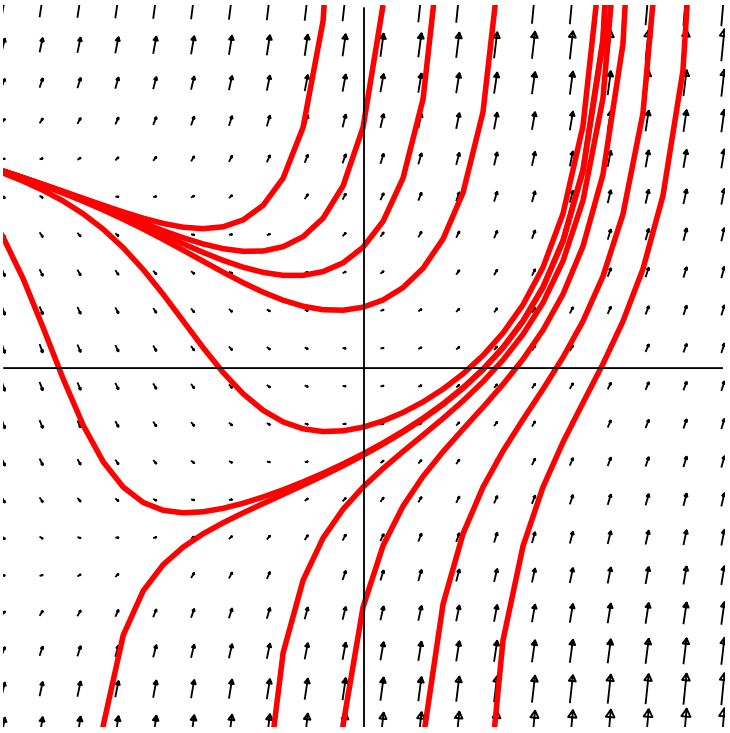

Figure 14.1. The vector field plot representing the differential equation $\frac{dy}{dx} = x + y^2$, along with some solution curves.

- Tangent vectors differentiate functions at a single point; thus they take real-valued functions to real numbers. Vector fields can differentiate the entire function everywhere at once, and thus they take real-valued functions to

real-valued functions. Any space (in particular the space of smooth real-valued functions on a manifold) can be understood better by looking at operators from that space to itself.

Of course the existence of vector fields on general manifolds relies on the constructions of Chapter 13, using a bump function. The following proposition follows using the same proof as that of Theorem 13.1.4 in Section 13.2.

Proposition 14.1.1. *Let M be a smooth manifold and $U \subset M$ a subset of M on which TU is trivial. Let $X\colon U \to TU$ be a smooth section. Then for any $p \in M$ there is a neighborhood V of p and a vector field $\tilde{X}\colon M \to TM$ such that $\tilde{X}$ agrees with X on V.*

Since TU is trivial, there are smooth vector fields $E_1, \dots, E_n\colon U \to TU$ such that $\{E_1(p), \dots, E_n(p)\}$ is a basis for T_pM at each $p \in U$. Hence we obtain every vector field on U as a linear combination of these fields:

$$X(p) = \sum_{k=1}^{n} f^k(p)\, E_k(p)$$

for some smooth functions $f^k\colon U \to \mathbb{R}$. Such a vector field gets extended to $\tilde{X}$ by using a bump function which is identically one in V and identically zero outside U.

Recall that a vector field on M is smooth if it is smooth in charts on M and TM. The following criteria are frequently more useful.

Proposition 14.1.2. *Let M be an n-dimensional manifold with tangent bundle TM, and let $\pi\colon TM \to M$ be the projection. A map $X\colon M \to TM$ with $\pi \circ X = id$ is smooth if and only if in any coordinate chart (ϕ, U) with $X|_U$ expressed as*

$$X(p) = \sum_{k=1}^{n} a^k(p)\, \frac{\partial}{\partial x^k}\Big|_p$$

the functions $a^k\colon U \to \mathbb{R}$ are smooth.

Proof. In coordinates (ϕ, U) on M and (Φ, TU) on TM we have

$$\Phi \circ X \circ \phi^{-1}(x^1, \dots, x^n) = \big(x^1, \dots, x^n, a^1 \circ \phi^{-1}(x^1, \dots, x^n), \dots, a^n \circ \phi^{-1}(x^1, \dots, x^n)\big).$$

This is C^∞ on $\mathbb{R}^n$ if and only if each $a^k \circ \phi^{-1}\colon \mathbb{R}^n \to \mathbb{R}$ is C^∞, which is precisely the definition of smoothness for $a^k \circ U \to \mathbb{R}$. □

Observe that if we have a vector field X defined on the whole space M and a smooth function $f\colon M \to \mathbb{R}$, then we get a function $X(f)\colon M \to \mathbb{R}$: we just differentiate f at each point, so that $X(f)(p) = X_p(f)$, using the characterization of vectors as differential operators from definition (10.3.1). (We sometimes write $X(p)$ as X_p precisely because this notation would otherwise be awkward.) We should check that the resulting $X(f)$ is actually smooth.[30]

Proposition 14.1.3. *Suppose X is a function from M to TM with $\pi \circ X = id$. Then X is a smooth vector field if and only if whenever f is a smooth function, $X(f)$ is also a smooth function.*

[30]This does not follow from the Chain Rule.

Proof. First let X be a smooth vector field and f a smooth function; we will show that $X(f)$ is smooth. In any coordinate chart (ϕ, U) we have $X(p) = \sum_k a^k(p) \frac{\partial}{\partial x^k}\big|_p$ for each $p \in U$, where $a^k \colon U \to \mathbb{R}$ is smooth by Proposition 14.1.2. Thus

$$X(f) \circ \phi^{-1}(x^1, \dots, x^n) = \sum_k a^k \circ \phi^{-1}(x^1, \dots, x^n) \frac{\partial (f \circ \phi^{-1})}{\partial x^k}(x^1, \dots, x^n)$$

by definition of the operator $\frac{\partial}{\partial x^k}\big|_p$. So $X(f) \circ \phi^{-1}$ is a sum of products of C^∞ functions on $\mathbb{R}^n$ and hence also C^∞, which means $X(f)$ is smooth by definition.

Conversely suppose that X is a (not necessarily smooth) function from M to TM with $\pi \circ X = \mathrm{id}$, but that whenever f is smooth then $X(f)$ is also smooth. Then we may still write $X_p = \sum_k a^k(p) \frac{\partial}{\partial x^k}\big|_p$ for all p in a coordinate chart (ϕ, U). Let f be a function which is identically zero outside U and such that $f = \phi^j$ on some open neighborhood V of a particular point. Then $\frac{\partial}{\partial x^k}\big|_p(f) = \frac{\partial \phi^j \circ \phi^{-1}}{\partial x^k}\big|_p = \delta^j_k$, which means that $X(f)(p) = a^j(p)$ everywhere on V. Since $X(f)$ is smooth on M, a^j is smooth on V, and since there is such a V around every point of U, we see that a^j is smooth on U. Hence by Proposition 14.1.2 we know that X is a smooth vector field. □

Thus a smooth vector field X is a linear operator from the (infinite-dimensional) vector space of C^∞ functions to itself, $f \mapsto X(f)$. Such operators are completely characterized by the product rule.

Proposition 14.1.4. *If X is a vector field on M, then for any smooth functions $f, g \colon M \to \mathbb{R}$, we have the product rule*

$$X(fg) = fX(g) + gX(f). \tag{14.1.1}$$

Conversely, if D is a linear operator from the space of smooth functions to itself which satisfies the product rule (14.1.1)*, then there is some vector field X such that $D(f) = X(f)$ for all smooth $f \colon M \to \mathbb{R}$.*

Proof. First let X be a smooth vector field with f and g smooth functions on M. Let $p \in M$, and let $v = X(p) \in T_pM$. Then by definition (10.3.1), we have for any representative curve $\gamma \colon (-\varepsilon, \varepsilon) \to M$ of v that

$$\begin{aligned} v(fg) = \frac{d}{dt}\big(f(\gamma(t))g(\gamma(t))\big)\Big|_{t=0} &= f(\gamma(0)) \frac{d}{dt} g(\gamma(t))\Big|_{t=0} + g(\gamma(0)) \frac{d}{dt} f(\gamma(t))\Big|_{t=0} \\ &= f(p)v(g) + g(p)v(f). \end{aligned}$$

We conclude that $X_p(fg) = f(p)X_p(g) + g(p)X_p(f)$ for every p, which implies (14.1.1).

Now suppose D is a linear operator from smooth functions to smooth functions satisfying (14.1.1). The first thing we want to show is that for any point p, the number $D(f)(p)$ depends only on the values of f near p, so that we can use a bump function to localize (and thus assume that f is supported in a coordinate chart). To do this, suppose h is a function which is identically zero in some open set U containing p (but possibly nonzero outside it). Choose a smooth bump function ξ on M such that $\xi = 1$ on an open set $W \ni p$ and $\xi = 0$ on the complement of U. Then $h\xi \equiv 0$ since for any point $q \in M$, either $q \in U$ and thus $h(q) = 0$, or $q \notin U$ and thus $\xi(q) = 0$. We thus have $\xi(p)D(h)(p) + h(p)D(\xi)(p) = 0$; since $h(p) = 0$ and $\xi(p) = 1$, we conclude that $D(h)(p) = 0$.

Thus when trying to compute $D(f)(p)$ for some $f\colon M \to \mathbb{R}$, we can instead compute $D(\tilde{f})(p)$ where $\tilde{f}$ is also smooth on M, equal to f in some neighborhood of p, and identically zero outside a coordinate chart, for the argument above shows that $D(f-\tilde{f})(p) = 0$. Now in a coordinate chart (ϕ, U) around p satisfying $\phi(p) = \mathbf{0}$, we can write

$$f(q) = a + \sum_{k=1}^{n} \phi^k(q) g_k(q) \tag{14.1.2}$$

where $a = f(p)$ is a constant and $g_k\colon U \to \mathbb{R}$ are smooth functions given explicitly by

$$g_k(q) = \int_0^1 \frac{\partial(f \circ \phi^{-1})}{\partial x^k}\Big(t\phi^1(q), \ldots, t\phi^n(q)\Big)\, dt. \tag{14.1.3}$$

The reason is that if $\tilde{f} = f \circ \phi^{-1}$ on $\mathbb{R}^n$, then

$$\tilde{f}(\mathbf{x}) - \tilde{f}(\mathbf{0}) = \int_0^1 \frac{d}{dt}\tilde{f}(t\mathbf{x})\, dt = \sum_{k=1}^{n} x^k \int_0^1 \frac{\partial \tilde{f}}{\partial x^k}(t\mathbf{x})\, dt,$$

and translating back to the manifold gives (14.1.2).

Now using (14.1.2) we get

$$D(f) = D(a) + \sum_{k=1}^{n} g_k D(\phi^k) + \phi^k D(g_k).$$

Notice that $D(a) = 0$ by the following trick: $D(a) = aD(1)$, but

$$D(1) = D(1 \cdot 1) = 1 \cdot D(1) + 1 \cdot D(1) = 2D(1)$$

so that $D(1) = 0$. Evaluating $D(f)$ at p and using $\phi^k(p) = 0$, we thus get

$$D(f)(p) = \sum_{k=1}^{n} g_k(p) D(\phi^k)(p).$$

But notice that by formula (14.1.3) we have

$$g_k(p) = \frac{\partial(f \circ \phi^{-1})}{\partial x^k}(\mathbf{0}) = \frac{\partial}{\partial x^k}\Big|_p (f).$$

Letting $a^k = D(\phi^k)(p)$ and $X_p = \sum_{k=1}^n a^k \frac{\partial}{\partial x^k}\big|_p$, we see that $D(f)(p) = X_p(f)$ for every smooth function f.

Doing the same thing for every point $p \in M$, we obtain a function $p \mapsto X_p \in T_pM$ such that $D(f)(p) = X_p(f)$ for all $p \in M$. Hence $p \mapsto X_p$ is a vector field on M if and only if it is smooth. However by assumption we know that whenever f is smooth then $D(f)$ is smooth, and since $D(f) = X(f)$ for every function f, we conclude by Proposition 14.1.3 that X is a vector field on M. □

Any linear operator on the vector space of C^∞ functions satisfying the product rule is called a *derivation*. What's nice about having done all the work to classify vector fields as derivations is that it makes it easy to decide what can or can't be a vector field. Classically a vector field was a map from $\mathbb{R}^n$ to $\mathbb{R}^n$ which gave the components in any coordinate chart, but any such map could only be valid if the components transformed in the right way under a change of coordinates. The present approach makes clear exactly which objects will end up satisfying this

coordinate-invariance without ever having to actually do it. In the next Section we will discuss some constructions that are much easier to understand as derivations.

14.2. **Constructions using vector fields.** The space of vector fields is of course a vector space over $\mathbb{R}$: given a number $a \in \mathbb{R}$ and a smooth vector field X on M, we can certainly multiply aX (by doing so in each tangent space) and get another smooth vector field. Similarly we can add vector fields together by adding their values in each tangent space. In this Section we discuss constructions that make sense on vector fields but which do not make sense on a general vector space. The characterization using derivations turns out to be extremely useful.

Proposition 14.2.1. *Suppose M is a smooth manifold and that X and Y are vector fields on M. Then $Z = [X, Y]$ defined by $Z(f) = X\big(Y(f)\big) - Y\big(X(f)\big)$ is also a vector field. This operator is called the* Lie bracket *of the vector fields, and it satisfies the identities*

$$[Y, X] + [X, Y] = 0 \quad \textit{(antisymmetry), and} \tag{14.2.1}$$

$$[[X, Y], Z] + [[Z, X], Y] + [[Y, Z], X] = 0 \quad \textit{(Jacobi identity).} \tag{14.2.2}$$

Proof. Clearly the operator $f \mapsto Z(f)$ is linear, and if f is smooth then $Z(f)$ is also smooth. So to prove that Z is a smooth vector field, we just have to check the Leibniz rule (14.1.1). The computation is easy:

$$\begin{aligned} Z(fg) &= X\big(Y(fg)\big) - Y\big(X(fg)\big) \\ &= X\big(gY(f) + fY(g)\big) - Y\big(gX(f) + fX(g)\big) \\ &= X(g)Y(f) + gX\big(Y(f)\big) + X(f)Y(g) + fX\big(Y(g)\big) - Y(g)X(f) \\ &\qquad - gY\big(X(f)\big) - Y(f)X(g) - fY\big(X(g)\big) \\ &= g[X, Y](f) + f[X, Y](g) \\ &= gZ(f) + fZ(g). \end{aligned}$$

The antisymmetry (14.2.1) is trivial, and the Jacobi identity comes from expanding and canceling: for any smooth f we have

$$\Big([[X, Y], Z] + [[Z, X], Y] + [[Y, Z], X]\Big)(f) = \Big((XYZ - YXZ - ZXY + ZYX) + (ZXY - XZY - YZX + YXZ) + (YZX - ZYX - XYZ + XZY)\Big)(f) = 0.$$

Hence the combination of these brackets is the zero operator on smooth functions, which means it must be the zero vector field. □

The space of all vector fields on M is a linear space, and under the Lie bracket it becomes a *Lie algebra*: in general a Lie algebra is a vector space with a bilinear operation $(x, y) \mapsto [x, y]$ satisfying the properties (14.2.1)–(14.2.2). These objects have very interesting properties that are widely studied on their own independently of differential geometry or manifolds, although the original motivation for them was in the study of vector fields and differential equations, as we will discuss in the next section. The space of all vector fields on a smooth manifold is an *infinite-dimensional Lie algebra*, but imposing extra conditions on it (such as only working with vector fields that respect a group symmetry of M) leads to the finite-dimensional Lie algebras which are more familiar. This will come up again later.

It may seem mysterious that the operation $[\cdot,\cdot]$ should be anything special on vector fields, but the significance of the characterization Proposition 14.1.4 is that there are very few ways of combining vector fields to get another one. Let's get a better sense of this operation by looking at it in coordinates. It's already nontrivial in one dimension, so to keep things simple we will work on a one-dimensional manifold (either $\mathbb{R}$ or S^1).

Example 14.2.2. Suppose for the moment that we are working on $\mathbb{R}$ (the case of S^1 is similar; we just delete a point). Any two vector fields can be written in coordinates as $X_p = a(x)\,\frac{\partial}{\partial}\big|_p$ and $Y_p = b(x)\,\frac{\partial}{\partial x}\big|_p$, where x is the coordinate of p. (Actually on $\mathbb{R}$, both p and x are the same thing, but we should keep consistency with our more general notation.) Let Z be the Lie bracket $Z = [X,Y]$: then for any function $f\colon \mathbb{R}\to\mathbb{R}$ we have

$$\begin{aligned} Z_p(f) &= X_p\Big(Y(f)\Big) - Y_p\Big(X(f)\Big) = a(x)\frac{d}{dx}\Big(b(x)f'(x)\Big) - b(x)\frac{d}{dx}\Big(a(x)f'(x)\Big) \\ &= a(x)\Big(b'(x)f'(x) + b(x)f''(x)\Big) - b(x)\Big(a'(x)f'(x) + a(x)f''(x)\Big) \\ &= \big[a(x)b'(x) - b(x)a'(x)\big]f'(x). \end{aligned}$$

Since this is true for every f we must have

$$Z_p = \big[a(x)b'(x) - b(x)a'(x)\big]\,\frac{\partial}{\partial x}\Big|_p. \tag{14.2.3}$$

There are obviously other ways to combine two vector fields in one dimension: we can multiply $a(x)$ and $b(x)$ and obtain $R_p = a(x)b(x)\,\frac{\partial}{\partial x}\big|_p$. The reason the product (14.2.3) is useful and this product is not is that the Lie bracket makes sense independent of coordinates, while the product P depends very much on coordinates. To see this, suppose we have a coordinate transformation given by $y = \phi(x)$ with inverse $x = \psi(y)$. Now in the y-coordinate we have

$$X_p = a(x)\,\frac{\partial}{\partial x}\Big|_p = a(x)\,\frac{dy}{dx}\Big|_{x(p)}\,\frac{\partial}{\partial y}\Big|_p = \phi'(x)a(x)\,\frac{\partial}{\partial y}\Big|_p = \phi'\big(\psi(y)\big)a\big(\psi(y)\big)\,\frac{\partial}{\partial y}\Big|_p.$$

Thus $\tilde a(y) = \phi'(\psi(y))a(\psi(y))$ is the coefficient for X in the y-coordinate. Similarly we can express $Y_p = \tilde b(y)\,\frac{\partial}{\partial y}\big|_p$ where $\tilde b(y) = \phi'(\psi(y))b(\psi(y))$. You can write this more simply as $\tilde a(y) = \phi'(x)a(x)$ and $\tilde b(y) = \phi'(x)b(x)$ as long as you keep in mind that this only makes sense if you remember that y is a function of x and vice versa.

If we were trying to express the naïve product R in y-coordinates, we'd just multiply the coefficients $\tilde a(y)$ and $\tilde b(y)$ to get

$$\tilde R_p = \tilde a(y)\tilde b(y)\,\frac{\partial}{\partial y}\Big|_p.$$

So for the product to mean anything, we need $R_p = \tilde R_p$. However

$$\tilde R_p = \phi'(\psi(y))^2 a(\psi(y))b(\psi(y))\,\frac{\partial}{\partial y}\Big|_p$$

while

$$R_p = \phi'(\psi(y))a(\psi(y))b(\psi(y))\,\frac{\partial}{\partial y}\Big|_p$$

and these are not the same unless $\phi' \equiv 1$. Vectors exist on the manifold independently of coordinates, and so everything that you do to vectors *must* be independent of coordinates.

Although it's kind of a pain to go through, you can't really understand this unless you actually see how the Lie bracket given by (14.2.3) really *is* the same regardless of which coordinate system we use. That is, if I wrote $X = \tilde{a}(y)\,\frac{\partial}{\partial y}$ and $Y = \tilde{b}(y)\,\frac{\partial}{\partial y}$ and computed the Lie bracket using (14.2.3) in y-coordinates as

$$\tilde{Z}_p = \big[\tilde{a}(y)\tilde{b}'(y) - \tilde{b}(y)\tilde{a}'(y)\big]\,\frac{\partial}{\partial y}\Big|_p,$$

then I can convert to x coordinates using the transition formulas $\frac{\partial}{\partial y}\big|_p = \psi'(y)\,\frac{\partial}{\partial x}\big|_p$ and $\tilde{a}(y) = \phi'(\psi(y))a(\psi(y))$ and $\tilde{b}(y) = \phi'(\psi(y))b(\psi(y))$. I get
$\tilde{b}'(y) = \phi''(\psi(y))b(\psi(y))\psi'(y)+\phi'(\psi(y))b'(\psi(y))\psi'(y) = \psi'(y)\big[\phi''(x)b(x)+\phi'(x)b'(x)\big]$
and a similar formula for $\tilde{a}'(y)$, so that

$$\begin{aligned}
\tilde{Z}_p &= \psi'(y)\Big[\phi'(x)a(x)\phi''(x)b(x) + \phi'(x)a(x)\phi'(x)b'(x) \\
&\qquad\qquad - \phi'(x)b(x)\phi''(x)a(x) - \phi'(x)b(x)\phi'(x)a'(x)\Big]\,\psi'(y)\,\frac{\partial}{\partial x}\Big|_p \\
&= \psi'(y)^2\phi'(x)^2\big[a(x)b'(x) - b(x)a'(x)\big]\,\frac{\partial}{\partial x}\Big|_p \\
&= \big[a(x)b'(x) - b(x)a'(x)\big]\,\frac{\partial}{\partial x}\Big|_p \\
&= Z_p,
\end{aligned}$$

using the fact that $\phi'(x)\psi'(y) = 1$ by the Inverse Function Theorem. ☺

Classically we would have defined the Lie bracket in coordinates as (14.2.3) and done the computation above to show that it did not actually depend on choice of coordinates. Using the derivation approach means we get coordinate-invariance for free, and so the change-of-coordinates formula is automatic. The philosophy is essentially that since coordinates are just particular functions on the manifold, understanding what an operation does to smooth functions is equivalent to understanding what it does to arbitrary coordinate charts, and studying things on overlaps of charts gets replaced with patching together things defined locally by using bump functions or partitions of unity. The more abstract approach avoids index computations and thus coordinate changes, at the price of having all the objects actually being operators that are defined indirectly, by what they do to other objects.

We have seen that if a linear operator on the space of smooth functions satisfies the Leibniz rule (14.1.1), then it must come from a vector field and must be a first-order differential operator. Let's see what a zero-order differential operator looks like. Such an operator L would satisfy the equation

$$L(fg) = fL(g) \quad \text{for all smooth } f, g\colon M \to \mathbb{R}. \tag{14.2.4}$$

All operators allow us to pull out constants, but few operators allow us to pull out entire functions like (14.2.4). Operators which do are called *tensorial* in differential geometry. We can also characterize such operators.

Proposition 14.2.3. *Suppose L is a linear operator from the space of smooth functions $f\colon M \to \mathbb{R}$, such that $L(fg) = fL(g)$ for all smooth functions f and g. Then there is a smooth function $h\colon M \to \mathbb{R}$ such that $L(g) = hg$ for all smooth functions g.*

Proof. Let $\mathbf{1}$ be the smooth function which takes the value 1 everywhere on M, and define $h = L(1)$. Then h is smooth, and for any smooth g we have

$$L(g) = L(g \cdot \mathbf{1}) = gL(\mathbf{1}) = gh.$$

□

It will be interesting to do the same sort of thing on other spaces: for example the operators from the space of vector fields to the space of smooth functions which are tensorial will end up being exactly the covector fields in the covector bundle, and checking tensoriality gives us a way of defining such covector fields without using coordinates.

We can do one other thing at the moment with vector fields as differential operators. Suppose we have a smooth map $F\colon M \to N$ from one smooth manifold to another. We get an operation from smooth functions on N to smooth functions on M: given any $g\colon N \to \mathbb{R}$, the function $f = g \circ F\colon M \to \mathbb{R}$ is also smooth. We sometimes write $F^*g = g \circ F$ for reasons that will become clear later. Given a vector field X on M and a function g on N, we get a smooth function $X(f) = X(g \circ F)$ on M. We might hope that $X(f)$ is equal to $\tilde{g} \circ F$ for some function $\tilde{g}\colon N \to \mathbb{R}$, for then we could consider $g \mapsto \tilde{g}$ as a linear first-order differential operator on N, and that would give us a vector field on N associated to X. Unfortunately this doesn't work: if F is not surjective, the candidate function $\tilde{g}$ is only determined on the subset $F[M] \subset N$, not on all of N.

Example 14.2.4. Consider $M = \mathbb{R}$ and $N = \mathbb{R}^2$, with F the immersion $F(t) = (\cos t, \sin t)$ so that $F[M] = S^1 \subset \mathbb{R}^2$. Let X on M be the vector field $X = \frac{\partial}{\partial t}$. Given a function $g\colon \mathbb{R}^2 \to \mathbb{R}$, we have

$$X(g \circ F) = \frac{\partial}{\partial t}\Big(g(\cos t, \sin t)\Big) = -\sin t\, g_x(\cos t, \sin t) + \cos t\, g_y(\cos t, \sin t).$$

Of course this is a function on $\mathbb{R}$, but we can ask whether it is equal to $\tilde{g} \circ F$ for some function $\tilde{g}\colon \mathbb{R}^2 \to \mathbb{R}$. And it is: for example $\tilde{g}(x, y) = -y\, g_x(x, y) + x\, g_y(x, y)$ would work. But also $\overline{g}(x, y) = (x^2 + y^2)\tilde{g}(x, y)$ would work: any other function which agrees with $\tilde{g}$ on the circle would give $\tilde{g} \circ F = X(g)$ on M.

A vector field Y on N such that $Y(g) = \tilde{g}$ (or in other words, such that $Y(g) \circ F = X(g \circ F)$) is given by

$$Y_{(x,y)} = -y \left.\frac{\partial}{\partial x}\right|_{(x,y)} + x \left.\frac{\partial}{\partial y}\right|_{(x,y)}.$$

Again another vector field $\tilde{Y}$ on N such that $\tilde{Y}(g) \circ F = g$ is given by $\tilde{Y} = QY$ where Q is a smooth function which is equal to one everywhere on the unit circle. The lesson is that a vector field on M and a map $F\colon M \to N$ gives us part of a vector field on N; it's defined on the subset $F[M]$ of N. ☺

So clearly what we are trying to do doesn't work unless $F[M]$ is surjective. Here's another example to show what fails if F is not injective.

Example 14.2.5. Suppose $M = \mathbb{R}^2$ and $N = \mathbb{R}$ with $F\colon M \to N$ given by $F(x, y) = x$. Given a vector field X on M, we try to define a vector field Y on N which is related in some natural way. We can assume that

$$X_{(x,y)} = a(x, y) \left.\frac{\partial}{\partial x}\right|_{(x,y)} + b(x, y) \left.\frac{\partial}{\partial y}\right|_{(x,y)}$$

for smooth functions a and b on $\mathbb{R}^2$. To figure out what an associated Y should be on $\mathbb{R}$, we start with a function $g\colon N \to \mathbb{R}$ and try to differentiate it.

Given a g on N which we may express as $g(t)$ for $t \in \mathbb{R}$, define $f\colon M \to \mathbb{R}$ by $f = g \circ F$. Then $f(x,y) = g(x)$, so that $X(f)$ is given by

$$X_{(x,y)}(f) = a(x,y)\frac{\partial}{\partial x}g(x) + b(x,y)\frac{\partial}{\partial y}g(x) = a(x,y)g'(x).$$

For this to come from a function on N, it should only depend on the x variable, which of course it does not for a general vector field X (corresponding to a general function a of two variables). ☺

Suppose we tried to go the other way around: consider a smooth map $F\colon M \to N$ and take a vector field Y on N. How could we get a related vector field X on M? Well X would need to differentiate real-valued functions $f\colon M \to \mathbb{R}$, and I know how to differentiate real-valued functions $g\colon N \to \mathbb{R}$, so I want to generate a g given an f. Given f, I can define $g(q)$ by $f(p)$ if $q = F(p)$, but this only makes sense if F is one-to-one: if $F(p_1) = F(p_2) = q$ with $p_1 \neq p_2$, then there is a smooth function on M with $f(p_1) = 1$ and $f(p_2) = 0$, so $g(q)$ can't be defined.

Hence there's nothing natural you can do in general to move vector fields from one manifold M to another manifold N unless your map between them is actually invertible. And if it is, you might as well assume it's a diffeomorphism. However in special cases it may happen that you can push a vector field from one manifold to another.

Definition 14.2.6. Suppose M and N are smooth manifolds, with X a vector field on M and Y a vector field on N. If $F\colon M \to N$ is a smooth map, we say that Y *is F-related to X* if

$$F_*(X_p) = Y_{F(p)} \quad \text{for every } p \in M. \tag{14.2.5}$$

If F is a diffeomorphism, then every Y on N is F-related to a unique vector field X on M and vice versa, and we write $Y = F_\# X$, the *push-forward of X*. In other words,

$$(F_\# X)_p = F_*(X_{F^{-1}(p)}). \tag{14.2.6}$$

Here are some applications of this concept.

- If $\iota\colon M \to N$ is an immersion and we have a vector field Y on N, we can try to find a vector field X on M such that Y is ι-related to X. We will need $\iota_*(X_p) = Y_{\iota(p)}$ for every $p \in M$, and thus it is necessary that $Y_{\iota(p)} \in \iota_*[T_pM]$ for every $p \in M$. If this happens, then there is a unique smooth X such that Y is ι-related to X. (Exercise.)
- If $\tau\colon M \to N$ is a surjective submersion (that is, $\tau_*\colon T_pM \to T_{\tau(p)}N$ is surjective for every $p \in M$), and we are given a vector field X on M, we can try to find a vector field Y on N that is τ-related to X. We need $\tau_*(X_p) = Y_{\tau(p)}$ for every $p \in M$, and thus it is necessary that for $p_1, p_2 \in M$ with $\tau(p_1) = \tau(p_2)$ we have $\tau_*(X_{p_1}) = \tau_*(X_{p_2})$. We can then show that there is a unique smooth Y on N which is τ-related to X. (Exercise.)
- Suppose $F\colon M \to N$ is a quotient map, where N is the quotient by a discrete group action $\{\phi_g \mid g \in G\}$ which is free and proper. Then the condition that $\tau_*(X_{p_1}) = \tau_*(X_{p_2})$ whenever $\tau(p_1) = \tau(p_2)$ is that for every

$g \in G$, we have $(\phi_g)_*(X_p) = X_{\phi_g(p)}$. In other words, X is ϕ_g-related to itself for every $g \in G$. (Exercise.)

- Suppose G is a Lie group and that X is a left-invariant vector field; that is, for some $v \in T_eG$ we have $X_g = (L_g)_* v$ for every $g \in G$. Then for every $h \in G$, the vector field X is (L_h)-related to itself since

$$(L_h)_*(X_g) = (L_h)_*(L_g)_*(v) = (L_{h\cdot g})_*(v) = X_{h\cdot g} = X_{L_h(g)}.$$

  Conversely if X is (L_h)-related to itself for every $h \in G$ then X must be a left-invariant vector field. (Exercise.)

In Example 14.2.4, the vector field $Y = -y\,\frac{\partial}{\partial x} + y\,\frac{\partial}{\partial y}$ on $\mathbb{R}^2$ is F-related to $X = \frac{\partial}{\partial t}$ on $\mathbb{R}$ (along with many other fields Y). In Example 14.2.5, the vector field $Y = g(x)\,\frac{\partial}{\partial x}$ on $\mathbb{R}$ is F-related to $X = a(x)\,\frac{\partial}{\partial x} + b(x,y)\,\frac{\partial}{\partial y}$ on $\mathbb{R}^2$ for any function $b\colon \mathbb{R}^2 \to \mathbb{R}$.

Let's work out the push-forward in an explicit example.

Example 14.2.7. Consider the diffeomorphism $\eta\colon \mathbb{R}^2 \to \mathbb{R}^2$ given in Cartesian coordinates by $\eta(x,y) = (y - x^2, x)$, and the vector field $X = 2xy\,\frac{\partial}{\partial x} + 3y^2\,\frac{\partial}{\partial y}$. Let us compute $\eta_\# X$ at a point (x,y).

To do this, we think of the domain as being given in (u,v) coordinates and the range as being given in (x,y) coordinates. (We do this in spite of the fact that we are really working with the exact same coordinate system. Otherwise it's far too easy to get confused.) So given the point (x,y), we need to work at the inverse image $(u,v) = \eta^{-1}(x,y)$. Since $(x,y) = \eta(u,v) = (v - u^2, u)$, we can easily solve for u and v to see that the inverse is $(u,v) = \eta^{-1}(x,y) = (y, x + y^2)$. Then in (u,v) coordinates, the vector field X is given by (just changing all x's to u's and all y's to v's, because we're just renaming coordinates):

$$X_{(u,v)} = 2uv\,\frac{\partial}{\partial u}\Big|_{(u,v)} + 3v^2\,\frac{\partial}{\partial v}\Big|_{(u,v)}.$$

Now writing $(x,y) = \eta(u,v) = (v - u^2, u)$, we have

$$\begin{aligned}
(\eta_\# X)_{(x,y)} &= \eta_*(X_{\eta^{-1}(x,y)}) = \eta_*(X_{(u,v)}) \\
&= \eta_*\left(2uv\,\frac{\partial}{\partial u}\Big|_{(u,v)} + 3v^2\,\frac{\partial}{\partial v}\Big|_{(u,v)}\right) \\
&= 2uv\,\eta_*\left(\frac{\partial}{\partial u}\Big|_{(u,v)}\right) + 3v^2\,\eta_*\left(\frac{\partial}{\partial v}\Big|_{(u,v)}\right) \\
&= 2uv\,\left(\frac{\partial x}{\partial u}\Big|_{(u,v)}\,\frac{\partial}{\partial x}\Big|_{(x,y)} + \frac{\partial y}{\partial u}\Big|_{(u,v)}\,\frac{\partial}{\partial y}\Big|_{(x,y)}\right) \\
&\qquad + 3v^2\,\left(\frac{\partial x}{\partial v}\Big|_{(u,v)}\,\frac{\partial}{\partial x}\Big|_{(x,y)} + \frac{\partial y}{\partial v}\Big|_{(u,v)}\,\frac{\partial}{\partial y}\Big|_{(x,y)}\right) \\
&= 2uv\,\left(-2u\,\frac{\partial}{\partial x}\Big|_{(x,y)} + \frac{\partial}{\partial y}\Big|_{(x,y)}\right) + 3v^2\,\left(\frac{\partial}{\partial x}\Big|_{(x,y)}\right) \\
&= (3v^2 - 4u^2 v)\,\frac{\partial}{\partial x}\Big|_{(x,y)} + 2uv\,\frac{\partial}{\partial y}\Big|_{(x,y)}.
\end{aligned}$$

Having obtained this formula, all we need to do is substitute $(u, v) = \eta^{-1}(x, y) = (y, x + y^2)$ to get $(\eta_\# X)_{(x,y)}$:

$$(\eta_\# X)_{(x,y)} = (x + y^2)(3x - y^2) \left.\frac{\partial}{\partial x}\right|_{(x,y)} + 2y(x + y^2) \left.\frac{\partial}{\partial y}\right|_{(x,y)}.$$

☺

The following proposition shows us how to understand the notion of "F-related" in terms of the basic operation of vector fields on smooth functions.

Lemma 14.2.8. *Suppose $F\colon M \to N$ is smooth. Let X be a vector field on M and Y a vector field on N such that Y is F-related to X. Then for any smooth function $g\colon N \to \mathbb{R}$, we have $Y(g) \circ F = X(g \circ F)$. Conversely if $X(g \circ F) = Y(g) \circ F$ for every smooth $g\colon N \to \mathbb{R}$, then Y is F-related to X.*

Proof. The functions $X(g \circ F)$ and $Y(g) \circ F$ are both functions from M to $\mathbb{R}$, and we can check the condition by checking at each point $p \in M$: that is, we need to show $X_p(g \circ F) = Y_{F(p)}(g)$ for every $p \in M$ and every smooth $g\colon N \to \mathbb{R}$ if and only if $Y_{F(p)} = F_*(X_p)$. But this just follows from the definition of F_* via equation (11.1.1). □

Since we defined the Lie bracket in terms of the derivation operator on smooth functions, we can best understand the Lie bracket of F-related vector fields in terms of the operation on smooth functions given by Lemma 14.2.8.

Proposition 14.2.9. *Suppose $F\colon M \to N$ is smooth. Let X_1 and X_2 be vector fields on M, and let Y_1 and Y_2 be vector fields on N, and suppose that Y_i is F-related to X_i for $i = 1, 2$. Then $[Y_1, Y_2]$ is F-related to $[X_1, X_2]$.*

Proof. We use Lemma 14.2.8: given a smooth function $g\colon N \to \mathbb{R}$, we have $X_i(g \circ F) = Y_i(g) \circ F$. Therefore we have

$$\begin{aligned}
[X_1, X_2](g \circ F) &= X_1\big(X_2(g \circ F)\big) - X_2\big(X_1(g \circ F)\big) \\
&= X_1\big(Y_2(g) \circ F\big) - X_2\big(Y_1(g) \circ F\big) \\
&= Y_1\big(Y_2(g)\big) \circ F - Y_2\big(Y_1(g)\big) \circ F \\
&= [Y_1, Y_2](g) \circ F.
\end{aligned}$$

□

One application of Proposition 14.2.9 is to compute Lie brackets of vector fields on a submanifold of Euclidean space. For example if $\iota\colon M \to \mathbb{R}^N$ is the embedding, and we have vector fields Y_1 and Y_2 on $\mathbb{R}^N$, then it is easy to compute $[Y_1, Y_2]$ in the global Cartesian coordinate system. If each Y_i is F-related to X_i for some vector field X_i on M (as can always be arranged locally), then $[Y_1, Y_2]$ is F-related to $[X_1, X_2]$, and this tells us what $[X_1, X_2]$ is without having to resort to coordinates on M.

Another application is to left-invariant vector fields on a Lie group. Recall from above that if X on G is left-invariant iff X is (L_h)-related to itself for every $h \in G$. Hence if X and Y are both left-invariant, then $[X, Y]$ is L_h-related to itself for every $h \in G$, and thus $[X, Y]$ is *also* a left-invariant vector field. Now the space of left-invariant vector fields is isomorphic to the tangent space T_eG (since every vector in T_eG generates a left-invariant field by left-translations). Hence the Lie bracket gives us an operation on the finite-dimensional vector space T_eG. Every

finite-dimensional Lie algebra (vector space equipped with a Lie bracket satisfying (14.2.1)–(14.2.2)) is the tangent space of some Lie group, with the bracket operation generated by the Lie bracket of the left-invariant vector fields. (This is rather involved to prove, however.)

We now want to consider another thing that's usually done with vector fields in calculus: using them to define differential equations. In fact the motivation for Lie brackets of vector fields came originally from Lie's work in trying to understand symmetries of differential equations.

14.3. **Vector fields as differential equations.** Suppose we have a vector field X on M, which we will mostly denote by $p \mapsto X_p$ instead of $p \mapsto X(p)$ from now on, to avoid confusion with $X(f)$ for a smooth function f. We can consider solution curves satisfying

$$\frac{d\gamma}{dt} = X_{\gamma(t)}, \qquad \gamma(0) = p, \tag{14.3.1}$$

which makes sense since both sides are in $T_{\gamma(t)}M$ for every t. In a coordinate chart (ϕ, U) with $X_p = \sum_{i=1}^n a^i(p) \frac{\partial}{\partial x^i}\big|_p$, we can write this as

$$\sum_{i=1}^n \frac{d}{dt}(\phi^i \circ \gamma) \frac{\partial}{\partial x^i}\Big|_{\gamma(t)} = \sum_{i=1}^n (a^i \circ \phi^{-1}) \circ (\phi \circ \gamma) \frac{\partial}{\partial x^i}\Big|_{\gamma(t)},$$

so that the solution curves satisfy the first-order system

$$\frac{d(x^i \circ \gamma)}{dt} = (a^i \circ \phi^{-1})\big((x^1 \circ \gamma)(t), \cdots, (x^n \circ \gamma)(t)\big) \text{ for each } i. \tag{14.3.2}$$

Writing $\tilde{a}^i = a^i \circ \phi^{-1}$ so that each $\tilde{a}^i \colon \mathbb{R}^n \to \mathbb{R}$ is C^∞, and forgetting about the curve γ in the manifold, this becomes the first-order system

$$\frac{dx^i}{dt} = \tilde{a}^i(x^1, \ldots, x^n), \qquad x^i(0) = x_0^i,$$

where $x_0^i = \phi^i(p)$.

(Recall that such a system is called "autonomous" or "time-independent" since the right-hand sides do not depend explicitly on t. This feature will be very important.) The advantage of (14.3.1) over (14.3.2) is that, while we can solve (14.3.2) using ODE techniques, the solution might only be valid in some small open set (corresponding to the coordinate chart); on the other hand, we can use the coordinate-invariant expression (14.3.1) to patch together solutions in different charts.

Example 14.3.1. As an example, consider the vector field X on $\mathbb{R}^2$ defined in Cartesian coordinates by $X_p = -y \frac{\partial}{\partial x}\big|_p + x \frac{\partial}{\partial y}\big|_p$. Then in Cartesian coordinates, the solution curves $\big(x(t), y(t)\big)$ satisfy

$$\frac{dx}{dt} = -y, \quad \frac{dy}{dt} = x, \qquad x(0) = x_o, \quad y(0) = y_o.$$

We can solve this by differentiating the first equation:

$$\frac{d^2x}{dt^2} = -\frac{dy}{dt} = -x.$$

This implies $x(t) = A\cos t + B \sin t$. Now we must have $x(0) = A = x_o$ and $x'(0) = B = -y(0) = -y_o$, so that the unique solution is

$$x(t) = x_o \cos t - y_o \sin t, \qquad y(t) = x_o \sin t + y_o \cos t. \tag{14.3.3}$$

On the other hand, we can also write the vector field X in polar coordinates as $X_p = \frac{\partial}{\partial\theta}\big|_p$ since

$$\frac{\partial}{\partial\theta} = \frac{\partial x}{\partial\theta}\frac{\partial}{\partial x} + \frac{\partial y}{\partial\theta}\frac{\partial}{\partial y} = -r\sin\theta\,\frac{\partial}{\partial x} + r\cos\theta\,\frac{\partial}{\partial y} = -y\,\frac{\partial}{\partial x} + x\,\frac{\partial}{\partial y}.$$

Then the differential equation is

$$\frac{dr}{dt} = 0, \quad \frac{d\theta}{dt} = 1, \qquad r(0) = r_o, \quad \theta(0) = \theta_o.$$

This is even easier to solve: we have

$$r(t) = r_o \qquad \text{and} \qquad \theta(t) = \theta_o + t. \tag{14.3.4}$$

Of course this only works when $r_o \neq 0$ (otherwise polar coordinates aren't defined); but we already know the solution when $(x_o, y_o) = (0,0)$ is $x(t) = 0, y(t) = 0$. Hence we have patched together the polar solution with the Cartesian solution. Explicitly, using the transformation formulas $x = r\cos\theta$ and $y = r\sin\theta$, we see that the two solutions are the same: starting with (14.3.4), we get equation (14.3.3).

$$\begin{aligned} x(t) &= r(t)\cos\big(\theta(t)\big) = r_o\cos(\theta_o + t) \\ &= r_o\cos\theta_o\cos t - r_o\sin\theta_o\sin t = x_o\cos t - y_o\sin t, \\ y(t) &= r(t)\sin\big(\theta(t)\big) = r_o\sin(\theta_o + t) \\ &= r_o\sin\theta_o\cos t + r_o\cos\theta_o\sin t = y_o\cos t + x_o\sin t. \end{aligned}$$

☺

We can check that this is true in general: if we change coordinates, we still have the same solution. In fact this must be true based on our original definition $\gamma'(t) = X_{\gamma(t)}$: a solution in one coordinate system must also be a solution in any other. Here we will do a direct coordinate-transformation argument.

Proposition 14.3.2. *If* $\mathbf{x}$ *and* $\mathbf{y}$ *are two overlapping* C^∞*-compatible coordinate systems with* $X = \sum_{k=1}^n a^k(x^1,\cdots,x^n)\,\frac{\partial}{\partial x^k} = \sum_{j=1}^n b^j(y^1,\cdots,y^n)\,\frac{\partial}{\partial y^j}$*, then solutions of*

$$\frac{dx^k}{dt} = a^k\big(\mathbf{x}(t)\big) \qquad \text{and} \qquad \frac{dy^j}{dt} = b^j\big(\mathbf{y}(t)\big)$$

are the same, in the sense that whenever $\mathbf{x}_o = \psi(\mathbf{y}_o)$*, we have*

$$\mathbf{x}(t) = \psi\big(\mathbf{y}(t)\big),$$

where $\psi = \mathbf{x}\circ\mathbf{y}^{-1}$ *is the coordinate transition function.*

Proof. By the existence and uniqueness theorem for ordinary differential equations, Theorem 5.2.6, we know that two curves $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ defined on $(-\varepsilon,\varepsilon)$ with the same initial condition $\mathbf{x}_1(0) = \mathbf{x}_2(0)$ and satisfying the same differential equation on $(-\varepsilon,\varepsilon)$ must actually be identical. So we just have to check that $\mathbf{x}(t)$ and $\psi\circ\mathbf{y}(t)$ satisfy the same differential equation. This follows from the transition formula Proposition 10.4.3, giving $a^k\big(\mathbf{x}\big) = \sum_{j=1}^n \frac{\partial x^k}{\partial y^j}\big|_{\mathbf{y}} b^j\big(\mathbf{y}\big)$. Thus we have

$$\frac{d}{dt}\psi^k\big(\mathbf{y}(t)\big) = \sum_{j=1}^n \frac{\partial x^k}{\partial y^j}\,\frac{dy^j}{dt} = \sum_{j=1}^n \frac{\partial x^k}{\partial y^j} b^j\big(\mathbf{y}(t)\big) = a^k\big(\mathbf{x}(t)\big) = a^k\big(\psi(\mathbf{y}(t))\big).$$

So $\psi\circ\mathbf{y}(t)$ satisfies the same equation as $\mathbf{x}(t)$, and has the same initial condition, and hence must actually be the same. □

The main thing we used in the proof above is that the components of vectors change in the same way as derivatives of curves. So when we change the coordinates of the curve on the left side and change the coordinates of the vector on the right side, they cancel each other out. This is not at all surprising, since we *defined* vectors originally as derivatives of curves. So the only reason we want Proposition 14.3.2 is that we are actually constructing solutions in a particular set of coordinates using Theorem 5.2.6, and we need to know that what we get is actually coordinate-independent.

By Theorem 5.2.6 (the Fundamental Theorem of ODEs), we know that in a neighborhood U of any point $p_o \in M$, we have some time interval $(-\varepsilon, \varepsilon)$ such that for any $p \in U$, there is a unique solution $\gamma_p \colon (-\varepsilon, \varepsilon) \to M$ of (14.3.1) satisfying $\gamma_p(0) = p$. Furthermore, for any fixed t, the function $p \mapsto \gamma_p(t)$ is smooth (in the sense that its coordinate functions are smooth). As a result, we can define the *flow map* Φ_t, although perhaps not on the entire space. The only really new thing here is that, instead of looking at this as the solution in coordinates, we are looking at it as the solution on the manifold (which happens to have coordinate representations).

Definition 14.3.3. If X is a vector field on M, then for every $p_o \in M$, there is an $\varepsilon > 0$ and an open set $U \ni p_o$ such that for every $t \in (-\varepsilon, \varepsilon)$, we have a smooth *flow map* $\Phi \colon (-\varepsilon, \varepsilon) \times U \to M$ defined by $\Phi(t, p) = \gamma_p(t)$, where $\gamma_p(t)$ is the solution of (14.3.1). In other words, we have

$$\frac{\partial \Phi}{\partial t}(t, p) = X_{\Phi(t,p)}, \qquad \Phi(0, p) = p. \tag{14.3.5}$$

It is common to write the flow map as $\Phi_t \colon U \to M$, i.e., $\Phi_t(p) = \Phi(t, p)$ if we want to hold t fixed.

Example 14.3.4. If X is the vector field $X = -y \frac{\partial}{\partial x} + x \frac{\partial}{\partial y}$ defined before, then (from our explicit solution) we know

$$\Phi_t(x, y) = (x \cos t - y \sin t, y \cos t + x \sin t),$$

and Φ_t happens to be defined, for every $t \in \mathbb{R}$, on all of $\mathbb{R}^2$. ☺

This example is not typical: usually the flow is not defined for all time.

Example 14.3.5. For a more typical example, consider the vector field $X(x) = x^2 \frac{\partial}{\partial x}$ on $\mathbb{R}$, for which the differential equation is $x'(t) = x(t)^2$. The solution is $x(t) = x_o/(1 - x_o t)$, so that the flow map is

$$\Phi_t(x) = x/(1 - xt),$$

which is only defined as long as $xt < 1$. (Of course, the formula makes sense for $xt > 1$ as well, but once the solution blows up there's no reason to consider anything beyond that time; hence we only consider the interval containing $t = 0$ on which the solution is smooth.) If $x > 0$, then $\Phi_t(x)$ is defined for all $t < 1/x$, while if $x < 0$, then $\Phi_t(x)$ is defined for all $t > -1/x$. As $x \to \pm\infty$, the time interval on which $\Phi_t(x)$ can be defined symmetrically shrinks to zero, so that in this case, our ε from Definition 14.3.3 is $\varepsilon = 1/|x|$. (When $x = 0$ the solution is $\Phi_t(0) = 0$, which is defined for all $t \in \mathbb{R}$.) ☺

Thus in general, ε depends on the location $p_o \in M$. In particular, we can't even say that there is a universal ε so that $\Phi_t \colon M \to M$ is defined for all $t \in (-\varepsilon, \varepsilon)$. In fact if there were such an ε, then we would have the flow map defined globally, i.e.,

$\Phi_t \colon M \to M$ would exist for all $t \in \mathbb{R}$; see Corollary 14.4.2 later on. This obviously makes things harder, and most of our computations involving flows will have to take place localized near a particular point and also localized in time, just to have the flow maps defined.

We now obtain a very useful alternative view of a vector field as a differential operator. Recall that every vector is the derivative of some curve γ, and that we differentiate functions at a particular point $p \in M$ in the direction of a particular $v \in T_pM$ by writing

$$v(f) = \frac{d}{dt} f\big(\gamma(t)\big)\Big|_{t=0}, \qquad \text{where } \gamma(0) = p \text{ and } \gamma'(0) = v.$$

From a certain point of view, therefore, every "partial derivative operator" is really just an ordinary derivative. Flows allow us to view the function $X(f)$ (which takes the partial derivative of f at each p in direction X_p) as an ordinary derivative as well.

Proposition 14.3.6. *Let X be a vector field and let Φ_t be the local flow. Then for any smooth function $f \colon M \to \mathbb{R}$, we have*

$$X(f)(p) = X_p(f) = \frac{\partial}{\partial t} f\big(\Phi_t(p)\big)\Big|_{t=0}.$$

Proof. Fix $p \in M$. This is literally nothing but the definition (10.3.1) applied to the function f using the curve $t \mapsto \Phi_t(p)$ since the derivative of this curve at time $t = 0$ is precisely the vector X_p. □

As we have seen again and again, there is no calculus but Calculus I, and Differential Geometry is its prophet. Every function we ever need to differentiate is really a function of a real variable if we think about it in the right way.

Now I want to remind you how we actually solve differential equations. The techniques you've learned have basically been ad hoc; pick an equation and hope you're lucky enough to have a technique in the textbook which solves it. One of the things we're eventually aiming for is finding a way to systematize these techniques, which was actually Lie's original motivation for inventing Lie groups. But first I want to make clear that *every* ordinary differential equation (in standard form, i.e., solved for the highest derivatives) actually represents the flow of a vector field.

- First, suppose the coefficients in a system of differential equations depend explicitly on time, as in

$$\begin{aligned} \frac{dx^1}{dt} &= F^1\big(x^1(t), \cdots, x^m(t), t\big), \\ &\vdots \\ \frac{dx^m}{dt} &= F^m\big(x^1(t), \cdots, x^m(t), t\big). \end{aligned}$$

  We want to transform this into an autonomous system, and we use the following trick. Define a new function $x^{m+1}(t) = t$. Then we can write the original nonautonomous differential equation (for m functions) as the

autonomous differential equation (for $(m+1)$ functions) as:

$$\begin{aligned}\frac{dx^1}{dt} &= F^1\big(x^1(t),\cdots,x^m(t),x^{m+1}(t)\big)\\ &\vdots\\ \frac{dx^m}{dt} &= F^m\big(x^1(t),\cdots,x^m(t),x^{m+1}(t)\big)\\ \frac{dx^{m+1}}{dt} &= 1,\end{aligned}$$

corresponding to the vector field

$$X = F^1(x^1,\cdots,x^m,x^{m+1})\,\frac{\partial}{\partial x^1} + \cdots + F^m(x^1,\cdots,x^m,x^{m+1})\,\frac{\partial}{\partial x^m} + \frac{\partial}{\partial x^{m+1}}.$$

- Second, any high-order ordinary differential equation can be written as a system of first-order differential equations by introducing the derivatives as new variables. For example, for the differential equation

$$\frac{d^4x}{dt^4} + 3\frac{dx}{dt}\frac{d^2x}{dt^2} - 2x(t) = 0,$$

we can introduce the functions $x^1(t) = x(t)$, $x^2(t) = \frac{dx}{dt}$, $x^3(t) = \frac{d^2x}{dt^2}$, and $x^4(t) = \frac{d^3x}{dt^3}$; then the differential equation becomes

$$\begin{aligned}\frac{dx^1}{dt} &= x^2(t)\\ \frac{dx^2}{dt} &= x^3(t)\\ \frac{dx^3}{dt} &= x^4(t)\\ \frac{dx^4}{dt} &= -3x^2(t)x^3(t) + 2x^1(t),\end{aligned}$$

corresponding to the vector field

$$X = x^2\,\frac{\partial}{\partial x^1} + x^3\,\frac{\partial}{\partial x^2} + x^4\,\frac{\partial}{\partial x^3} + (-3x^2x^3 + 2x^1)\,\frac{\partial}{\partial x^4}.$$

In general, any k^{th}-order differential equation for one variable becomes a vector field on $\mathbb{R}^k$, and a system of j equations, each of order k, becomes a vector field on $\mathbb{R}^{jk}$.

Using both tricks, we can write the general standard-form system of differential equations (consisting of m equations, each of order at most k, and possibly nonautonomous), as a first-order system of up to $mk+1$ autonomous differential equations; in other words, a vector field on $\mathbb{R}^{mk+1}$.

So here are the standard techniques for solving ODEs.

- The differential equation corresponding to a one-dimensional vector field can always be solved explicitly: we have $\frac{dx}{dt} = f(x)\,\frac{\partial}{\partial x}$, and we can separate the variables to get $\int \frac{dx}{f(x)} = \int dt$. Thus if g is an antiderivative of $1/f$, we obtain $g(x) = t + g(x_o)$, with solution $x(t) = g^{-1}\big(t + g(x_o)\big)$. Hence the flow is $\Phi_t(x_o) = g^{-1}\big(t + g(x_o)\big)$. The time of existence will depend on the

interval on which g is invertible. (We know g is invertible near x_o by the inverse function theorem, since $g'(x_o) = 1/f(x_o) \neq 0$.)

For example, if $\frac{dx}{dt} = 1 + x^2$, then integrating gives $\arctan x = t + \arctan x_o$, so that the flow is $\Phi_t(x_o) = \tan(t + \arctan x_o) = \dfrac{x_o + \tan t}{1 - x_o \tan t}$. Again, this will only be defined for t sufficiently small (and how small depends on x_o).

From a different point of view, we can think of the function g as giving a coordinate change on $\mathbb{R}$ to a new coordinate $y = g(x)$. In the coordinate y, the differential equation becomes

$$\frac{dy}{dt} = g'\big(x(t)\big)\,\frac{dx}{dt} = \frac{1}{f\big(x(t)\big)}\,f\big(x(t)\big) = 1,$$

with solution $y(t) = y_o + t$. This is valid only where the coordinate change is valid (which will not be on all of $\mathbb{R}$ in general). The notion of solving a differential equation by changing coordinates is extremely important, as we will see.

In terms of vector fields, we have rewritten the field $X = f(x)\,\frac{\partial}{\partial x}$ in the form $X = \frac{\partial}{\partial y}$ by solving the equation $dy/dx = 1/f(x)$.

- Linear systems, given by $\mathbf{x}'(t) = A\mathbf{x}(t)$. The solution with $\mathbf{x}(0) = \mathbf{x}_o$ is $\mathbf{x}(t) = e^{tA}\mathbf{x}_o$, where

$$e^{tA} = \sum_{k=0}^{\infty} \frac{t^n}{n!} A^n.$$

As an example, consider the system $\frac{dx}{dt} = x + y$, $\frac{dy}{dt} = y$. This is a linear system with $A = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$. We can compute the powers of A inductively: $A^k = \left(\begin{smallmatrix} 1 & k \\ 0 & 1 \end{smallmatrix}\right)$, so that

$$e^{tA} = \begin{pmatrix} \sum_{k=0}^{\infty} \frac{t^k}{k!} & \sum_{k=0}^{\infty} \frac{kt^k}{k!} \\ 0 & \sum_{k=0}^{\infty} \frac{t^k}{k!} \end{pmatrix} = \begin{pmatrix} e^t & \sum_{k=1}^{\infty} t\,\frac{t^{k-1}}{(k-1)!} \\ 0 & e^t \end{pmatrix} = \begin{pmatrix} e^t & te^t \\ 0 & e^t \end{pmatrix}.$$

As a result, the solution of the equation with $x(0) = x_o$ and $y(0) = y_o$ is $x(t) = x_o e^t + y_o t e^t$ and $y(t) = y_o e^t$; in flow form, we get $\Phi_t(x_o, y_o) = (x_o e^t + y_o t e^t, y_o e^t)$.

The method of undetermined coefficients (guessing the form of a solution and plugging in to find the actual coefficients) also works to solve this particular differential equation. However understanding the matrix exponential makes remembering the rules for undetermined coefficients unnecessary. In addition it allows us to write the solution quite simply as $\mathbf{x}(t) = e^{tA}\mathbf{x}_o$, even if the explicit formula for e^{tA} is complicated.

In general one computes the matrix exponential as follows: write A in Jordan form as $A = PJP^{-1}$; then $\exp(tA) = P\exp(tJ)P^{-1}$, and the exponential of a matrix in Jordan form is built from the exponentials of Jordan blocks, which are fairly easy to compute directly.

We can also think of this example in terms of new coordinates. We have computed that $\Phi(t, x, y) = (xe^t + yte^t, ye^t)$ where t is time and the initial condition is (x, y). Now consider the new coordinates (u, v) given by $(x, y) = \Phi(u, 0, v)$. Explicitly we have $x = vue^u$ and $y = ve^u$, and solving for (u, v) gives $u = x/y$ and $v = ye^{-x/y}$. Hence the vector field

$X = (x+y)\frac{\partial}{\partial x} + y\frac{\partial}{\partial y}$ which generated our differential equation takes the form

$$\begin{aligned} X &= (x+y)\frac{\partial}{\partial x} + y\frac{\partial}{\partial y} \\ &= (vue^u + ve^u)\left(\frac{\partial u}{\partial x}\frac{\partial}{\partial u} + \frac{\partial v}{\partial x}\frac{\partial}{\partial v}\right) + ve^u\left(\frac{\partial u}{\partial y}\frac{\partial}{\partial u} + \frac{\partial v}{\partial y}\frac{\partial}{\partial v}\right) \\ &= (vue^u + ve^u)\left(\frac{1}{y}\frac{\partial}{\partial u} - e^{-x/y}\frac{\partial}{\partial v}\right) \\ &\qquad + ve^u\left(-\frac{x}{y^2}\frac{\partial}{\partial u} + (e^{-x/y} + \tfrac{x}{y}e^{-x/y})\frac{\partial}{\partial v}\right) \\ &= \frac{\partial}{\partial u}. \end{aligned}$$

In these coordinates the differential equation is $u'(t) = 1$ and $v'(t) = 0$, which we view as the simplest possible system. This coordinate change is not unique: more generally we could take a generic curve $v \mapsto \gamma(v) \in \mathbb{R}^2$ and build coordinates using $(u, v) \mapsto \Phi(u, \gamma(v))$, and again we will have $X = \frac{\partial}{\partial u}$.

- For higher-dimensional nonlinear equations, things can be more complicated. However, in the right coordinates, they simplify. For example, consider the differential equation $\frac{dx}{dt} = x(x^2+y^2)$, $\frac{dy}{dt} = y(x^2+y^2)$. You can probably guess that this equation would look simpler in polar coordinates, based on seeing the term (x^2+y^2). Indeed, writing the vector field as $X = x(x^2+y^2)\frac{\partial}{\partial x} + y(x^2+y^2)\frac{\partial}{\partial y}$ and using the polar coordinate transformation formula, we get

$$\begin{aligned} X &= x(x^2+y^2)\left(\frac{\partial r}{\partial x}\frac{\partial}{\partial r} + \frac{\partial \theta}{\partial x}\frac{\partial}{\partial \theta}\right) + y(x^2+y^2)\left(\frac{\partial r}{\partial y}\frac{\partial}{\partial r} + \frac{\partial \theta}{\partial y}\frac{\partial}{\partial \theta}\right) \\ &= r^3\cos\theta\left(\cos\theta\frac{\partial}{\partial r} - \frac{\sin\theta}{r}\frac{\partial}{\partial \theta}\right) + r^3\sin\theta\left(\sin\theta\frac{\partial}{\partial r} + \frac{\cos\theta}{r}\frac{\partial}{\partial \theta}\right) \\ &= r^3\frac{\partial}{\partial r}. \end{aligned}$$

As a result, the differential equations in polar coordinates are $\frac{dr}{dt} = r^3$ and $\frac{d\theta}{dt} = 0$, with solution $r(t) = \dfrac{r_o}{\sqrt{1-2r_o^2 t}}$, $\theta(t) = \theta_o$. The flow in rectangular coordinates is then $\Phi_t(x_o, y_o) = \left(\dfrac{x_o}{\sqrt{1-2(x_o^2+y_o^2)t}}, \dfrac{y_o}{\sqrt{1-2(x_o^2+y_o^2)t}}\right)$.

The fact that the equation simplifies in polar coordinates comes from the fact that the solutions commute with rotations: if Ψ_s represents the family of rotations by angle s, i.e., $\Psi_s(x_o, y_o) = (x_o\cos s + y_o\sin s, -x_o\sin s + y_o\cos s)$, then we will have

$$\Phi_t \circ \Psi_s(x_o, y_o) = \Psi_s \circ \Phi_t(x_o, y_o)$$

for every s and t. It's worth checking this explicitly. It tells us that if we first rotate the initial condition, then solve the differential equation, it's the same as solving the equation first with unrotated initial condition, then

rotating. Later in this section we'll be able to figure out how we might have detected this sort of thing without already knowing what the flow Φ_t is.

Finally, we notice that in this example, we reduced a two-dimensional system to a one-dimensional system; we could use the technique for one-dimensional systems (another coordinate change) to obtain new coordinates $(\rho, \psi) = (-1/(2r^2), \theta)$ in which the differential equations are $\rho'(t) = 1$ and $\psi'(t) = 0$.

14.4. **Flows and one-parameter groups.** The most important thing about the (locally-defined) flow operators Φ_t from Definition 14.3.3 is that they form a group, in the sense that Φ_0 is the identity and $(\Phi_{t_1} \circ \Phi_{t_2})(p) = \Phi_{t_1+t_2}(p)$ (as long as t_1 and t_2 are small enough that this is actually defined). The inverse of each Φ_t is obviously Φ_{-t}, which is defined as long as t is close enough to 0. Hence the map $t \mapsto \Phi_t$ is a group homomorphism. We can check that this is true for some of the flows defined above:

- If $\Phi_t(x_o) = \tan(t + \arctan x_o)$, then

$$\begin{aligned}\Phi_{t_1} \circ \Phi_{t_2}(x_o) &= \tan\left(t_1 + \arctan\left[\Phi_{t_2}(x_o)\right]\right) \\ &= \tan\left(t_1 + \arctan\left[\tan(t_2 + \arctan x_o)\right]\right) \\ &= \tan(t_1 + t_2 + \arctan x_o) \\ &= \Phi_{t_1+t_2}(x_o),\end{aligned}$$

  although this formula is only valid if $-\frac{\pi}{2} < t_2 + \arctan x_o < \frac{\pi}{2}$ and also $-\frac{\pi}{2} < t_1 + t_2 + \arctan x_o < \frac{\pi}{2}$.
- If $\Phi_t = e^{tA}$ for a matrix A, then the composition of functions is simply matrix multiplication: $\Phi_{t_1} \circ \Phi_{t_2} = e^{t_1 A} e^{t_2 A} = e^{(t_1+t_2)A} = \Phi_{t_1+t_2}$.
- If $\Phi_t(x_o, y_o) = \dfrac{(x_o, y_o)}{\sqrt{1 - 2(x_o^2 + y_o^2)t}}$, then

$$\begin{aligned}(\Phi_{t_1} \circ \Phi_{t_2})(x_o, y_o) &= \frac{(x_o, y_o)}{\sqrt{1 - 2(x_o^2 + y_o^2)t_2}\sqrt{1 - 2\left(\frac{x_o^2+y_o^2}{1-2(x_o^2+y_o^2)t_2}\right)t_1}} \\ &= \frac{(x_o, y_o)}{\sqrt{1 - 2(x_o^2 + y_o^2)(t_1 + t_2)}} \\ &= \Phi_{t_1+t_2}(x_o, y_o).\end{aligned}$$

  Again this only makes sense if t_1 and t_2 are small enough.

Now we prove this in general.

Proposition 14.4.1. *Let X is a vector field on M and Φ_t the local flows as in Definition 14.3.3, then for every $p \in M$ there is a real number $\varepsilon > 0$ such that, whenever $t_1, t_2, t_1 + t_2 \in (-\varepsilon, \varepsilon)$, we have*

$$\Phi_{t_1+t_2}(p) = (\Phi_{t_1} \circ \Phi_{t_2})(p).$$

In particular if Φ_t is defined on all of M for all $t \in \mathbb{R}$, we have $\Phi_{t_1+t_2} = \Phi_{t_1} \circ \Phi_{t_2}$.

Proof. By definition of Φ_t, we know that for every $p \in M$, the curve $t \mapsto \Phi_t(p)$ is the unique solution of the equation $\gamma'(t) = X_{\gamma(t)}$ with initial condition $\gamma(0) = p$.

Fix $\tau \in \mathbb{R}$ small enough that $q = \Phi_\tau(p)$ exists, and consider the two curves $\gamma_1(t) = \Phi_{t+\tau}(p)$ and $\gamma_2(t) = \Phi_t(q)$. Both of these curves satisfy the differential equation

$$\gamma_i'(t) = X_{\gamma(t)}, \qquad \gamma(0) = q,$$

and therefore they must be the same for all t for which both are defined; in other words we have $\Phi_{t+\tau}(p) = \Phi_t(\Phi_\tau(p))$ for all values of t where both sides are defined. Since τ was an arbitrary time close to zero, this equation holds for any two sufficiently small times. □

We would like to view this algebraically as a homomorphism into the group of diffeomorphisms of M, but the problem is that we have only defined the maps Φ_t locally on subsets of M, and we don't know that there is any time for which Φ_t is actually globally defined on M. There is a good reason for this.

Corollary 14.4.2. *Suppose that there is an $\varepsilon > 0$ such that the local flows Φ from Definition 14.3.3 are defined on $\Phi\colon (-\varepsilon, \varepsilon) \times M \to M$. Then in fact the flow map can be extended to $\Phi\colon \mathbb{R} \times M \to M$, and the maps $t \mapsto \Phi_t$ are global homomorphisms from the additive group $\mathbb{R}$ to the diffeomorphism group of M (under composition).*

Proof. Suppose Φ_t is defined for $t \in (-\varepsilon, \varepsilon)$ on all of M. We can extend Φ_t to be defined for $t \in (-2\varepsilon, 2\varepsilon)$ by setting $\Phi_t(p) = \Phi_{t/2}(\Phi_{t/2}(p))$; the group property implies that $\Phi_t(p)$ is still the unique solution of (14.3.5) on this larger interval. Repeating this process, we can clearly extend the interval indefinitely to all of $\mathbb{R}$. □

We can also prove the converse of this theorem: any one-parameter group of diffeomorphisms must arise as the flow of a vector field. The statement is a bit complicated since we have to worry about the flows being only local: the simpler version of the hypothesis is that $\Phi\colon \mathbb{R} \times M \to \mathbb{R}$ is smooth and satisfies $\Phi_{t_1+t_2} = \Phi_{t_1} \circ \Phi_{t_2}$ for all $t_1, t_2 \in \mathbb{R}$.

Proposition 14.4.3. *Suppose that Φ_t is a family of maps defined on some open domains $U_t \subset M$ for $t > 0$, with $(t, p) \mapsto \Phi_t(p)$ smooth on its domain, and such that $U_0 = M$ and Φ_0 is the identity on M. Suppose that the union of all U_t is M, and that $\Phi_{t_1+t_2} = \Phi_{t_1} \circ \Phi_{t_2}$ at all points where both sides are defined. Then there is a C^∞ vector field X defined on M such that Φ_t is the flow of X.*

Proof. In order for a vector field X to generate the flow Φ_t, we must have by Definition 14.3.3 for all $t \in (-\varepsilon, \varepsilon)$ and for all $p \in M$. In particular when $t = 0$, we must have

$$X_p = \frac{\partial \Phi(t, p)}{\partial t}\Big|_{t=0} \qquad \text{for all } p \in M.$$

Let us define a vector field X by this formula for all p; then we just have to check that (14.3.5) is valid.

So as in Proposition 14.4.1, we take any fixed τ, and observe that by the group equation we have for any $p \in M$ that

$$\frac{\partial}{\partial t}\Big|_{t=\tau} \Phi_t(p) = \frac{\partial}{\partial t}\Big|_{t=0} \Phi_{t+\tau}(p) = \frac{\partial}{\partial t}\Big|_{t=0} \Phi_t\big(\Phi_\tau(p)\big) = X_{\Phi_\tau(p)}$$

by definition of X. □

It is quite common to study non-autonomous differential equations in terms of time-dependent vector fields: then we have

$$\frac{\partial \Phi_t(p)}{\partial t} = X\big(t, \Phi_t(p)\big), \qquad \Phi_0(p) = p,$$

and as before this will have a solution for any fixed p defined in some time interval. However we will no longer have the group structure: in general, $\Phi_{t_1+t_2} \neq \Phi_{t_1} \circ \Phi_{t_2}$. For our purposes, it will be preferable to just work with time-independent vector fields by using the extra-variable trick.

So far everything we have discussed is perfectly valid for any manifold. Unfortunately, it can frequently happen that the flow Φ_t of a vector field X is only defined for a small time interval $(-\varepsilon, \varepsilon)$ in a small neighborhood of each point, with the ε depending on the point, as for example when $M = \mathbb{R}$ and $X = x^2 \frac{\partial}{\partial x}$. However if M is a *compact* manifold, then the flow must be defined for all time. This is a very convenient property. So far we've always been talking about manifolds as more complicated than Euclidean space, but from the point of view of differential equations, compact manifolds are much simpler than Euclidean space. Thus for example people who work in differential equations even on Euclidean space frequently prefer working with periodic equations (which make sense on the torus $\mathbb{T}^n$), so that various global things can be guaranteed.

Theorem 14.4.4. *If M is a compact n-dimensional manifold, and X is a smooth vector field on M, then there is a global flow, i.e., maps $\Phi_t \colon M \to M$ for all $t \in \mathbb{R}$ such that $\frac{\partial \Phi_t(p)}{\partial t} = X(\Phi_t(p))$ for all t.*

Proof. By the local existence theorem, for every point p there is an open set U and a positive number ε such that whenever $q \in U$, the solution $\Phi_t(q)$ is defined on $(-\varepsilon, \varepsilon)$. Cover M by finitely many of these open sets $U_1, \cdots, U_N$, and set ε to be the minimum of $\varepsilon_1, \cdots, \varepsilon_N$. Then we know that for any point q on M whatsoever, the flow map $\Phi_t(q)$ is defined for the same $(-\varepsilon, \varepsilon)$. So $\Phi_t \colon M \to M$ is defined for $t \in (-\varepsilon, \varepsilon)$. By Corollary 14.4.2, the map Φ_t is thus defined for every $t \in \mathbb{R}$. □

Compactness is useful to ensure global existence of flows, but sometimes other tools are useful. For example, if the vector field is uniformly bounded or Lipschitz in some norm we can prove global existence directly.

14.5. **Straightening vector fields.** At the end of Section 14.3, we discussed explicit solutions of some typical differential equations which all have in common the fact that the equations become simpler in better coordinates.

- For the one-variable autonomous differential equation $\frac{dx}{dt} = f(x)$, we are dealing with the vector field $X = f(x) \frac{\partial}{\partial x}$. The coordinate change $y = \int dx/f(x)$ yields $X = \frac{\partial}{\partial y}$.
- For the n-dimensional linear system $\frac{dx}{dt} = x + y$, $\frac{dy}{dt} = y$, the coordinate change $u = x/y$ and $v = ye^{-x/y}$ changes the vector field's components from $X = (x+y)\frac{\partial}{\partial x} + y\frac{\partial}{\partial y}$ to $X = \frac{\partial}{\partial u}$.
- For the two-variable differential equation $\frac{dx}{dt} = x(x^2+y^2)$, $\frac{dy}{dt} = y(x^2+y^2)$, the coordinate change $\rho = -1/[2(x^2+y^2)]$, $\psi = \arctan(y/x)$ changes the vector field's components from $X = (x^2+y^2)\big(x\frac{\partial}{\partial x} + y\frac{\partial}{\partial y}\big)$ to $X = \frac{\partial}{\partial \rho}$.

In all of these cases, we have "straightened" the vector field out, so that in the new coordinate system, we are just dealing with a particular coordinate vector.

If we can do this, we can always solve the differential equation. In some sense, therefore, solving differential equations explicitly is just a matter of finding the right coordinates. Abstractly, we can always do this (since a differential equation always has a solution). Practically this may or may not be effective. The following makes this notion precise. It will also give us a very convenient tool to use when working with vector fields abstractly. Although the proof seems fairly complicated, you should observe that it's just a generalization of the exact same technique we used to straighten the vector field $X = (x+y)\frac{\partial}{\partial x} + y\frac{\partial}{\partial y}$ at the end of Section 14.3.

Theorem 14.5.1. *Let X be any vector field on an n-dimensional manifold M. For every $p \in M$ such that $X(p) \neq 0$, there is a coordinate chart $(\mathbf{y} = \psi, V)$ with $V \ni p$, such that $X = \frac{\partial}{\partial y^1}$ on V.*

Thus the flow in ψ-coordinates is given by

$$\psi \circ \Phi_t \circ \psi^{-1}(y^1_o, y^2_o, \cdots, y^n_o) = (y^1_o + t, y^2_o, \cdots, y^2_o),$$

as long as t is sufficiently small that $\Phi_t(\psi^{-1}(y^1_o, \cdots, y^n_o)) \in V$.

Proof. The technique to prove this is to essentially start with the flow (which we know to exist by Theorem 5.2.6) and construct the coordinates from that, using the Inverse Function Theorem 5.2.4.

Given a vector field X and a local flow Φ_t defined for $(-\varepsilon, \varepsilon)$ in some neighborhood of p, let $(\phi = \mathbf{x}, U)$ be a coordinate chart around p. (We can assume $\phi(p) = \mathbf{0}$ and that $\phi[U] = \mathbb{R}^n$ with Φ_t defined on all of U, by shrinking around p if necessary.) Let $\Xi_t = \phi \circ \Phi_t \circ \phi^{-1}$ be the flow as a map in coordinates, with $\Xi_t \colon \mathbb{R}^n \to \mathbb{R}^n$. If in these coordinates we have

$$X \circ \phi^{-1}(x^1, \ldots, x^n) = \sum_{k=1}^n a^k(x^1, \ldots, x^n) \left.\frac{\partial}{\partial x^k}\right|_q, \qquad \text{where } \phi(q) = (x^1, \ldots, x^n),$$

then the coordinate flow Ξ_t satisfies

(14.5.1)
$$\frac{\partial}{\partial t}\Xi^k_t(x^1, \ldots, x^n) = a^k\big(\Xi^1_t(x^1, \ldots, x^n), \ldots, \Xi^n_t(x^1, \ldots, x^n)\big) \quad \text{for } 1 \leq k \leq n.$$

By assumption $X_p \neq 0$, and so in coordinate components we have $a^k(\mathbf{0}) \neq 0$ for at least one k (since $\phi(p) = \mathbf{0}$ and a^k are the components of X in a basis). We can assume without loss of generality that $a^1(\mathbf{0}) \neq 0$. We will use this fact in a bit when we apply the Inverse Function Theorem.

Recall from Definition 14.3.3 that we can write $\Xi_t(p) = \Xi(t, p)$, where Ξ is smooth jointly; i.e., the flow map $\Xi \colon (-\varepsilon, \varepsilon) \times \mathbb{R}^n \to \mathbb{R}^n$ is smooth. Define $F \colon (-\varepsilon, \varepsilon) \times \mathbb{R}^{n-1} \to \mathbb{R}^n$ by

$$(x^1, \ldots, x^n) = F(y^1, y^2, \ldots, y^n) = \Xi(y^1, 0, y^2, \ldots, y^n). \tag{14.5.2}$$

I want to show that F is invertible near the origin, for then I will know that $\mathbf{y} = \psi = F^{-1} \circ \phi$ is a new coordinate chart on M. Before I do this, however, let's check that this chart actually solves the problem. In other words, the flow in $\mathbf{y}$-coordinates should be very simple. Now since $\mathbf{x} = F(\mathbf{y})$, and since Ξ_t was the flow in $\mathbf{x}$-coordinates, the flow Σ_t in $\mathbf{y}$-coordinates will be

(14.5.3)
$$\begin{aligned}\Sigma_t(y^1, \ldots, y^n) &= F^{-1} \circ \Xi_t \circ F(y^1, \ldots, y^n) = F^{-1} \circ \Xi_t \circ \Xi_{y^1}(0, y^2, \ldots, y^n) \\ &= F^{-1} \circ \Xi_{t+y^1}(0, y^2, \ldots, y^n) = (t + y^1, y^2, \ldots, y^n).\end{aligned}$$

This is a very simple formula for the flow Σ_t in $\mathbf{y}$-coordinates. Furthermore as in the proof of Proposition 14.4.3, the vector field can be recovered from the flow simply by taking the time-derivative of the flow at time $t = 0$. Hence if the flow looks like (14.5.3), then the vector field must look like

$$X_q = X \circ \psi^{-1}(y^1, \dots, y^n) = \frac{\partial}{\partial y^1}\Big|_q, \qquad \text{where } \psi(q) = F^{-1}(\phi(q)) = (y^1, \dots, y^n),$$

for any q in the coordinate neighborhood. This is the straightening we want.

So we just have to prove that F is invertible near the origin, and by the Inverse Function Theorem 5.2.4, it is of course sufficient to show that $DF(\mathbf{0})$ is an invertible matrix. We have by the definition (14.5.2) that the components of $DF(\mathbf{0})$ are $\dfrac{\partial x^j}{\partial y^i}\Big|_{\mathbf{0}}$, where

$$\frac{\partial x^j}{\partial y^1}(0, \dots, 0) = \frac{\partial \Xi^j}{\partial y^1}(y^1, 0, y^2, \dots, y^n)\Big|_{\mathbf{y}=\mathbf{0}} = a^j(0, \dots, 0)$$

using (14.5.1) (since the y^1-direction is the time direction). All other partial derivatives can be computed by first setting $y^1 = 0$ to simplify: since Ξ_0 is the identity, we have $\Xi^j(0, 0, y^2, \dots, y^n) = y^j$ if $j > 1$ and $\Xi^1(0, 0, y^2, \dots, y^n) = 0$. Thus we get

$$\frac{\partial x^j}{\partial y^i}(\mathbf{0}) = \frac{\partial \Xi^j}{\partial y^i}(\mathbf{0}) = \frac{\partial y^j}{\partial y^i} = \delta_i^j \qquad \text{if } i > 1 \text{ and } j > 1$$

while

$$\frac{\partial x^1}{\partial y^i} = 0 \qquad \text{if } i > 1.$$

We thus see that $DF(\mathbf{0})$ is the matrix

$$DF(\mathbf{0}) = \begin{pmatrix} a^1(\mathbf{0}) & 0 & \cdots & 0 \\ a^2(\mathbf{0}) & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a^n(\mathbf{0}) & 0 & \cdots & 1 \end{pmatrix},$$

where the lower right corner is the $(n-1) \times (n-1)$ identity matrix. Obviously this is invertible if $a^1(\mathbf{0}) \neq 0$, which is why we assumed this earlier. (If it had not been, we could have used another hyperplane instead of $(0, y^2, \dots, y^n) \in \mathbb{R}^n$.) Thus F is invertible in some neighborhood of the origin, and $\psi = F^{-1} \circ \phi$ defines a genuine coordinate chart on M in a neighborhood of p such that $X = \frac{\partial}{\partial y^1}$ everywhere on that neighborhood. □

Theorem 14.5.1 tells us that every vector field can be straightened in a neighborhood of a point where it is nonzero, and hence in the right coordinate system every differential equation can be solved explicitly. Of course the construction of the right coordinates relies on having a formula for the flow map already, and so this is certainly a circular argument: it definitely does *not* give any explicit formula for the solution of a general differential equation. It's primarily useful if you are given a particular vector field and want to do something else with it: you can always assume it is of this simple form rather than something complicated. (We will use this at the end of this Chapter.)

Practically we almost never solve differential equations by guessing a coordinate chart in which $X = \frac{\partial}{\partial y^1}$, although in principle if we were very lucky we always could. Practically we find symmetries of the vector field and then use coordinate

transformations to peel off one direction at a time, until in the end the vector field does become trivial. We did this at the end of Section 14.3 for the vector field $X = (x^2 + y^2)\,(x\,\frac{\partial}{\partial x} + y\,\frac{\partial}{\partial y})$ on $\mathbb{R}^2$. Changing into polar coordinates, we saw that $X = r^3\,\frac{\partial}{\partial r}$, for which the flow can be found explicitly since the differential equation doesn't depend on the θ coordinate. The fact that this worked depended on knowing the definition of polar coordinates already. We could compute the flow of the vector field explicitly, and we knew the rotations of the plane explicitly, and so we could check that the flow commuted with rotations. In general this doesn't work since we don't know the flow and we may not even know the symmetries. We want to understand, as Lie did, how to determine whether the flow of a vector field commutes with some symmetries without having to actually find the flow.

Example 14.5.2. Now what you may have already noticed is that the rotational symmetries in the plane (in our third differential equations example) are also a group: if

$$\Psi_s(x_o, y_o) = \big(x_o \cos s - y_o \sin s, x_o \sin s + y_o \cos s\big),$$

then we have $\Psi_{s_1+s_2} = \Psi_{s_1} \circ \Psi_{s_2}$ for every $s_1, s_2 \in \mathbb{R}$ by the sum rule for sine and cosine. Thus by Proposition 14.4.3, Ψ_s must actually come from a vector field. It's not hard to see what this field must be; in fact, since in matrix form we have

$$\Psi_s = \begin{pmatrix} \cos s & -\sin s \\ \sin s & \cos s \end{pmatrix} = \exp\left[s \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \right],$$

we expect to have the vector field generated by the matrix $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, which means

$$U = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}, \quad \text{or in differential geometric notation,} \quad U = -y\,\frac{\partial}{\partial x} + x\,\frac{\partial}{\partial y}.$$

Thus by our first example, the differential equation is $\frac{dx}{dt} = -y$, $\frac{dy}{dt} = x$, corresponding to the vector field $U = -y\,\frac{\partial}{\partial x} + x\,\frac{\partial}{\partial y}$. Since this vector field is the basic rotation field, it should not surprise you at all that in polar coordinates we have $U = \frac{\partial}{\partial \theta}$. ☺

What we are aiming for now is Lie's basic discovery: the fact that the flow Φ_t of a vector field X and a group of symmetries Ψ_s commutes (i.e., $\Phi_t \circ \Psi_s = \Psi_s \circ \Phi_t$ for all t and s) can be expressed directly in terms of a condition on the vector fields X and the generating field U of Ψ_s. First, observe that for any fixed s and any flows Φ_t and Φ_s forming a group, the conjugations $\Xi_t = \Psi_s \circ \Phi_t \circ \Psi_{-s}$ also form a group: letting $F = \Psi_s$ for a fixed s, we have

$$\begin{aligned} \Xi_{t_1} \circ \Xi_{t_2} &= F \circ \Phi_{t_1} \circ F^{-1} \circ F \circ \Phi_{t_2} \circ F^{-1} \\ &= F \circ \Phi_{t_1} \circ \Phi_{t_2} \circ F^{-1} = F \circ \Phi_{t_1+t_2} \circ F^{-1} = \Xi_{t_1+t_2}, \end{aligned}$$

for every t_1 and t_2. As a result, there must be some vector field that generates Ξ_t. The next Proposition shows what it is.

Proposition 14.5.3. *Suppose $F\colon M \to M$ is a diffeomorphism. Let X be a smooth vector field on M, and let $Y = F_{\#}X$ be the push-forward defined by formula (14.2.6). If the local flow of X is Φ_t, then the local flow of Y is $\Xi_t = F \circ \Phi_t \circ F^{-1}$.*

Proof. Let $g\colon M \to \mathbb{R}$ be a smooth function. Then we understand Y if we understand $Y(g)$ by Lemma 14.2.8, and we understand $Y(g)$ if we understand the flow of Y by Proposition 14.3.6. So let us begin with the flow $\Xi_t = F \circ \Phi_t \circ F^{-1}$, which

satisfies the homomorphism property as we just computed, and thus is the flow of some vector field Y by Proposition 14.4.3. We want to show that $Y = F_{\#}X$. By Proposition 14.3.6 we have

$$Y_q(g) = \frac{\partial}{\partial t}(g \circ \Xi_t(q))\Big|_{t=0} = \frac{\partial}{\partial t}\big(g \circ F \circ \Phi_t \circ F^{-1}(q)\big) = X_{F^{-1}(q)}(g \circ F) = (F_{\#}X)_q(g),$$

where we use Lemma 14.2.8 in the last step. Since this is true for every g and q, we conclude that $Y = F_{\#}X$. □

There are a few things worth noticing about this Proposition.

- Despite the symmetry of the flow compositions $\Xi_t = F \circ \Phi_t \circ F^{-1}$, the compositions of the corresponding vector field are not symmetric: $F_{\#}X = F_*(X \circ F^{-1})$. We take the derivative of F on the outside but not on the inside. The reason is because the vector field comes from differentiating the flow with respect to t; to do this, we do not have to differentiate the inner composition at all (since it doesn't depend on t), but we do have to differentiate the outer composition using the Chain Rule (since it depends on t implicitly).
- Also, by our definition of vector field, we are *forced* to use the formula (14.2.6): if we want to transform a vector using a map F, we need to use the derivative F_* of the map; however, the derivative doesn't go into the correct tangent space to get a vector field. Thus if we want a vector at p, we need to go backwards and start with the vector at $F^{-1}(p)$; then we push forward with the only possible F-derivative F_* (the one that starts at $F^{-1}(p)$ and must therefore end up at p). So the seemingly complicated formula (14.2.6) is the only one that even makes sense, given our definitions.
- Although the definition of $F_{\#}X$ looks most natural when we write it in terms of the corresponding flows, recall that the flow of a vector field is generally not defined globally, even for short times. Thus it is better to use the formula (14.2.6) as a definition since it is computed directly in terms of the field, not the flow.

Using this tool, we now can say that if Ψ_s is a group of symmetries (for convenience, we can assume they are defined for all $s \in \mathbb{R}$), then for each s we have a vector field $(\Psi_s)_{\#}X$ which measures how much Ψ_s changes X. If the flow of X commutes with the symmetries, then we have $\Phi_t = \Psi_s \circ \Phi_t \circ \Psi_{-s}$ for every t and s, and hence $(\Psi_s)_{\#}X = X$ for all s. In general, then, we can measure the degree to which Ψ_s commutes with Φ by taking the derivative of $(\Psi_s)_{\#}X$ with respect to s. This makes sense because, for each fixed p, the vector $\big((\Psi_s)_{\#}X\big)_p$ lies in T_pM for every s. So we can take derivatives without worrying about curves: we're just differentiating a curve that lives in a single tangent space, so there's no concern about the base points of the vectors moving.

Proposition 14.5.4. *Suppose X is a vector field, and suppose that Ψ_s is a one-parameter group of diffeomorphisms of M with generating vector field U. Then we can define the* Lie derivative of X with respect to U *as a vector field $\mathcal{L}_U X$ satisfying*

$$(\mathcal{L}_U X)_p = -\frac{\partial}{\partial s}\Big|_{s=0}\big((\Psi_s)_{\#}X\big)_p. \tag{14.5.4}$$

In particular, this quantity only depends on U, not on the flow Ψ_s.

Furthermore, we can compute $\mathcal{L}_U X$ explicitly using the operator formula

$$\mathcal{L}_U X = [U, X] \tag{14.5.5}$$

where $[U, X]$ is the Lie bracket defined by Proposition 14.2.1.

Proof. This is surprisingly difficult to prove in a general coordinate chart, but if we use correctly adapted coordinates $\mathbf{y} = \psi$ where $U = \frac{\partial}{\partial y^1}$, it becomes much easier.[31] In such a chart, write

$$X \circ \psi^{-1}(y^1, \dots, y^n) = \sum_{k=1}^n a^k(y^1, \dots, y^n) \left.\frac{\partial}{\partial y^k}\right|_{\psi^{-1}(y^1,\dots,y^n)}.$$

Since in adapted coordinates the flow Ψ_s takes the form

$$\psi \circ \Psi_s \circ \psi^{-1}(y^1, y^2, \dots, y^n) = (y^1 + s, y^2, \dots, y^n),$$

we have

$$(\Psi_s)_* \left(\left.\frac{\partial}{\partial y^i}\right|_q \right) = \left.\frac{\partial}{\partial y^i}\right|_{\Psi_s(q)}$$

for every i and every q. Hence in particular we know that

$$\begin{aligned}\big((\Psi_s)_\# X\big)_q = (\Psi_s)_*\big(X_{\Psi_{-s}(q)}\big) &= (\Psi_s)_* \left(\sum_{k=1}^n a^k(\psi(\Psi_{-s}(q))) \left.\frac{\partial}{\partial y^k}\right|_{\Psi_{-s}(q)} \right) \\ &= \sum_{k=1}^n a^k\big(\psi(\Psi_{-s}(q))\big) \left.\frac{\partial}{\partial y^k}\right|_q = \sum_{k=1}^n a^k(y^1 - s, y^2, \dots, y^n) \left.\frac{\partial}{\partial y^k}\right|_q.\end{aligned}$$

Differentiate this with respect to s, then set $s = 0$, to obtain

$$\left.\frac{\partial}{\partial s}\big(\Psi_s)_\# X\big)_q\right|_{s=0} = \frac{\partial}{\partial s} \sum_{k=1}^n a^k(y^1 - s, y^2, \dots, y^n) \left.\frac{\partial}{\partial y^k}\right|_q = -\sum_{k=1}^n \frac{\partial a^k}{\partial y^1}(y^1, \dots, y^n) \left.\frac{\partial}{\partial y^k}\right|_q.$$

Hence by definition we have

$$(\mathcal{L}_U X)_q = \sum_{k=1}^n \frac{\partial a^k}{\partial y^1}(y^1, \dots, y^n) \left.\frac{\partial}{\partial y^k}\right|_q, \qquad \text{where } \psi(q) = (y^1, \dots, y^n).$$

On the other hand, computing $[U, X]$ we obtain

$$\begin{aligned}[U, X]_q(f) &= \frac{\partial}{\partial y^1}\Big(\sum_{k=1}^n a^k(y^1, \dots, y^n) \frac{\partial (f \circ \psi^{-1})}{\partial y^k} \Big) - \sum_{k=1}^n a^k(y^1, \dots, y^n) \frac{\partial^2 (f \circ \psi^{-1})}{\partial y^1 \partial y^k} \\ &= \sum_{k=1}^n \frac{\partial a^k}{\partial y^1}(y^1, \dots, y^n) \left.\frac{\partial (f \circ \psi^{-1})}{\partial y^k}\right|_{\mathbf{y}(q)},\end{aligned}$$

which implies that

$$[U, X]_q = \sum_{k=1}^n \frac{\partial a^k}{\partial y^1}(y^1, \dots, y^n) \left.\frac{\partial}{\partial y^k}\right|_q \qquad \text{where } \psi(q) = (y^1, \dots, y^n).$$

We thus conclude that $[U, X] = \mathcal{L}_U X$. □

[31] This is one of the best reasons to use the Straightening Theorem 14.5.1.

It ends up being a nice surprising coincidence that two different invariant formulas (14.5.4) and (14.5.5) end up being the same, especially since we proved it using a very special coordinate system. Of course, since both objects are defined invariantly, the fact that they are equal in one coordinate chart means that they are equal in all coordinate charts. The important thing was to establish that both objects were actually vector fields (in the sense that they could differentiate functions on the manifold), and once we know they exist as vector fields, it does not matter what coordinates we use to compute them.

We now return to the original situation of determining whether flows commute based on whether the corresponding vector fields commute.

Proposition 14.5.5. *Suppose X and U are vector fields on M with local flows Φ_t and Ψ_s respectively. Then we have $\Phi_t \circ \Psi_s = \Psi_s \circ \Phi_t$ for all t and s where this makes sense if and only if we have $[U, X] = 0$.*

Proof. Suppose $\Phi_t \circ \Psi_s = \Psi_s \circ \Phi_t$ for all values of t and s where it makes sense. Then if $\Xi_t = \Psi_s \circ \Phi_t \circ \Psi_{-s}$, we know that $\Xi_t = \Phi_t$. This implies by Proposition 14.5.3 that $(\Psi_s)_\# X = X$ for all s. Differentiate with respect to s and set $s = 0$ to obtain $\mathcal{L}_U X = 0$, which implies that $[U, X] = 0$.

Conversely assume $[U, X] = 0$. Then we conclude that $\mathcal{L}_U X = 0$, which means that $\frac{\partial}{\partial s}(\Psi_s)_\# X = 0\big|_{s=0} = 0$. However since Ψ_s is a group, we can conclude that $\frac{\partial}{\partial s}(\Psi_s)_\# X = 0$ for *every* s. Thus $(\Psi_s)_\# X = X$ since this is true at $s = 0$. Hence the flow of each side is the same, but we know that the flow of the left side is $\Psi_s \circ \Phi_t \circ \Psi_{-s}$ if the flow of the right side is Φ_t by Proposition 14.5.3. □

Example 14.5.6. If $X = (x^2 + y^2)\left(x \frac{\partial}{\partial x} + y \frac{\partial}{\partial y}\right)$ and $U = -y \frac{\partial}{\partial x} + x \frac{\partial}{\partial y}$, then we can compute directly that $[U, X] = 0$ without even knowing the flows of U or X. This tells us that if we choose coordinates such that $U = \frac{\partial}{\partial y^1}$, and if X is expressed in those coordinates by $X = f(y^1, y^2) \frac{\partial}{\partial y^1} + g(y^1, y^2) \frac{\partial}{\partial y^2}$, then $\frac{\partial f}{\partial y^1} = \frac{\partial g}{\partial y^1} = 0$. We conclude that the components of X actually depend only on y^2, so we can write the differential equation generating its flow as

$$\frac{dy^1}{dt} = f(y^2) \qquad \frac{dy^2}{dt} = g(y^2).$$

The second equation can generically be solved explicitly for y^2, which leads to a solution of the first equation for y^1. ☺

The lesson is that if X is some arbitrary vector field such that $[U, X] = 0$ for some vector field U, and if the flow of U can be computed explicitly, then X simplifies greatly in the coordinates generated by the flow of U. So the general strategy for finding the flow of X explicitly is to find symmetries generated by vector fields U with $[U, X] = 0$, and if X is preserved by these symmetries, then we can reduce X to an effectively-lower-dimensional vector field. The technique of Lie works in greater generality, but this is the basic idea.

Now the proofs above are correct, but the drawback is that it's hard to see how it might actually occur to anyone that Lie brackets are related to commuting vector fields. Let's work this out in the general one-dimensional case (without any intelligent choice of coordinates) to get a sense of it.

Example 14.5.7. Suppose $M = \mathbb{R}$ with vector fields $X = f(x)\frac{\partial}{\partial x}$ and $U = g(x)\frac{\partial}{\partial x}$. Then the flow of X satisfies

$$\frac{\partial \Phi}{\partial t}(t,x) = f\big(\Phi(t,x)\big), \qquad \Phi(0,x) = x.$$

We can solve this differential equation in a power series since the conditions give $\Phi(0,x) = x$ and $\frac{\partial \Phi}{\partial t}(0,x) = f(x)$. We thus obtain the solution

$$\Phi(t,x) = x + tf(x) + O(t^2).$$

Similarly for the flow Ψ of U we get

$$\Psi(s,x) = x + sg(x) + O(s^2).$$

Now expand the functions $\Phi\big(t,\Psi(s,x)\big)$ and $\Psi\big(s,\Phi(t,x)\big)$ in power series in both s and t. We get

$$\begin{aligned}\Phi\big(t,\Psi(s,x)\big) &= \Psi(s,x) + tf\big(\Psi(s,x)\big) + O(t^2)\\ &= x + sg(x) + O(s^2) + tf(x + sg(x) + O(s^2)) + O(t^2)\\ &= x + sg(x) + tf(x) + tsf'(x)g(x) + O(s^2+t^2).\end{aligned}$$

Similarly we get

$$\Psi\big(s,\Phi(t,x)\big) = x + sg(x) + tf(x) + tsg'(x)f(x) + O(s^2+t^2),$$

and thus in coordinates we have

$$\Phi\big(t,\Psi(s,x)\big) - \Psi\big(s,\Phi(t,x)\big) = st\big[g(x)f'(x) - f(x)g'(x)\big].$$

Hence if the flows commute, we must have $gf' - fg' = 0$, which is the coordinate expression of $[U,X]$ from Example 14.2.2. It thus becomes obvious that the flows can only commute if the Lie bracket is zero; it is less obvious that the Lie bracket being zero implies the flows commute, but this eventually comes from the fact that flows are a one-parameter group. ☺

We conclude with the following application, showing how to straighten multiple commuting vector fields simultaneously.

Theorem 14.5.8. *Suppose X and Y are vector fields on M such that $[X,Y] = 0$. Then in a neighborhood of any point p, there is a coordinate chart such that $X = \frac{\partial}{\partial y^1}$ and $Y = \frac{\partial}{\partial y^2}$ everywhere on the coordinate domain.*

Proof. There are two ways to prove this. One is to apply the Straightening Theorem 14.5.1 to obtain a chart $\mathbf{x}$ such that $X = \frac{\partial}{\partial x^1}$. Then if Y is expressed in coordinates as $Y_q = \sum_k b^k(x^1,\dots,x^n)\frac{\partial}{\partial x^k}$ for $\phi(q) = (x^1,\dots,x^n)$, we can check that

$$[X,Y] = \sum_k \frac{\partial b^k}{\partial x^1}(x^1,\dots,x^n)\frac{\partial}{\partial x^k}.$$

So the condition $[X,Y] = 0$ means that each component b^k of Y is independent of x^1, and thus we can view Y as actually defined on an $(n-1)$-dimensional manifold. Then applying the straightening trick again, we can rearrange the coordinates $(x^2,\dots,x^n)$ to straighten out Y without changing X.

The other way to proceed is to consider a parametrized $(n-2)$-dimensional surface in M given by $F\colon \Omega \subset \mathbb{R}^{n-2} \to M$ such that neither X nor Y are ever contained in any tangent space to $F[\Omega]$. Then for any coordinate chart ϕ, the map

$\phi \circ \Phi_{y^1}(\Psi_{y^2}\big(F(y^3, \ldots, y^n)\big))$ is locally an invertible map from an open subset of $\mathbb{R}^n$ to itself and thus generates a coordinate chart. The vector field X is $\frac{\partial}{\partial y^1}$ in these coordinates obviously, and since Φ commutes with Ψ, we can also conclude that Y is $\frac{\partial}{\partial y^2}$. □

Clearly we can apply the same sort of procedure given any number of commuting vector fields. We thus obtain the following useful characterization: if there are k vector fields X_i such that near every point there is a coordinate chart with $X_i = \frac{\partial}{\partial x^i}$ for $1 \leq i \leq k$, then $[X_i, X_j] = 0$ for all i and j; conversely if these fields commute then there is a chart near every point such that $X = \frac{\partial}{\partial x^i}$.

15. Differential forms

"You will find that many of the truths we cling to depend greatly on our own point of view."

15.1. **The cotangent space T_p^*M and 1-forms.** Denote by $\mathcal{F}(M)$ the space of smooth functions $f\colon M \to \mathbb{R}$ and by $\chi(M)$ the space of smooth vector fields on M. In Sections 14.1–14.2, we saw how to characterize functions as linear operators $L\colon \mathcal{F}(M) \to \mathcal{F}(M)$ satisfying $L(fg) = fL(g)$ for all $f, g \in \mathcal{F}(M)$ (that is, tensorial operators). We also saw how to characterize vector fields as derivations, that is, linear operators $D\colon \mathcal{F}(M) \to \mathcal{F}(M)$ such that $D(fg) = fD(g) + gD(f)$ for all $f, g \in \mathcal{F}(M)$. We thus characterized $\mathcal{F}(M)$ as certain operators on $\mathcal{F}(M)$, and $\chi(M)$ as certain other operators on $\mathcal{F}(M)$. It is natural to look at linear operators from $\chi(M)$ to $\mathcal{F}(M)$. Operators $L\colon \chi(M) \to \mathcal{F}(M)$ that satisfy tensoriality, that is $L(fX) = fL(X)$ for all $X \in \chi(M)$ and $f \in \mathcal{F}(M)$, must arise from pointwise-defined linear operators from each T_pM to $\mathbb{R}$, and hence correspond to covector spaces; putting these together leads to a cotangent bundle, and sections are covector fields. And as we saw in Chapter 4, once we can build a covector space out of a vector space, we can build spaces of tensors as well. Having the tangent bundle and cotangent bundle will then give us tensor bundles in a very natural way.

Let $g\colon M \to \mathbb{R}$ be a smooth function. Consider the operator $L\colon \chi(M) \to \mathcal{F}(M)$ given by $L(X) = X(g)$. Certainly L is tensorial in the sense of (14.2.4), since for any smooth $f\colon M \to \mathbb{R}$, we have $L(fX) = (fX)(g) = fX(g) = fL(X)$. In previous cases, once we found a large class of operators which satisfied some nice algebraic property, it was easy to prove that all operators with that algebraic property must be of the form we found. In this case it will not be true that all tensorial operators from $\chi(M)$ to $\mathcal{F}(M)$ are given by this simple form, and that failure is one of the deepest results of Differential Geometry.

To understand this, let's first try to understand what should be happening point by point, in order to understand how to build a bundle structure.

Definition 15.1.1. Suppose $p \in M$ is any point, and let f be a germ of a function at p. That is, there is some open set $U \ni p$ and a smooth function $f\colon U \to \mathbb{R}$. We define $df_p\colon T_pM \to \mathbb{R}$ to be the operator $df|_p(v) = v(f)$. The operator $df|_p$ is linear, and hence $df|_p \in T_p^*M$, the dual space of T_pM. It is called the *differential of f at p*. In general, elements of T_p^*M are usually called *cotangent vectors* or 1-forms.

All we'd have to check is that $df|_p$ is linear, but that actually follows exactly from the definition of the vector space structure on T_pM given by Definition 10.3.6. The only strange thing about this is that we write T_p^*M instead of $(T_pM)^*$, but the former notation is more convenient since we will soon want to put them together.

Example 15.1.2. Suppose $f\colon \mathbb{C} \to \mathbb{R}$ is given by $f(z) = \operatorname{Im}(e^z)$. Then in coordinates we have $f(x, y) = e^x \sin y$. Writing $v = a\,\frac{\partial}{\partial x}|_0 + b\,\frac{\partial}{\partial y}|_0$, we have

$$df|_0(v) = v(f) = a\,\frac{\partial(e^x \sin y)}{\partial x}(0,0) + b\,\frac{\partial(e^x \sin y)}{\partial y}(0,0) = b.$$

Hence the cotangent vector satisfies

$$df|_0\left(\frac{\partial}{\partial x}\Big|_0\right) = 0, \qquad df|_0\left(\frac{\partial}{\partial y}\Big|_0\right) = 1,$$

or in other words it is the dual covector to $\frac{\partial}{\partial y}|_0$. ☺

Suppose g is a smooth function. Then there is a map dg given by $p \mapsto dg|_p \in T_p^*M$. If we put all the cotangent spaces T_p^*M into a bundle T^*M, then dg would be a section of this bundle, and we would have globally that for any vector field X on M, there would be a function $dg(X)$ which at $p \in M$ is given by $dg_p(X_p)$. To understand what this bundle needs to be, we first want to construct an explicit basis at each point, which allows us to build the local trivializations.

Proposition 15.1.3. *Suppose $(\mathbf{x} = \phi, U)$ is a coordinate chart on a manifold M, and let $dx^k|_p$ denote the differentials of the coordinate functions x^k at p as in Definition 15.1.1. Then $\{dx^k|_p\}$ is the dual basis for T_p^*M of the basis $\{\frac{\partial}{\partial x^j}\big|_p\}$ of T_pM. In other words, we have*

$$dx^k|_p\left(\frac{\partial}{\partial x^j}\Big|_p\right) = \delta_j^k. \tag{15.1.1}$$

Furthermore, for any germ f of a function at p, we have

$$df|_p = \sum_{k=1}^{n} \frac{\partial(f \circ \phi^{-1})}{\partial x^k}\Big|_{\phi(p)} dx^k|_p. \tag{15.1.2}$$

We often write just

$$df = \sum_{k=1}^{n} \frac{\partial f}{\partial x^k}\, dx^k$$

if no confusion can result.

Proof. By definition we have

$$dx^k|_p\left(\frac{\partial}{\partial x^j}\Big|_p\right) = \frac{\partial}{\partial x^j}\left(\phi^k \circ \phi^{-1}(x^1, \ldots, x^n)\right)\Big|_{\phi(p)} = \frac{\partial x^k}{\partial x^j}\Big|_{\phi(p)} = \delta_j^k|_{\phi(p)} = \delta_j^k.$$

This is what we wanted to show; since we already know abstractly that T_p^*M is n-dimensional, it follows that $dx^k|_p$ must be its basis.

Now let us compute $df|_p$ in terms of this basis for an arbitrary f. By definition we have

$$df|_p\left(\frac{\partial}{\partial x^j}\Big|_p\right) = \frac{\partial}{\partial x^j}\Big|_p(f) = \frac{\partial(f \circ \phi^{-1})}{\partial x^j}\Big|_{\phi(p)}.$$

Therefore both sides of

$$df|_p = \sum_{k=1}^{n} \frac{\partial(f \circ \phi^{-1})}{\partial x^k}\Big|_{\phi(p)} dx^k|_p,$$

give the same result when applied to the vector $\frac{\partial}{\partial x^j}\big|_p$, for any j, and thus they must actually represent the same covector. □

Having the general formula (15.1.2) then implies that if f happens to be a component y^j of a coordinate chart, then we can express $dy^j|_p$ in terms of the basis $\{dx^k|_p\}$. This formula is the dual of the formula from Corollary 10.4.5; in fact once we know how basis vectors in T_pM change under a change of coordinates, we

clearly must know exactly how the dual vectors change as in Chapter 4. Hence the following corollary is not only obvious but has two different obvious proofs.

Corollary 15.1.4. *If $(\mathbf{x}, U)$ and $\mathbf{y}, V)$ are two coordinate charts with $p \in U \cap V$, then the covectors in the respective dual bases of T_p^*M, written as $\{dx^1|_p, \dots, dx^n|_p\}$ and $\{dy^1|_p, \dots, dy^n|_p\}$, satisfy the transition formula*

$$dx^j\big|_p = \sum_{k=1}^n \frac{\partial x^j}{\partial y^k}\Big|_{\mathbf{y}(p)} dy^k\big|_p. \tag{15.1.3}$$

Observe how natural this notation looks; in fact in calculus the same formula appears, although the notation "dx" is never really defined. *This* is the proper way to think of differentials: as operators on vectors. The only price we pay is the rather awkward-looking formula (15.1.1), which you should get used to.

There is one more thing that is easy to check while we are still working with cotangent vectors at a single point. In Definition 15.1.1, we started with a function f defined in a neighborhood of p and obtained an element $df|_p$ of the cotangent vector space T_p^*M. Having obtained the basis for the cotangent space in Proposition 15.1.3, we can now establish the converse of this: every cotangent vector in T_pM is $df|_p$ for some smooth $f\colon M \to \mathbb{R}$. Since the function is globally-defined, we don't need to work with germs anymore.

Proposition 15.1.5. *Let M be a smooth n-dimensional manifold, and let $p \in M$ be any point. Then every element $\alpha \in T_p^*M$ is given by $df|_p$ for some (non-unique) smooth function $f\colon M \to \mathbb{R}$.*

Proof. Let (ϕ, U) be a coordinate chart in a neighborhood of p. By Proposition 15.1.3, every element $\alpha \in T_p^*M$ is given by

$$\alpha = \sum_{k=1}^n a_k\, dx^k|_p$$

for some numbers $a_1, \dots, a_n$. Let $\zeta\colon M \to \mathbb{R}$ be a smooth bump function which is identically equal to one in some neighborhood V of p with $V \subset U$. Define

$$f(q) = \zeta(q) \sum_{k=1}^n a_k \phi^k(q) \quad \text{if } q \in U \text{ and zero otherwise.}$$

Then f is smooth, and $f|_V(q) = \sum_{k=1}^n a_k \phi^k(q)$, so that

$$df|_p = \sum_{k=1}^n \sum_{j=1}^n a_k \frac{\partial \phi^k \circ \phi^{-1}}{\partial x^j}\Big|_{\mathbf{x}=\phi(p)} dx^j|_p = \sum_{k=1}^n \sum_{j=1}^n a_k \delta_j^k dx^j|_p = \sum_{k=1}^n a_k dx^k|_p = \alpha.$$

□

Given a smooth function f defined in a neighborhood of p for which p is not a critical point (i.e., $df|_p \neq 0$), we can use the Inverse Function Theorem to construct a coordinate chart $(x^1, \dots, x^n)$ on some neighborhood V of p such that $f = x^1$ on V. Hence every cotangent vector at p can always be assumed to be $dx^1|_p$ in the right coordinates. Of course we could do the same thing with tangent vectors: every $v \in T_pM$ can be written in the correct coordinate chart as $v = \frac{\partial}{\partial x^1}\big|_p$.

But we can actually do much more with vectors: recall from the Straightening Theorem 14.5.1 that for any vector field V with $V_p \neq 0$, we can choose a coordinate

chart around p such that $V_q = \frac{\partial}{\partial x^1}\big|_q$ for *every* $q \in V$. We might hope that the same sort of thing works for fields of covectors: that is, given a smooth map of covectors, there is locally a coordinate chart such that $\omega_q = dx^1|_q$ for every q in some open set. This is not at all true, and in fact this failure is why so many of the interesting things one can do in Differential Geometry involve cotangent vectors rather than tangent vectors. We will see this in the next few Chapters.

15.2. **The cotangent bundle T^*M and 1-form fields.** We have obtained the derivative of a function f at a point p, as a covector (1-form) in T^*_pM, which is a linear function on T_pM. This is analogous to starting with a function $f\colon \mathbb{R} \to \mathbb{R}$ and defining the derivative at a particular point, $f'(a)$. In calculus, once we do this, we quickly move to thinking of f' as *another* function $f'\colon \mathbb{R} \to \mathbb{R}$, just by letting the number a vary. We'd like to perform the same generalization here: to have a derivative operator d which takes a function f to a field of 1-forms defined on all of M, by letting the base point p vary.

What exactly is such an object? Well, first, we want to patch together all of the cotangent spaces T^*_pM to get a single cotangent bundle T^*M, in exactly the same way as we did in Definition 12.1.4 to get the tangent bundle TM. It will end up being just another vector bundle, with different trivializations. Our definition is inspired in the same way as it was for TM: we begin with coordinate charts (ϕ, U) on M and obtain charts $(\tilde{\Phi}, T^*U)$ on T^*M by just copying the components in the coordinate basis.

Definition 15.2.1. If M is a smooth n-dimensional manifold, the *cotangent bundle* T^*M is defined to be the disjoint union of all the cotangent spaces T^*_pM. For every coordinate chart (ϕ, U) on M, we define a coordinate chart $(\tilde{\Phi}, T^*U)$ by the formula

$$\tilde{\Phi}\left(\sum_{k=1}^{n} a_k\, dx^k|_p\right) = (\phi(p), a_1, \ldots, a_n) \in \mathbb{R}^{2n}. \tag{15.2.1}$$

The topology is generated by declaring a set $\Omega \subset TM$ to be open if and only if $\tilde{\Phi}[\Omega \cap T^*U]$ is open for every chart $(\tilde{\Phi}, T^*U)$, and this topology makes all the functions $\tilde{\Phi}$ continuous and smooth and gives T^*M the structure of a smooth manifold.

The related local trivializations $\zeta\colon T^*U \to U \times \mathbb{R}^n$ given by

$$\zeta(\left(\sum_{k=1}^{n} a_k\, dx^k|_p\right) = (p, a_1, \ldots, a_n) \in \mathbb{R}^{2n}$$

make T^*M into an n-dimensional vector bundle over M as in Definition 12.1.5.

As in the Definition 12.1.4 of the tangent bundle, the main thing to check is that the transition maps are smooth on $\mathbb{R}^{2n}$. Using the transition formula from Corollary 15.1.4, we see that if the same covector α is expressed in two different charts

$$\alpha = \sum_{k=1}^{n} a_k\, dx^k|_p = \sum_{j=1}^{n} b_j\, dy^j|_p,$$

then the coefficients are related by

$$\alpha = \sum_{k=1}^{n}\sum_{j=1}^{n} a_k \left.\frac{\partial x^k}{\partial y^j}\right|_{\mathbf{y}(p)} dy^j\big|_p = \sum_{j=1}^{n} b_j\, dy^j\big|_p,$$

so that we have

$$b_j = \sum_{k=1}^{n} a_k \left.\frac{\partial x^k}{\partial y^j}\right|_{\mathbf{y}(p)}.$$

Hence we can write the transition map as

$$\begin{aligned}(y^1,\dots,y^n,b_1,\dots,b_n) &= \tilde{\Phi}\circ\Psi^{-1}(x^1,\dots,x^n,a_1,\dots,a_n)\\ &= \left(\phi\circ\psi^{-1}(x^1,\dots,x^n), \sum_{k=1}^{n} a_k\,\frac{\partial x^k}{\partial y^1},\dots,\sum_{k=1}^{n} a_k\,\frac{\partial x^k}{\partial y^n}\right).\end{aligned}$$

Note that from this perspective the functions $\frac{\partial x^k}{\partial y^j}$ are actually being viewed as functions of $\mathbf{x}$; hence we are really viewing them as the inverse matrix of $\frac{\partial y^j}{\partial x^k}(\mathbf{x})$, which exists and is smooth by the Inverse Function Theorem. Hence the transition maps are smooth and in particular continuous, so that they preserve the topological structure.

Now let's discuss the spaces of vector fields and 1-forms. These will all have a vector-space structure which is infinite-dimensional. The topology is a bit involved (there are many inequivalent choices, and no reason to prefer one over another), and it is not really necessary for anything we do, so we will ignore it.

Definition 15.2.2. Suppose M is an n-dimensional manifold.

- The *space of smooth functions* on M with values in $\mathbb{R}$ is denoted by $\mathcal{F}(M)$; the vector space operation is $(af+bg)(p) = af(p)+bg(p)$.
- The *space of vector fields* on M is denoted by $\chi(M)$ and consists of all smooth vector fields defined on all of M. The vector space operation is $(aX+bY)_p = aX_p + bY_p$.
- The *space of* 1*-forms* on M is denoted by $\Omega^1(M)$ and consists of all smooth sections of the cotangent bundle T^*M. The vector space operation is $(a\alpha_p + b\beta_p)(X_p) = a\alpha_p(X_p)+b\beta_p(X_p)$ for all $X_p \in T_pM$. Since "1-form" can mean either a single covector in T_p^*M or a section of the cotangent bundle, we will sometimes use "covector field" or "1-form field" in case it's ambiguous.

Given a 1-form ω and a vector field X, we can apply ω_p to X_p at each point to get a real number $\omega_p(X_p)$. Hence globally we can apply ω to X to get a function on M. Obviously the resulting function does not depend on the choice of coordinates used to describe it, but that's because we went to so much trouble to define vector fields and 1-forms in a coordinate-invariant way. Classically one would have written (in two different coordinate charts $\mathbf{x}$ and $\mathbf{y}$) the formulas $\omega = \sum_j \alpha_j\,dx^j = \sum_k \beta_k\,dy^k$ and $X = \sum_j v^j\,\frac{\partial}{\partial x^j} = \sum_k w^k\,\frac{\partial}{\partial y^k}$, using the transition formulas (10.4.6) and (15.1.3) to relate β to α and w to v. Then using the awkward formula (15.1.1), we can check that $\sum_j \alpha_j v^j = \sum_k \beta_k w^k$ and hence the common value gives a well-defined notion of $\omega(X)$. Let's see how this actually works in practice.

Example 15.2.3. Consider as an example the vector field X on $\mathbb{R}^2$ given in Cartesian coordinates by $X = x\,\frac{\partial}{\partial x} - y\,\frac{\partial}{\partial y}$, and the 1-form $\omega = xy\,dx + y^2\,dy$. Then $\omega(X)$ is a function on $\mathbb{R}^2$, given in Cartesian coordinates by

$$\omega(X)(x,y) = (xy\,dx + y^2\,dy)\left(x\,\frac{\partial}{\partial x} - y\,\frac{\partial}{\partial y}\right) = x^2y - y^3.$$

Let's check that we get the same function in polar coordinates. By the transition formulas (15.1.3) and (10.4.6), the 1-form and vector field become

$$\omega = r^2 \cos\theta \sin\theta \,(\cos\theta\, dr - r\sin\theta\, d\theta) + r^2 \sin^2\theta\,(\sin\theta\, dr + r\cos\theta\, d\theta) = r^2 \sin\theta\, dr$$

and

$$\begin{aligned} X &= r\cos\theta \left(\cos\theta \frac{\partial}{\partial r} - \frac{\sin\theta}{r}\frac{\partial}{\partial\theta}\right) - r\sin\theta\left(\sin\theta\frac{\partial}{\partial r} + \frac{\cos\theta}{r}\frac{\partial}{\partial\theta}\right) \\ &= r\cos 2\theta \frac{\partial}{\partial r} - \sin 2\theta \frac{\partial}{\partial\theta}. \end{aligned}$$

Thus we have in polar coordinates that

$$\omega(X) = r^2 \sin\theta\, dr \left(r\cos 2\theta \frac{\partial}{\partial r} - \sin 2\theta \frac{\partial}{\partial\theta}\right) = r^3 \sin\theta\cos 2\theta.$$

This agrees with our previous computation $\omega(X) = x^2y - y^3 = y(x^2 - y^2) = r^3\sin\theta\cos 2\theta$, as expected. ☺

Just as we did in Propositions 14.1.2–14.1.3, we can characterize smoothness of 1-forms in terms of smoothness of their components in a chart or in terms of whether they give a smooth function when applied to a smooth vector field. These characterizations are both more elegant and easier to use.

Proposition 15.2.4. *Let M be an n-dimensional manifold with cotangent bundle T^*M. Suppose $\omega\colon M \to T^*M$ is a (not necessarily smooth) function such that $\omega_p \in T_p^*M$ for every $p \in M$. Then ω is a smooth 1-form if and only if around every point of M there is a coordinate chart (ϕ, U) such that*

$$\omega_p = \sum_{k=1}^{n} \alpha_k(p)\, dx^k|_p$$

with the functions $\alpha_k\colon U \to \mathbb{R}$ all smooth.

Furthermore, ω is a smooth 1-form if and only if for every smooth vector field X on M, the function $\omega(X)$ defined by $p \mapsto \omega_p(X_p)$ is smooth.

Proof. The first part follows exactly as in the proof of Proposition 14.1.2: to check smoothness, we take a chart (ϕ, U) on M and the corresponding chart $(\tilde\Phi, T^*U)$ on T^*M, where

$$\tilde\Phi\circ\omega\circ\phi^{-1}(x^1,\dots,x^n) = \big(x^1,\dots,x^n, \alpha_1\circ\phi^{-1}(x^1,\dots,x^n),\dots,\alpha_n\circ\phi^{-1}(x^1,\dots,x^n)\big).$$

We must therefore have $\alpha_k \circ \phi^{-1}\colon \mathbb{R}^n \to \mathbb{R}$ smooth, which means $\alpha_k\colon U \to \mathbb{R}$ is smooth.

To prove the second part, note that if ω and X are smooth then the coefficients α_k of ω and the coefficients a^k of X are smooth, and in a coordinate neighborhood we have

$$\omega(X)(p) = \omega_p(X_p) = \sum_{k=1}^{n} \alpha_k(p) a^k(p), \tag{15.2.2}$$

which is a sum of products of smooth functions and hence smooth on U. Since $\omega(X)$ is smooth on each coordinate chart, it is smooth on M. Conversely suppose $\omega(X)$ is smooth for every smooth X. Then in a neighborhood of any point p contained in a chart (ϕ, U), there is a smooth vector field X_j such that $(X_j)_q = \frac{\partial}{\partial x^j}\big|_q$ for all q in some neighborhood V of p. Then on V we have $\omega(X_j)(q) = \alpha_j(q)$, so that α_j

is smooth on V. Since we can do this around every point of U, it follows that α_j is smooth on U, and since this is true for every j and every chart, we conclude that ω is smooth by the first part of the proposition. □

The proof shows us that every smooth 1-form ω generates a linear operator from the space of vector fields to the space of functions by $X \mapsto \omega(X)$. This action is often known as "contraction" because in coordinates it is given by (15.2.2), which eliminates the components as a dot product. We can characterize such operators completely using tensoriality. First we need a Lemma which gives us a Taylor expansion locally; we already used this to prove Proposition 14.1.4.

Lemma 15.2.5. *Let $p \in M$, and let X be a smooth vector field defined in a coordinate neighborhood (ϕ, U) of M such that $X_p = 0$ and $\phi(p) = 0$. Then there are smooth vector fields Y_k defined on U such that*

$$X_q = \sum_{k=1}^{n} \phi^k(q)(Y_k)_q \quad \text{for all } q \in U. \tag{15.2.3}$$

Proof. Just write $X_q = \sum_{j=1}^n v^j(q) \frac{\partial}{\partial x^j}$, and expand each a^j by the formula (14.1.2) to get

$$v^j(q) = \sum_{k=1}^{n} \phi^k(q) g_k^j(q),$$

since $v^j(p) = 0$. Then define $Y_j = \sum_{j=1}^n g_k^j \frac{\partial}{\partial x^j}$. □

Theorem 15.2.6. *If ω is a 1-form then the linear operator $L_\omega \colon \chi(M) \to \mathcal{F}(M)$ given by $L_\omega(X) = \omega(X)$ is tensorial, i.e., it satisfies*

$$L_\omega(fX) = f\omega(X) \tag{15.2.4}$$

for every $f \in \mathcal{F}(M)$. Conversely every linear operator $L \colon \chi(M) \to \mathcal{F}(M)$ which is tensorial is L_ω for some smooth 1-form.

Proof. Tensoriality of ω is immediate from the definition: we have

$$\omega(fX)(p) = \omega_p(f(p)X_p) = f(p)\omega_p(X_p)$$

since ω is just acting in each tangent space.

To prove the converse, we first want to show the essential property that if X is a smooth vector field such that for some $p \in M$ we have $X_p = 0$, then $\omega(X)(p) = 0$. (In fact this can be used as the definition of tensoriality.) First note that given X defined on M, we can multiply by a bump function ζ and obtain $L(\zeta X)(p) = \zeta(p)L(X)(p) = L(x)(p)$, so that we can assume X is supported in a coordinate neighborhood (ϕ, U) of p in order to compute $L(X)(p)$. Without loss of generality we can assume that $\phi(p) = 0$. Then we can use Lemma 15.2.5 to write $X = \sum_k \varphi^k Y_k$ in this coordinate neighborhood, and obtain

$$L(X)(p) = \sum_k L(\varphi^k Y_k)(p) = \sum_k \varphi^k(p) L(Y_k)(p) = 0$$

since each $\varphi^k(p) = 0$.

The implication is that if X and Y are two vector fields such that at some point p we have $X_p = Y_p$, then we must also have $L(X)(p) = L(Y)(p)$. So the value of $L(X)$ at p depends only on X_p. Now for each p define $\omega_p \colon T_pM \to \mathbb{R}$ by the formula $\omega_p(v) = L(X)(p)$ where X is any vector field on M such that $X_p = v$; as

we have just checked, this does not depend on choice of X. We can also check that ω_p is linear on vectors since L is linear on vector fields, and thus $\omega_p \in T_p^*M$.

We thus obtain a function $\omega\colon M \to T^*M$ such that $\omega_p \in T_p^*M$ for each p. Now for every smooth vector field X, we have $\omega(X) = L(X)$ for every vector field X by construction of ω, and since $L(X)$ is a smooth function for every X, we conclude by Proposition 15.2.4 that ω is a smooth 1-form. □

The simplest type of 1-form is our original motivating example: the derivative of a function. The following proposition just rephrases what we already know.

Proposition 15.2.7. *Suppose $f\colon M \to \mathbb{R}$ is a smooth function. Let $df\colon M \to T^*M$ denote the map $p \mapsto df|_p$ where each $df|_p$ is as defined in Definition 15.1.1. Then df is a smooth 1-form.*

Proof. We know that $df|_p \in T_p^*M$ for each p, so all we need to know by Proposition 15.2.4 is that $df(X)$ is a smooth function whenever X is a smooth vector field. But since $df(X) = X(f)$, we know this is true by Proposition 14.1.3: every smooth vector field differentiates smooth functions to give another smooth function. □

After all that, let's work out some 1-forms explicitly.

Example 15.2.8. First let's compute df for an explicit function f. Again we will work on $M = \mathbb{C}$; consider the coordinate-free function $f(z) = \operatorname{Re}(z^3)$. In Cartesian coordinates $\mathbf{x} = (x, y)$, we have $f \circ \mathbf{x}^{-1}(x, y) = \operatorname{Re}(x + iy)^3 = x^3 - 3xy^2$. By Proposition 15.1.3, we have

$$df = \frac{\partial f}{\partial x}\,dx + \frac{\partial f}{\partial y}\,dy = (3x^2 - 3y^2)\,dx - 6xy\,dy.$$

On the other hand, if we work in polar coordinates $\mathbf{u} = (r, \theta)$, then $f \circ \mathbf{u}^{-1}(r, \theta) = \operatorname{Re}(re^{i\theta})^3 = r^3 \cos 3\theta$, so that

$$df = \frac{\partial f}{\partial r}\,dr + \frac{\partial f}{\partial y}\,d\theta = 3r^2 \cos 3\theta\,dr - 3r^3 \sin 3\theta\,d\theta.$$

This must be the same as df in Cartesian coordinates; in fact the transition formula is (15.1.3), which gives

$$\begin{aligned}
df &= (3x^2 - 3y^2)\,dx - 6xy\,dy \\
&= (3x^2 - 3y^2)\left(\frac{\partial x}{\partial r}\,dr + \frac{\partial x}{\partial \theta}\,d\theta\right) - 6xy\left(\frac{\partial y}{\partial r}\,dr + \frac{\partial y}{\partial \theta}\,d\theta\right) \\
&= 3r^2(\cos^2\theta - \sin^2\theta)(\cos\theta\,dr - r\sin\theta\,d\theta) - 6r^2\cos\theta\sin\theta(\sin\theta\,dr + r\cos\theta\,d\theta) \\
&= \left[3r^2\cos 2\theta\cos\theta - 3r^2\sin 2\theta\sin\theta\right]dr + \left[-3r^3\cos 2\theta\sin\theta - 3r^3\sin 2\theta\cos\theta\right]d\theta \\
&= 3r^2\cos 3\theta\,dr - 3r^3\sin 3\theta\,d\theta,
\end{aligned}$$

as expected. ☺

Of course, not every 1-form is df for some function f, just as not every vector field is the gradient of a function.

Example 15.2.9. Consider the 1-form ω on $\mathbb{R}^2$ written in rectangular coordinates as

$$\omega = h(x, y)\,dx + j(x, y)\,dy \tag{15.2.5}$$

for some functions h and j. If $\omega = df$ for some function f, then since $df = \dfrac{\partial f}{\partial x}\,dx + \dfrac{\partial f}{\partial y}\,dy$, we must have

$$h(x,y) = \frac{\partial f}{\partial x}(x,y) \qquad \text{and} \qquad j(x,y) = \frac{\partial f}{\partial y}(x,y).$$

Since mixed partials commute by Theorem 5.1.9, we have

$$\frac{\partial h}{\partial y}(x,y) = \frac{\partial^2 f}{\partial y \partial x}(x,y) = \frac{\partial^2 f}{\partial x \partial y}(x,y) = \frac{\partial j}{\partial x}(x,y). \tag{15.2.6}$$

Obviously this is not true for an arbitrary ω.

For example, if $\omega_1 = 2xy\,dx + x^2\,dy$ then ω_1 satisfies condition (15.2.6), and in fact we have $\omega_1 = df_1$ where $f_1(x,y) = x^2 y$. On the other hand if $\omega_2 = -x^2\,dx + 2xy\,dy$, then we have $h_y(x,y) = 0$ and $j_x(x,y) = 2y$, so ω_2 cannot be df_2 for any function $f_2 \colon \mathbb{R}^2 \to \mathbb{R}$. ☺

In fact we can prove that on $\mathbb{R}^2$, any 1-form satisfying (15.2.6) must be the differential of a function on $\mathbb{R}^2$, as follows. This is a special case of the Poincaré Lemma.

Proposition 15.2.10. *Suppose ω is a 1-form on $\mathbb{R}^2$ with $\omega_{(x,y)} = h(x,y)\,dx + j(x,y)\,dy$ such that $\dfrac{\partial h}{\partial y}(x,y) = \dfrac{\partial j}{\partial x}(x,y)$ everywhere on $\mathbb{R}^2$. Then there is a smooth function $f \colon \mathbb{R}^2 \to \mathbb{R}$ such that $\omega = df$; in other words we have $h = \dfrac{\partial f}{\partial x}$ and $j = \dfrac{\partial f}{\partial y}$ everywhere on $\mathbb{R}^2$.*

Proof. We are expecting $h = f_x$, so just define the function to be the integral of h. That is, set $\tilde{f}(x,y) = \int_0^x h(u,y)\,du$ and see what happens. Obviously we have $\tilde{f}_x = h$, and by Theorem 5.3.4, we have

$$\frac{\partial \tilde{f}}{\partial y}(x,y) = \int_0^x \frac{\partial h}{\partial y}(u,y)\,du = \int_0^x \frac{\partial j}{\partial u}(u,y)\,du = j(x,y) - j(0,y).$$

This isn't quite what we want, so we correct by this function of y: set

$$f(x,y) = \tilde{f}(x,y) + \int_0^y j(0,v)\,dv$$

to obtain

$$\frac{\partial f}{\partial x} = \frac{\partial \tilde{f}}{\partial x} = h(x,y) \qquad \text{and} \qquad \frac{\partial f}{\partial y} = \frac{\partial \tilde{f}}{\partial y} + j(0,y) = j(x,y),$$

as desired. □

Remark 15.2.11. You may actually recognize this as a technique from a Differential Equations class: given a differential equation on $\mathbb{R}^2$ of the form

$$\frac{dx}{dt} = -N(x,y) \qquad \frac{dy}{dt} = M(x,y),$$

we can "cross-multiply" and write $\omega = M(x,y)\,dx + N(x,y)\,dy$, where the integral curves $\gamma(t)$ all satisfy $\omega(\gamma'(t)) = 0$.

If the components satisfy $\frac{\partial M}{\partial y} = \frac{\partial N}{\partial x}$ everywhere, then the differential equation is called "exact," and by the Poincaré Lemma Proposition 15.2.10, we have $\omega = df$ for some function $f(x,y)$. Since $\omega(\gamma'(t)) = 0$ for all t on an integral curve γ, we

have $df(\gamma'(t)) = 0$ which means $\frac{d}{dt}f(\gamma(t)) = 0$. Hence f is constant on integral curves, so that if $\gamma(t) = (x(t), y(t))$ then $f(x(t), y(t)) = C$ for some constant C (determined by the initial conditions).

We then try to solve $f(x, y) = C$ for y in terms of x (which can typically be done, by the Implicit Function Theorem 5.2.2) as $y = g(x)$, and then the differential equation becomes $\frac{dx}{dt} = -N(x, g(x))$. From here we can separate to get

$$\frac{dx}{N(x, g(x))} = -dt,$$

integrate both sides, and obtain t as a function of x, which can hopefully be inverted to get $x(t)$ explicitly, from which we get $y(t) = g(x(t))$ and an explicit formula for the integral curve $\gamma(t)$.

This is one of the most common ways to solve an autonomous system of two differential equations explicitly.

15.3. **Tensoriality and tensor fields.** Recall that we classified 1-forms in Theorem 15.2.6 as tensorial linear operators from the space $\chi(M)$ to the space $\mathcal{F}(M)$. The notion of tensoriality enables us to easily generalize 1-forms to multilinear operators on vector fields, and using duality, to operators on 1-forms. This essentially enables us to automatically deal with sections of tensor bundles without defining them directly.

Definition 15.3.1. A *tensor field* T on M of order (r, s) is a linear operator

$$T\colon \chi(M) \oplus \cdots \oplus \chi(M) \oplus \Omega^1(M) \oplus \cdots \oplus \Omega^1(M) \to \mathcal{F}(M),$$

where there are r factors of vector fields and s factors of 1-forms, such that T is tensorial (linear over C^∞ functions) in each term.

A 1-form is a tensor field of order $(1, 0)$, while a vector field is a tensor field of order $(0, 1)$ (since we can think of vector fields as operating on 1-forms just by switching perspective).

Example 15.3.2. Here are some common examples of tensor fields.

- An *inner product* is a $(2, 0)$ tensor field g which is symmetric: that is, $g(X, Y) = g(Y, X)$ for all vector fields X and Y. If $g(X_p, X_p) > 0$ for every vector field X and every point p, then g is called a Riemannian metric. If $g(X_p, Y_p) = 0$ for every Y_p implies $X_p = 0$, then g is called a nondegenerate metric.
- A 2-form is a $(2, 0)$ tensor field ω which is antisymmetric: $\omega(X, Y) = -\omega(Y, X)$ for all vector fields X and Y. This is the same as saying that ω_p is a 2-form on T_pM for every $p \in M$, as in Definition 4.3.1. If $\omega(X_p, Y_p) = 0$ for every Y_p implies $X_p = 0$, then ω is called a symplectic form.
- More generally we can do this for any k-form by imposing antisymmetry. In particular an n-form is an $(n, 0)$ tensor field which is completely antisymmetric.
- A tensor field L of order $(1, 1)$ can be identified with a smoothly varying family of linear operators from each T_pM to itself: if we know what $L(v, \alpha)$ is for each $v \in T_pM$ and $\alpha \in T_p^*M$, then we can identify the operator $\alpha \mapsto L(v, \alpha)$ with an element of $(T_p^*M)^*$, which by Proposition 4.1.9 is naturally isomorphic to T_pM. So we have a map from each T_pM to T_pM. The smoothness condition says that whenever X and ω are smooth, so is

$L(X, \omega)$. Given a vector field X we can define a vector field Y by $L(X, \omega) = \omega(Y)$, and thus we can identify L with an operator $\tilde{L}$ for which $Y = \tilde{L}(X)$.
- As an example the evaluation tensor E defined by $E(X, \omega) = \omega(X)$ is tensorial as is easily checked, and is thus a smooth tensor field of order $(1, 1)$, which is identified with the identity operator $\tilde{E}$ from $\chi(M)$ to $\chi(M)$.
- Similarly using the identification Proposition 4.1.9 between $(T_p^*M)^*$ and T_pM, any tensor field of order $(k, 1)$ can be viewed as a linear operator from k smooth vector fields to a single smooth vector field; thus our definition of tensor fields is more general than it may seem at first.

☺

Given any type of tensor field with any given antisymmetry or antisymmetry, we can construct a corresponding bundle. The point is that the bundle is built exactly as we built the bundles TM and T^*M: we use coordinate charts on M to get coordinate basis elements of our tensor spaces exactly as in Section 4.2, and we obtain local trivializations just by retaining the base point in the manifold but writing the tensor at that base point in components. For example, in a coordinate chart on M every symmetric $(2, 0)$ tensor can be written

$$g_p = \sum_{i=1}^{n} g_{ii}\, dx^i|_p \otimes dx^i|_p + \sum_{1 \leq i < j \leq n} g_{ij}\, (dx^i|_p \otimes dx^j|_p + dx^j|_p \otimes dx^i|_p),$$

and the numbers g_{ij} for $1 \leq i \leq j \leq n$ form a coordinate chart of the bundle $\mathrm{Sym}(TM)$ of symmetric $(2, 0)$ tensors on M. Transition maps can be computed by writing $dx^i \otimes dx^j$ in terms of $dy^k \otimes dy^\ell$ to prove smoothness, and we thus obtain a smooth bundle structure. Similarly we can construct the bundle of 2-forms over M using the basis elements $dx^i|_p \wedge dx^j|_p$ for $i < j$, or the bundle of $(1, 1)$ tensors over M using the basis $dx^i|_p \otimes \frac{\partial}{\partial x^j}\big|_p$ for $1 \leq i, j \leq n$. Sections of these bundles will then be tensor fields as we have been describing in terms of tensorial linear operators, so at this stage we don't really get any new information by discussing the bundles directly.

You may have noticed that we've been switching back and forth between coordinate descriptions of objects and invariant descriptions. For the most part, one can get by just using invariant descriptions: for example, we've defined functions and vectors without using coordinates, and most other objects are defined in terms of those coordinate-independent objects. However, we always need to be able to compute things in coordinates, and that's why the transition formulas are important. But they're important for another reason: we'd like to actually define new objects, and sometimes the coordinate definition is more convenient. Once we know that a formula is independent of coordinates, we can expect that there is an invariant definition. But first we need to know what we're looking for.

A prime example of this is our next definition. So far we've seen that we can generalize the gradient of a function f by thinking of the 1-form df. (We will make the analogy between $\operatorname{grad} f$ and df explicit in Chapter 19, when we discuss inner products.) We're now interested in taking the derivative of a vector field. In vector calculus in $\mathbb{R}^3$, we have two possibilities: the curl or the divergence. Since we have already seen that the derivative of a function is most naturally thought of as a 1-form (not a vector field), it's natural to expect to take derivatives of 1-forms rather than vector fields.

Thinking about what such a derivative should look like, we observe that when we want to take a derivative of a function, we need to specify a direction (and thus a vector) in which to take the derivative. So when we want to take a derivative of a 1-form, it makes sense we'll have to specify a direction in which to do it: thus the derivative of a 1-form should be a linear map from vectors to 1-forms, or in other words, a map from a pair of vectors to real numbers. So we should get a tensor of type $(2,0)$. What does such a thing look like?

Example 15.3.3. A tensor of type $(2,0)$ is determined by its operation on two vector fields; since it must be linear over functions, it is determined by its operation on basis vectors. Specifically, take a coordinate system $(\mathbf{x}, U)$ and write vector fields X and Y as $X = \sum_{j=1}^n X^j \frac{\partial}{\partial x^j}$ and $Y = \sum_{k=1}^n Y^k \frac{\partial}{\partial x^k}$. Then if α is a tensor field of type $(2,0)$, we have

$$\begin{aligned}\alpha(X,Y)(p) &= \alpha\left(\sum_{j=1}^n X^j(p)\,\frac{\partial}{\partial x^j}\Big|_p, \sum_{k=1}^n Y^k(p)\,\frac{\partial}{\partial x^k}\Big|_p\right)\\ &= \sum_{j=1}^n\sum_{k=1}^n X^j(p)Y^k(p)\,\alpha\left(\frac{\partial}{\partial x^j}\Big|_p, \frac{\partial}{\partial x^k}\Big|_p\right),\end{aligned}$$

and if we define

$$\alpha_{jk}(p) = \alpha\left(\frac{\partial}{\partial x^j}\Big|_p, \frac{\partial}{\partial x^k}\Big|_p\right),$$

we can write

$$\alpha = \sum_{j=1}^n\sum_{k=1}^n \alpha_{jk}(p)\,dx^j|_p \otimes dx^k|_p,$$

where the direct product of 1-forms is defined as

$$dx^j|_p \otimes dx^k|_p(X_p, Y_p) = dx^j|_p(X_p)\cdot dx^k|_p(Y_p).$$

This direct product provides a basis for the tensors at each point just as in Chapter 4, and this works in any coordinate chart.

Now suppose we have a different coordinate chart $(\mathbf{y}, V)$ overlapping U. Then we can write $X = \sum_{j=1}^n \tilde{X}^j \frac{\partial}{\partial y^j}$ and $Y = \sum_{k=1}^n \tilde{Y}^k \frac{\partial}{\partial x^k}$, where the basis vectors transform according to (10.4.6), and we must have

$$\alpha(X,Y) = \sum_{j,k=1}^n \tilde{X}^j\tilde{Y}^k\,\alpha\big(\tfrac{\partial}{\partial y^j}, \tfrac{\partial}{\partial y^k}\big) = \sum_{j,k=1}^n \tilde{\alpha}_{jk}\tilde{X}^j\tilde{Y}^k.$$

We thus end up with

(15.3.1)

$$\tilde{\alpha}_{jk} = \alpha\left(\frac{\partial}{\partial y^j}, \frac{\partial}{\partial y^k}\right) = \sum_{i=1}^n\sum_{\ell=1}^n \frac{\partial x^i}{\partial y^j}\,\frac{\partial x^\ell}{\partial y^k}\,\alpha\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^\ell}\right) = \sum_{i,\ell=1}^n \frac{\partial x^i}{\partial y^j}\,\frac{\partial x^\ell}{\partial y^k}\,\alpha_{i\ell}.$$

☺

Suppose instead of starting with a tensor field of order $(2,0)$, we started with the coefficients α_{jk} in coordinate charts $(\mathbf{x}, U)$, such that whenever two charts overlapped, the coefficients in the charts were related by (15.3.1). We could then build the tensor field by setting $\alpha_I = \xi_I\alpha(X,Y)$ on each coordinate chart $(\mathbf{x}_I, U_I)$, where $\{\xi_I\}$ is a partition of unity with $\operatorname{supp}\xi_I \subset U_I$. It makes sense to compute

$\alpha(X,Y)$ in each coordinate chart since we have the coefficients α_{jk}, and since we are multiplying by ξ_I, we only ever need to compute in the chart U_I. Then we could define α by $\alpha = \sum_I \alpha_I$, obtaining a globally-defined smooth operator from $\chi(M) \otimes \chi(M) \to \mathcal{F}(M)$. Of course, if we could do this, then there would probably be an invariant way to define the tensor field as a tensorial operator, although it may be less obvious.

15.4. **The differential of a 1-form (in coordinates).** Now let's return to our motivating problem: differentiating a tensor field of type $(1,0)$ to obtain a tensor field of type $(2,0)$. What we have in mind, for a 1-form $\omega = \sum_{j=1}^n \omega_j(\mathbf{x})\, dx^j$, is something like a $(2,0)$ tensor

$$D\omega = \sum_{i=1}^n \sum_{j=1}^n \frac{\partial \omega_j}{\partial x^i}\, dx^i \otimes dx^j.$$

The question is whether this is coordinate-independent or not. (Notice that, unlike every other object we have defined so far, we are trying to define the operation by its coordinate components. The reason is that the invariant definition is far less obvious. Because we want to start in coordinates, we need to check explicitly that the object so obtained does not depend on coordinates. If so, then it represents something genuine.) Unfortunately this definition of $D\omega$ doesn't quite work. This will be our first negative result, but hey, if every result were positive, anyone could do mathematics. Fortunately, the proof will suggest the right definition.

Proposition 15.4.1. *Suppose $D\omega$ is defined for 1-forms ω in terms of coordinates by*

$$D\omega = \sum_{i=1}^n \sum_{j=1}^n \frac{\partial \omega_j}{\partial x^i}\, dx^i \otimes dx^j.$$

Then $D\omega$ cannot be a tensor of type $(2,0)$; in other words, it is not *true that given two coordinate charts $(\mathbf{x}, U)$ and $(\mathbf{y}, V)$ for which $\omega = \sum_{j=1}^n \omega_j(\mathbf{x})\, dx^j = \sum_{l=1}^n \widetilde{\omega}_\ell(\mathbf{y})\, dy^\ell$, we have*

$$\sum_{i=1}^n \sum_{j=1}^n \frac{\partial \omega_j}{\partial x^i}\, dx^i \otimes dx^j = \sum_{k=1}^n \sum_{\ell=1}^n \frac{\partial \widetilde{\omega}_\ell}{\partial y^k}\, dy^k \otimes dy^\ell.$$

Proof. We first use the transition formula (15.1.3) to get $\omega_j(\mathbf{x}) = \sum_{\ell=1}^n \widetilde{\omega}_\ell \circ (\mathbf{y}) \, \frac{\partial y^l}{\partial x^j}$. Then we simply plug in and compute.

$$\begin{aligned}
D\omega &= \sum_{i=1}^n \sum_{j=1}^n \frac{\partial}{\partial x^i}(\omega_j(\mathbf{x}))\, dx^i \otimes dx^j \\
&= \sum_{i=1}^n \sum_{j=1}^n \sum_{\ell=1}^n \frac{\partial}{\partial x^i}\left(\widetilde{\omega}_\ell \circ (\mathbf{y})\, \frac{\partial y^\ell}{\partial x^j}\right) dx^i \otimes dx^j \\
&= \sum_{i=1}^n \sum_{j=1}^n \sum_{\ell=1}^n \left(\frac{\partial y^\ell}{\partial x^j}\, \frac{\partial}{\partial x^i}\big(\widetilde{\omega}_\ell(\mathbf{y})\big) + \widetilde{\omega}_\ell(\mathbf{y})\, \frac{\partial^2 y^\ell}{\partial x^i \partial x^j}\right) dx^i \otimes dx^j \\
&= \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \sum_{\ell=1}^n \frac{\partial y^\ell}{\partial x^j}\, \frac{\partial y^k}{\partial x^i}\, \frac{\partial \widetilde{\omega}_\ell}{\partial y^k}\, dx^i \otimes dx^j + \sum_{i=1}^n \sum_{j=1}^n \sum_{\ell=1}^n \widetilde{\omega}_\ell(\mathbf{y})\, \frac{\partial^2 y^\ell}{\partial x^i \partial x^j}\, dx^i \otimes dx^j \\
&= \sum_{k=1}^n \sum_{\ell=1}^n \frac{\partial \widetilde{\omega}_\ell}{\partial y^k}\, dy^k \otimes dy^\ell + \sum_{i=1}^n \sum_{j=1}^n \sum_{l=1}^n \widetilde{\omega}_\ell(\mathbf{y})\, \frac{\partial^2 y^\ell}{\partial x^i \partial x^j}\, dx^i \otimes dx^j,
\end{aligned}$$

so that we finally obtain the transition formula

$$D\omega = \tilde{D}\widetilde{\omega} + \sum_{i=1}^n \sum_{j=1}^n \sum_{\ell=1}^n \widetilde{\omega}_\ell(\mathbf{y})\, \frac{\partial^2 y^\ell}{\partial x^i \partial x^j}\, dx^i \otimes dx^j. \tag{15.4.1}$$

The first term in (15.4.1) is what we'd want, but the second term is generally not zero. Thus $D\omega$ *is* coordinate-dependent. □

You should notice that we do *almost* get something invariant from this. The error term is a matrix like

$$R_{ij} = \sum_{\ell=1}^n \widetilde{\omega}_\ell(\mathbf{y})\, \frac{\partial^2 y^\ell}{\partial x^i \partial x^j},$$

and what's interesting is that this is a *symmetric* matrix. If we switch i and j, the only effect is to switch the order of a partial differentiation of a smooth function, and by Theorem 5.1.9 on mixed partials, this does not change the quantity.

We will find this happening fairly often: if we try to make a tensor out of something that is not a tensor, we end up with second partial derivatives, and these are always symmetric. Thus our error terms generally end up being symmetric. If we therefore take the *antisymmetric* part of such quantities, the error term disappears. It is for this reason that almost every tensor in differential geometry is defined in terms of antisymmetric operators. Let's see how this works.

Definition 15.4.2. If ω is a 1-form on $M \cong \mathbb{R}^n$, then we define an object $d\omega$ by the formula

$$\begin{aligned}
d\omega &= \sum_{i=1}^n \sum_{j=1}^n \left(\frac{\partial \omega_j}{\partial x^i} - \frac{\partial \omega_i}{\partial x^j}\right) dx^i \otimes dx^j \\
&= \sum_{i=1}^n \sum_{j=1}^n \frac{\partial \omega_j}{\partial x^i}(dx^i \otimes dx^j - dx^j \otimes dx^i) \qquad (15.4.2) \\
&= \sum_{i=1}^n \frac{\partial \omega_j}{\partial x^i}\, dx^i \wedge dx^j.
\end{aligned}$$

(Recall the definition of wedge product on forms from Definition 4.3.3.) The two formulas are equivalent since the second is obtained from the first by interchanging dummy indices in the second summand.

More generally if M is a smooth n-manifold, then we define $d\omega$ in each coordinate chart and use a partition of unity on M to assemble $d\omega$ globally, as discussed after Example 15.3.3.

Now we need to check that this actually is a tensor of type $(2,0)$, i.e., that it does not depend on the choice of coordinates.

Proposition 15.4.3. *If ω is a 1-form and $d\omega$ is defined in any coordinate system as in* (15.4.2)*, then $d\omega$ is independent of coordinates: if $(\mathbf{x}, U)$ and $(\mathbf{y}, V)$ are two coordinate charts with $\omega = \sum_{j=1}^n \omega_j \, dx^j = \sum_{\ell=1}^n \widetilde{\omega}_\ell \, dy^\ell$, then we have*

$$\sum_{i=1}^n \sum_{j=1}^n \left(\frac{\partial \omega_j}{\partial x^i} - \frac{\partial \omega_i}{\partial x^j} \right) dx^i \otimes dx^j = \sum_{k=1}^n \sum_{\ell=1}^n \left(\frac{\partial \widetilde{\omega}_\ell}{\partial y^k} - \frac{\partial \widetilde{\omega}_k}{\partial y^\ell} \right) dy^k \otimes dy^\ell.$$

Proof. The proof is exactly the same as that of Proposition 15.4.1, except the antisymmetry cancels out the error term this time. □

Let's work out something explicit.

Example 15.4.4. Consider $M \cong \mathbb{R}^3$, with a 1-form $\omega = u(x,y,z)\,dx + v(x,y,z)\,dy + w(x,y,z)\,dz$. Then

$$\begin{aligned} d\omega &= \sum_{i=1}^3 \sum_{j=1}^3 \frac{\partial \omega_j}{\partial x^i} \, dx^i \wedge dx^j \\ &= \frac{\partial u}{\partial x} \, dx \wedge dx + \frac{\partial u}{\partial y} \, dy \wedge dx + \frac{\partial u}{\partial z} \, dz \wedge dx + \frac{\partial v}{\partial x} \, dx \wedge dy + \frac{\partial v}{\partial y} \, dy \wedge dy \\ &\quad + \frac{\partial v}{\partial z} \, dz \wedge dy + \frac{\partial w}{\partial x} \, dx \wedge dz + \frac{\partial w}{\partial y} \, dy \wedge dz + \frac{\partial w}{\partial z} \, dz \wedge dz, \end{aligned}$$

which simplifies using antisymmetry to

$$d\omega = \left(\frac{\partial w}{\partial y} - \frac{\partial v}{\partial z} \right) dy \wedge dz + \left(\frac{\partial u}{\partial z} - \frac{\partial w}{\partial x} \right) dz \wedge dx + \left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \right) dx \wedge dy. \tag{15.4.3}$$

Note the similarity to the formula for the curl of a vector field. We will make this more explicit in Chapter 19.

Also observe that if $M \cong \mathbb{R}^2$, with a 1-form $\omega = u(x,y)\,dx + v(x,y)\,dy$, then

$$d\omega = \sum_{i=1}^2 \sum_{j=1}^2 \frac{\partial \omega_j}{\partial x^i} \, dx^i \wedge dx^j = \left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \right) dx \wedge dy.$$

Again this is similar to the curl of a two-dimensional vector field (which can be thought of as a function).

☺

The fact that the derivative of a tensor field of order $(1,0)$ ends up being an antisymmetric tensor field of order $(2,0)$ motivates us to consider antisymmetric tensor fields of any order and their derivatives, which we will do in the next Chapter. We will also obtain an invariant definition of the d operator without using coordinates.

16. The d operator

"This is where the fun begins."

16.1. **The differential of a 1-form (invariant).** We saw in the last section that if $\alpha = \sum_{i=1}^n \alpha_i\,dx^i$ is a 1-form, we can define a 2-form $d\alpha$ by the coordinate formula

$$d\alpha = \sum_{i,j=1}^n \frac{\partial \alpha_j}{\partial x^i}\,dx^i \wedge dx^j.$$

It happens that, by Proposition 15.4.3, this formula gives the same 2-form regardless of choice of coordinates. Our goal now is to find some genuinely invariant definition of $d\alpha$, without using coordinates.

Our first clue is obtained by computing

$$\begin{aligned} d\alpha\left(\frac{\partial}{\partial x^k}, \frac{\partial}{\partial x^\ell}\right) &= \sum_{i=1}^n \sum_{j=1}^n \frac{\partial \alpha_j}{\partial x^i}\,(dx^i \wedge dx^j)\left(\frac{\partial}{\partial x^k}, \frac{\partial}{\partial x^\ell}\right) \\ &= \sum_{i=1}^n \sum_{j=1}^n \frac{\partial \alpha_j}{\partial x^i}(\delta^i_k \delta^j_\ell - \delta^i_\ell \delta^j_k) \\ &= \frac{\partial \alpha_\ell}{\partial x^k} - \frac{\partial \alpha_k}{\partial x^\ell}. \end{aligned}$$

Now this formula tells us[32]

$$d\alpha(X,Y) = X\big(\alpha(Y)\big) - Y\big(\alpha(X)\big) \qquad \text{if } X = \frac{\partial}{\partial x^k} \text{ and } Y = \frac{\partial}{\partial x^\ell}, \tag{16.1.1}$$

since $\alpha\left(\frac{\partial}{\partial x^k}\right) = \alpha_k$. We might be tempted to use formula (16.1.1) to *define* $d\alpha$ in an invariant way, but this formula can't possibly be right. The left side should be tensorial in both X and Y separately (i.e., we can pull out C^∞ functions), while the right-hand side is not:

$$\begin{aligned} d\alpha(fX,Y) &= fX\big(\alpha(Y)\big) - Y\big(\alpha(fX)\big) \\ &= fX\big(\alpha(Y)\big) - Y\big(f\alpha(X)\big) \\ &= fX\big(\alpha(Y)\big) - \alpha(X)Y(f) - fY\big(\alpha(X)\big) \\ &= f d\alpha(X,Y) - Y(f)\alpha(X) \\ &\neq f d\alpha(X,Y), \quad \text{which can't be correct.} \end{aligned}$$

As a result, equation (16.1.1) is wrong in general. There is a missing term which happens to be zero when X and Y are both coordinate fields.

The clue is that we already know that if X and Y are vector fields on M, then so is the Lie bracket $[X,Y]$, by Proposition 14.2.1. However if X and Y are coordinate basis vector fields, with $X = \frac{\partial}{\partial x^k}$ and $Y = \frac{\partial}{\partial x^\ell}$, then for any function f we have

$$[X,Y](f) = \frac{\partial}{\partial x^k}\frac{\partial f}{\partial x^\ell} - \frac{\partial}{\partial x^\ell}\frac{\partial f}{\partial x^k} = 0$$

[32]In this formula and in the future, we will always use $X(f)$ to denote $df(X)$, i.e., this *never* means multiplication of the function f by the vector field X, but rather means the new function obtained by applying X to f as a differential operator.

since mixed partials commute by Theorem 5.1.9. This suggests that the missing term involves a Lie bracket of vector fields. Now at last we can define the coordinate-free version of $d\omega$ for a 1-form ω.

Definition 16.1.1. If ω is a 1-form on an n-manifold M, then we define the 2-form $d\omega$ on M by the formula[33]

$$d\omega(X,Y) = X\big(\omega(Y)\big) - Y\big(\omega(X)\big) - \omega([X,Y]), \tag{16.1.2}$$

for any two vector fields X and Y on M.

We need to check that this does actually define a 2-form, i.e., that it's antisymmetric and tensorial.

Proposition 16.1.2. *The 2-form $d\omega$ defined by* (16.1.2) *is an antisymmetric tensor of type $(2,0)$, i.e., we have*

$$\begin{aligned} d\omega(X,Y) &= -d\omega(Y,X) \\ d\omega(fX,Y) &= f d\omega(X,Y), \\ d\omega(X,fY) &= f d\omega(X,Y) \end{aligned}$$

for any vector fields X and Y, and any smooth function f.

Proof. Antisymmetry is obvious, since $[X,Y] = -[Y,X]$.

For tensoriality, we need to simplify $d\omega(fX,Y)$ for any function f. First, for any function h, we have by Definition 14.2.1 of the Lie bracket that

$$\begin{aligned} [fX,Y](h) &= fX\big(Y(h)\big) - Y\big(fX(h)\big) \\ &= fX\big(Y(h)\big) - Y(f)X(h) - fY\big(fX(h)\big) \\ &= f[X,Y](h) - Y(f)X(h), \end{aligned}$$

so that

$$[fX,Y] = f[X,Y] - Y(f)X. \tag{16.1.3}$$

As a result, then, we get

$$\begin{aligned} d\omega(fX,Y) &= fX\big(\omega(Y)\big) - Y\big(\omega(fX)\big) - \omega([fX,Y]) \\ &= fX\big(\omega(Y)\big) - Y\big(f\omega(X)\big) - \omega\big(f[X,Y] - Y(f)X\big) \\ &= fX\big(\omega(Y)\big) - Y(f)\omega(X) - fY\big(\omega(X)\big) - f\omega([X,Y]) + Y(f)\omega(X) \\ &= fX\big(\omega(Y)\big) - fY\big(\omega(X)\big) - f\omega([X,Y]) \\ &= f d\omega(X,Y). \end{aligned}$$

For the last equation, we have $d\omega(X,fY) = -d\omega(fY,X) = -f d\omega(Y,X) = f d\omega(X,Y)$. □

Observe what happened in the proof above: the incorrect formula (16.1.1) is *not* tensorial in X, and the Lie bracket is also *not* tensorial in X by formula (16.1.3). However, when we combine them in the right way as in (16.1.2), we get something that *is* tensorial in X. This happens quite often, usually in the same way. We need antisymmetric combinations of differential operators (because of difficulties like in Proposition 15.4.1), and then we need antisymmetric derivatives (Lie brackets) to cancel out terms, so that we end up with something tensorial.

[33] Again note that we are using X to differentiate the function $\omega(Y)$ and using Y to differentiate the function $\omega(X)$.

Example 16.1.3. Suppose $\omega = h(x,y)\,dx + j(x,y)\,dy$ on $\mathbb{R}^2$. Then $d\omega$ is a 2-form on $\mathbb{R}^2$, which means it is completely determined by what it does to any basis of vector fields.

So let $X = \frac{\partial}{\partial x}$ and $Y = \frac{\partial}{\partial y}$. Then $[X,Y] = 0$ since mixed partials commute, and furthermore we have

$$\begin{aligned} d\omega(X,Y) &= X\big(\omega(Y)\big) - Y\big(\omega(X)\big) - \omega([X,Y]) \\ &= X\big(j(x,y)\big) - Y\big(h(x,y)\big) \\ &= \frac{\partial j}{\partial x}(x,y) - \frac{\partial h}{\partial y}(x,y). \end{aligned}$$

By tensoriality, this completely determines $d\omega$, and we have

$$d\omega_p = \left(\frac{\partial j}{\partial x}(x,y) - \frac{\partial h}{\partial y}(x,y)\right)\,(dx|_p \wedge dy|_p)$$

where (x,y) are the coordinates of p.

Note that $d\omega = 0$ implies that $\omega = df$ for some function f by Proposition 15.2.10, and conversely that if $\omega = df$ then $d\omega = 0$. ☺

16.2. **The differential of a k-form.** It's pretty easy to generalize the d operator, which we've now defined for 0-forms (functions) by Proposition 15.2.7 and for 1-forms by Definition 16.1.1, to any k-form. We just have to differentiate antisymmetrically, then subtract off Lie brackets to get tensoriality.

Example 16.2.1. Suppose ω is a 2-form. Then $d\omega$ is a 3-form, and the first term in $d\omega(X,Y,Z)$ should be $X\big(\omega(Y,Z)\big)$. Once we know this, all the rest of the terms are determined. For example, this expression is already antisymmetric in Y and Z (since ω is), but to get antisymmetry between X and Y, and between X and Z, we need to have the cyclically permuted terms

$$X\big(\omega(Y,Z)\big) + Y\big(\omega(Z,X)\big) + Z\big(\omega(X,Y)\big). \tag{16.2.1}$$

Actually if you want, you can interpret this as the antisymmetrization of the $(3,0)$ tensor $D\otimes\omega$ as in Proposition 4.3.5, where we write $D\otimes\omega(X,Y,Z) = X\big(\omega(Y,Z)\big)$. This makes it look somewhat more like the curl operation in vector calculus, where we define the cross product $\mathbf{U}\times\mathbf{V}$ and then use the same formula $\nabla\times\mathbf{V}$ to define the curl as though the differential operator ∇ were another vector field. The terms (16.2.1) are then exactly what you'd get when you take the alternation of the tensor product of a 1-form with a 2-form, as we did just before Example 4.3.4. I'm not sure it's worth taking this too seriously, but it might help with intuition.

Of course although (16.2.1) is antisymmetric, it is not tensorial. This didn't matter when we were working with forms in a single tangent space as in Chapter 4, but now it has to be imposed. If we define $D\omega(X,Y,Z)$ to be (16.2.1), then we have

$$D\omega(fX,Y,Z) - fD\omega(X,Y,Z) = Y(f)\,\omega(Z,X) + Z(f)\,\omega(X,Y).$$

To obtain tensoriality, we recall the formula (16.1.3), which shows that if

$$\begin{aligned} d\omega(X,Y,Z) = {} & X\big(\omega(Y,Z)\big) + Y\big(\omega(Z,X)\big) + Z\big(\omega(X,Y)\big) \\ & - \omega([X,Y],Z) - \omega([Y,Z],X) - \omega([Z,X],Y), \end{aligned} \tag{16.2.2}$$

then

$$d\omega(fX,Y,Z) - f d\omega(X,Y,Z) = Y(f)\,\omega(Z,X) + Z(f)\,\omega(X,Y) \\ + Y(f)\,\omega(X,Z) - Z(f)\,\omega(X,Y) = 0,$$

as desired. And by antisymmetry, since $d\omega$ is tensorial in the first component, it is also tensorial in the second and third components. ☺

Let's now do the general case.

Definition 16.2.2. Suppose M is an n-dimensional manifold and ω is a k-form on M (with $k \leq n$). Then we define the $(k+1)$-form $d\omega$ to be the tensor of type $(k+1,0)$ satisfying, for any $k+1$ vector fields $X_1, X_2, \cdots, X_{k+1}$, the formula

$$d\omega(X_1,\cdots,X_{k+1}) = \sum_{j=1}^{k+1} (-1)^{j+1} X_j\big(\omega(X_1,\cdots,X_{j-1},X_{j+1},\cdots,X_{k+1})\big) \\ + \sum_{1\leq i<j\leq k+1} (-1)^{i+j} \omega\big([X_i,X_j],X_1,\cdots,X_{i-1},X_{i+1},\cdots,X_{j-1},X_{j+1},\cdots,X_{k+1}\big). \tag{16.2.3}$$

This somewhat strange formula is the way it is in order to actually get a $(k+1)$-form: the powers of (-1) ensure antisymmetry, while the second summand's Lie brackets ensure that $d\omega$ is a tensor. Let's verify these things.

Proposition 16.2.3. *If ω is a k-form, then $d\omega$ is a $(k+1)$-form. Specifically, for any $k+1$ vector fields X_1 through X_{k+1}, we have*

$$d\omega(X_1,\cdots,X_p,\cdots,X_q,\cdots,X_{k+1}) = -d\omega(X_1,\cdots,X_q,\cdots,X_p,\cdots,X_{k+1})$$

for every pair of indices p and q; furthermore, for any function f, we have

$$d\omega(X_1,\cdots,fX_p,\cdots,X_{k+1}) = f d\omega(X_1,\cdots,X_p,\cdots,X_{k+1})$$

for any index p.

Proof. This is one of those things where the basic ideas are pretty clear, but the details are a nightmare. Thus most authors write, "The proof is left to the reader." But you have to do something like this at least once, so we might as well do it here.

First let us verify antisymmetry. To make this a little easier and avoid getting lost in the indices, let's assume the two vector fields being interchanged are adjacent. This is no restriction, since any transposition permutation can be decomposed into transpositions of adjacent elements.[34] Furthermore, every transposition is composed of an *odd* number of adjacent transpositions. (Hence if one adjacent transposition reverses the sign, then so will every transposition.) So let us then prove that

$$d\omega(X_1,\cdots,X_p,X_{p+1},\cdots,X_{k+1}) = -d\omega(X_1,\cdots,X_{p+1},X_p,\cdots,X_{k+1}).$$

A nice trick for doing this is to show that $d\omega(X_1,\cdots,X_p,X_p,\cdots,X_{k+1}) = 0$ for any k vector fields; then we automatically have antisymmetry since if we knew

[34]For example, the transposition $(123) \mapsto (321)$ is composed of the elements $(123) \mapsto (213) \mapsto (231) \mapsto (321)$. The general result is one of the first theorems of discrete group theory.

this, we would also know that

$$\begin{aligned}0 &= d\omega(\cdots, X_p + X_q, X_p + X_q, \cdots)\\ &= d\omega(\cdots, X_p, X_p, \cdots) + d\omega(\cdots, X_p, X_q, \cdots)\\ &\qquad + d\omega(\cdots, X_q, X_p, \cdots) + d\omega(\cdots, X_q, X_q, \cdots)\\ &= d\omega(\cdots, X_p, X_q, \cdots) + d\omega(\cdots, X_q, X_p, \cdots).\end{aligned}$$

Thus we write

$$d\omega(X_1, \cdots, X_{k+1}) = \sum_{1\le j\le k+1} A_j + \sum_{1\le i<j\le k+1} B_{ij} \tag{16.2.4}$$

where

$$A_j = (-1)^{j+1} X_j\big(\omega(X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_{k+1})\big) \tag{16.2.5}$$

$$B_{ij} = (-1)^{i+j}\omega\big([X_i, X_j], X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_{j-1}, X_{j+1}, \cdots, X_{k+1}\big), \tag{16.2.6}$$

and analyze each of the terms A_j and B_{ij} supposing that $X_p = X_{p+1}$. There are several possibilities. For A_j, all terms except for when $j = p$ and $j = p+1$ must automatically vanish by antisymmetry of ω, since $X_p = X_{p+1}$ appears in two different arguments. Thus in the first sum of (16.2.4) we are left with only

$$\begin{aligned}\textstyle\sum_{j=1}^{k+1} A_j &= A_p + A_{p+1}\\ &= (-1)^{p+1} X_p\big(\omega(\cdots, X_{p-1}, X_{p+1}, X_{p+2}, \cdots)\big)\\ &\qquad + (-1)^{p+2} X_{p+1}\big(\omega(\cdots, X_{p-1}, X_p, X_{p+2}, \cdots)\big) = 0\end{aligned}$$

since $X_p = X_{p+1}$ and the signs $(-1)^{p+1}$ and $(-1)^{p+2}$ are always opposite.

Now we have to do the same thing with the second sum $\sum_{i<j} B_{ij}$ in (16.2.4). Again, if X_p and X_{p+1} both show up in any particular summand, then since ω is antisymmetric, that summand must be zero. So the only terms to worry about are those with $i = p$, $i = p+1$, $j = p$, or $j = p+1$. We have

$$\begin{aligned}\sum_{1\le i<j\le n} B_{ij} &= \sum_{j>p}^{n} B_{pj} + \sum_{j>p+1} B_{p+1,j} + \sum_{i<p} B_{ip} + \sum_{i<p+1} B_{i,p+1}\\ &= \sum_{j>p+1} (B_{pj} + B_{p+1,j}) + \sum_{i<p} (B_{ip} + B_{i,p+1}) + 2B_{p,p+1}.\end{aligned} \tag{16.2.7}$$

Now it's easy to see that $B_{pj} + B_{p+1,j} = 0$ if $j > p+1$ since

$$\begin{aligned}B_{pj} + B_{p+1,j} &= (-1)^{p+j}\omega([X_p, X_j], \cdots, X_{p-1}, X_{p+1}, X_{p+2}, \cdots, X_{j-1}, X_{j+1}, \cdots)\\ &\quad + (-1)^{p+j+1}\omega([X_{p+1}, X_j], \cdots, X_{p-1}, X_p, X_{p+2}, \cdots, X_{j-1}, X_{j+1}, \cdots),\end{aligned}$$

and this vanishes since $X_p = X_{p+1}$ and the terms have opposite signs. Similarly we see that $B_{ip} + B_{i,p+1} = 0$ if $i < p$. Finally we have

$$B_{p,p+1} = (-1)^{2p+1}\omega([X_p, X_{p+1}], X_1, \cdots, X_{p-1}, X_{p+2}, \cdots, X_k) = 0$$

by antisymmetry of the Lie bracket. Hence all terms of the sum (16.2.7) are zero, and thus we conclude by formula (16.2.4) that $d\omega(\cdots, X_p, X_{p+1}, \cdots) = 0$ if $X_p = X_{p+1}$. Antisymmetry of $d\omega$ is a consequence.

The next step is to prove linearity over the functions, so we want to compute $d\omega(X_1, \cdots, fX_p, \cdots, X_{k+1})$. Since we have already proved antisymmetry, it's enough to prove tensoriality in the first component: that is, we want to show

$$d\omega(fX_1, X_2, \cdots, X_{k+1}) = f\, d\omega(X_1, X_2, \cdots, X_{k+1}). \tag{16.2.8}$$

Write $\Omega = d\omega(fX_1, X_2, \cdots, X_{k+1}) - f\, d\omega(X_1, X_2, \cdots, X_{k+1})$. Set $\tilde{A}_j$ and $\tilde{B}_{ij}$ to be the terms in (16.2.5) and (16.2.6) with X_1 replaced by fX_1; then we have

$$\Omega = \sum_{j=1}^{k+1} \tilde{A}_j - fA_j + \sum_{1 \le i < j \le k+1} \tilde{B}_{ij} - fB_{ij}.$$

Now many of the terms in (16.2.4) are already tensorial. For example the only nontensorial part of B_{ij} is the Lie bracket, and so if $i \neq 1$ and $j > i$, then we immediately have $\tilde{B}_{ij} = fB_{ij}$. On the other hand, A_j is tensorial only in X_j component, which means we have $\tilde{A}_1 - fA_1 = 0$ automatically.

We have thus simplified to

$$\Omega = \sum_{j>1} \tilde{A}_j - fA_j + \sum_{j>1} \tilde{B}_{1j} - fB_{1j},$$

and we need to show that this is zero. Using the definition of A_j from (16.2.5) and tensoriality of ω, we have

$$\begin{aligned} \tilde{A}_j - fA_j &= (-1)^{j+1} \big[X_j\big(\omega(fX_1, \cdots, X_{j-1}, X_{j+1}, \cdots)\big) \\ &\qquad - fX_j\big(\omega(X_1, \cdots, X_{j-1}, X_{j+1}, \cdots)\big)\big] \\ &= (-1)^{j+1} X_j(f)\, \omega(X_1, \cdots, X_{j-1}, X_{j+1}, \cdots). \end{aligned}$$

Furthermore using the formula (16.1.3), we have

$$\begin{aligned} \tilde{B}_{1j} - fB_{1j} &= (-1)^{1+j} \big[\omega([fX_1, X_j], X_2, \cdots, X_{j-1}, X_{j+1}, \cdots) \\ &\qquad - f\, \omega([X_1, X_j], X_2, \cdots, X_{j-1}, X_{j+1}, \cdots)\big] \\ &= -(-1)^{1+j} X_j(f)\, \omega(X_1, X_2, \cdots, X_{j-1}, X_{j+1}, \cdots). \end{aligned}$$

Hence we conclude $\tilde{A}_j - fA_j + \tilde{B}_{1j} - fB_{1j} = 0$ for all $j > 1$, which implies $\Omega = 0$ and thus the tensoriality (16.2.8) as desired. □

16.3. **The differential in coordinates and its properties.** Despite the awkward definition and proof above of tensoriality, d is actually a very simple operator in terms of its properties. If we compute it in a coordinate basis, the formula is quite easy, and all the messy powers of (-1) disappear. In addition, if the vector fields X_j are coordinate vector fields, so that $X_j = \frac{\partial}{\partial x^{m_j}}$ for some $m_j \in \{1, \cdots, n\}$, then we will have $[X_i, X_j] = 0$ for all i and j. We have already seen what happens in simple cases: when ω is a 0-form (a function), we have by Proposition 15.1.3 that

$$d\omega = \sum_{j=1}^{n} \frac{\partial \omega}{\partial x^j}\, dx^j.$$

When $\omega = \sum_{i=1}^n \omega_i\, dx^i$ is a 1-form, we have by Definition 15.4.2 that

$$d\omega = \sum_{j=1}^{n} \frac{\partial \omega_i}{\partial x^j}\, dx^j \wedge dx^i.$$

Using the formula from Example 16.2.1, we can check that the same pattern holds.

Example 16.3.1. Consider for concreteness the 2-form ω on $\mathbb{R}^4$ given by

$$\omega = f(x^1, x^2, x^3, x^4)\, dx^2 \wedge dx^4.$$

Then $d\omega$ is a 3-form on $\mathbb{R}^4$, which means it is completely determined if we know

$$\omega_{123} = d\omega(\tfrac{\partial}{\partial x^1}, \tfrac{\partial}{\partial x^2}, \tfrac{\partial}{\partial x^3}), \quad \omega_{124} = d\omega(\tfrac{\partial}{\partial x^1}, \tfrac{\partial}{\partial x^2}, \tfrac{\partial}{\partial x^4}),$$
$$\omega_{134} = d\omega(\tfrac{\partial}{\partial x^1}, \tfrac{\partial}{\partial x^3}, \tfrac{\partial}{\partial x^4}), \quad \omega_{234} = d\omega(\tfrac{\partial}{\partial x^2}, \tfrac{\partial}{\partial x^3}, \tfrac{\partial}{\partial x^4}).$$

We thus need to use (16.2.2) four times, with our vector fields X, Y, and Z all equal to coordinate basis vector fields. Hence since mixed partials commute, we know $[X, Y] = [Y, Z] = [Z, X] = 0$.

We compute:

$$\omega_{123} = \tfrac{\partial}{\partial x^1}\big(\omega(\tfrac{\partial}{\partial x^2}, \tfrac{\partial}{\partial x^3})\big) + \tfrac{\partial}{\partial x^2}\big(\omega(\tfrac{\partial}{\partial x^3}, \tfrac{\partial}{\partial x^1})\big) + \tfrac{\partial}{\partial x^3}\big(\omega(\tfrac{\partial}{\partial x^1}, \tfrac{\partial}{\partial x^2})\big) = \tfrac{\partial}{\partial x^1}(0) = 0$$
$$\omega_{124} = \tfrac{\partial}{\partial x^1}\big(\omega(\tfrac{\partial}{\partial x^2}, \tfrac{\partial}{\partial x^4})\big) + \tfrac{\partial}{\partial x^2}\big(\omega(\tfrac{\partial}{\partial x^4}, \tfrac{\partial}{\partial x^1})\big) + \tfrac{\partial}{\partial x^4}\big(\omega(\tfrac{\partial}{\partial x^1}, \tfrac{\partial}{\partial x^2})\big) = \tfrac{\partial}{\partial x^1} f(x^1, x^2, x^3, x^4)$$
$$\omega_{134} = \tfrac{\partial}{\partial x^1}\big(\omega(\tfrac{\partial}{\partial x^3}, \tfrac{\partial}{\partial x^4})\big) + \tfrac{\partial}{\partial x^3}\big(\omega(\tfrac{\partial}{\partial x^4}, \tfrac{\partial}{\partial x^1})\big) + \tfrac{\partial}{\partial x^4}\big(\omega(\tfrac{\partial}{\partial x^1}, \tfrac{\partial}{\partial x^3})\big) = \tfrac{\partial}{\partial x^1}(0) = 0$$
$$\omega_{234} = \tfrac{\partial}{\partial x^2}\big(\omega(\tfrac{\partial}{\partial x^3}, \tfrac{\partial}{\partial x^4})\big) + \tfrac{\partial}{\partial x^3}\big(\omega(\tfrac{\partial}{\partial x^4}, \tfrac{\partial}{\partial x^2})\big) + \tfrac{\partial}{\partial x^4}\big(\omega(\tfrac{\partial}{\partial x^2}, \tfrac{\partial}{\partial x^3})\big) = -\tfrac{\partial}{\partial x^3} f(x^1, x^2, x^3, x^4).$$

We therefore conclude that

$$\begin{aligned} d\omega &= \omega_{123}\, dx^1 \wedge dx^2 \wedge dx^3 + \omega_{124}\, dx^1 \wedge dx^2 \wedge dx^4 \\ &\qquad + \omega_{134}\, dx^1 \wedge dx^3 \wedge dx^4 + \omega_{234}\, dx^2 \wedge dx^3 \wedge dx^4 \\ &= \frac{\partial f}{\partial x^1}\, dx^1 \wedge dx^2 \wedge dx^4 - \frac{\partial f}{\partial x^3}\, dx^2 \wedge dx^3 \wedge dx^4 \\ &= \left(\frac{\partial f}{\partial x^1}\, dx^1 + \frac{\partial f}{\partial x^3}\, dx^3\right) \wedge dx^2 \wedge dx^4 \\ &= \left(\frac{\partial f}{\partial x^1}\, dx^1 + \frac{\partial f}{\partial x^2}\, dx^2 + \frac{\partial f}{\partial x^3}\, dx^3 + \frac{\partial f}{\partial x^4}\, dx^4\right) \wedge dx^2 \wedge dx^4 \\ &= df \wedge dx^2 \wedge dx^3. \end{aligned}$$

Here of course we used antisymmetry of the wedge product in order to insert the terms $\frac{\partial f}{\partial x^2}\, dx^2$ and $\frac{\partial f}{\partial x^4}\, dx^4$ without changing the 3-form.

We have thus demonstrated the formula

$$d\big(f\, dx^2 \wedge dx^4\big) = df \wedge dx^2 \wedge dx^4,$$

and this formula generalizes to any k-form in any number of dimensions. ☺

Proposition 16.3.2. *Suppose $(\mathbf{x}, U)$ is a coordinate chart on M, and suppose ω is a k-form written in coordinates as*

$$\omega = \sum_{1 \le i_1 < \cdots < i_k \le n} \omega_{i_1 \cdots i_k}\, dx^{i_1} \wedge \cdots \wedge dx^{i_k}.$$

Then $d\omega$, computed from (16.2.3)*, is*

$$d\omega = \sum_{1 \le i_1 < \cdots < i_k \le n} \sum_{j=1}^{n} \frac{\partial \omega_{i_1 \cdots i_k}}{\partial x^j}\, dx^j \wedge dx^{i_1} \wedge \cdots \wedge dx^{i_k}. \tag{16.3.1}$$

Proof. By linearity of the d operator, it is enough to prove this in the case where

$$\omega = f\, dx^{i_1} \wedge \cdots \wedge dx^{i_k} \tag{16.3.2}$$

for some function f (which allows us to avoid rewriting the sum over all possible indices repeatedly). We can of course assume that $1 \leq i_1 < \cdots < i_k \leq n$. Then we want to prove that

$$d\left(f\, dx^{i_1} \wedge \cdots \wedge dx^{i_k}\right) = df \wedge dx^{i_1} \wedge \cdots \wedge dx^{i_k}. \tag{16.3.3}$$

We just need to compute both sides of (16.3.3) when applied to $k+1$ coordinate vector fields. So suppose we have indices m_1 through m_{k+1} and vector fields $X_i = \frac{\partial}{\partial x^{m_i}}$ for $1 \leq i \leq k+1$. Then we have $[X_i, X_j] = 0$ for every i and j, so the second sum in (16.2.3) is zero. Let's assume the indices are ordered so that $m_1 < m_2 < \cdots < m_{k+1}$ to keep things simple.

The rest of (16.2.3) becomes

$$\begin{aligned} d\omega(X_1, \cdots, X_{k+1}) &= \sum_{j=1}^{k+1} (-1)^{j+1} X_j\big(\omega(X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_{k+1})\big) \\ &= \sum_{j=1}^{k+1} (-1)^{j+1} \frac{\partial}{\partial x^{m_j}} \left[\omega\left(\frac{\partial}{\partial x^{m_1}}, \cdots, \frac{\partial}{\partial x^{m_{j-1}}}, \frac{\partial}{\partial x^{m_{j+1}}}, \cdots, \frac{\partial}{\partial x^{m_{k+1}}}\right)\right] \end{aligned}$$

To compute the term inside square brackets, we notice that since $\omega = f\, dx^{i_1} \wedge \cdots \wedge dx^{i_k}$, and since the indices i and m are both sorted from smallest to largest, the only way it can possibly be nonzero is if $m_1 = i_1$, …, $m_{j-1} = i_{j-1}$, $m_{j+1} = i_j$, …, $m_{k+1} = i_k$. We thus obtain

$$d\omega(X_1, \cdots, X_{k+1}) = \sum_{j=1}^{k+1} (-1)^{j+1} \delta^{i_1}_{m_1} \cdots \delta^{i_{j-1}}_{m_{j-1}} \delta^{i_j}_{m_{j+1}} \cdots \delta^{i_k}_{m_{k+1}} \frac{\partial f}{\partial x^{m_j}}. \tag{16.3.4}$$

Now let's try computing the right side of (16.3.3) applied to the same vector fields. We get

$$\sum_{j=1}^{n} \frac{\partial f}{\partial x^j}\, dx^j \wedge dx^{i_1} \wedge \cdots \wedge dx^{i_k} \left(\frac{\partial}{\partial x^{m_1}}, \cdots, \frac{\partial}{\partial x^{m_{k+1}}}\right).$$

The only way this is nonzero is if $\{j, i_1, \cdots, i_k\}$ and $\{m_1, \cdots, m_{k+1}\}$ are equal as sets. Now the i's and m's are both ordered, but j can be any number. Thus we must have $j = m_p$ for some p, and the rest of them must be $i_1 = m_1$, $i_2 = m_2$, …, $i_{p-1} = m_{p-1}$, $i_p = m_{p+1}$, …, $i_k = m_{k+1}$. We thus have

$$dx^j \wedge dx^{i_1} \wedge \cdots \wedge dx^{i_k} \left(\frac{\partial}{\partial x^{m_1}}, \cdots, \frac{\partial}{\partial x^{m_{k+1}}}\right) = (-1)^{p-1}\, \delta^{i_1}_{m_1} \cdots \delta^{i_{p-1}}_{m_{p-1}} \delta^{i_p}_{m_{p+1}} \cdots \delta^{i_k}_{m_{k+1}},$$

since we have to perform $p-1$ adjacent transpositions to put dx^j in the p^{th} spot. There are $(k+1)$ ways for this to happen, depending on which m_p happens to be equal to j. Thus the right side of (16.3.3) applied to $(X_1, \ldots, X_{k+1})$ becomes

$$\big(df \wedge dx^{i_1} \wedge \cdots \wedge dx^{i_k}\big)(X_1, \ldots, X_{k+1}) = \sum_{p=1}^{k+1} (-1)^{p-1} \frac{\partial f}{\partial x^{m_p}}\, \delta^{i_1}_{m_1} \cdots \delta^{i_{p-1}}_{m_{p-1}} \delta^{i_p}_{m_{p+1}} \cdots \delta^{i_k}_{m_{k+1}},$$

and this agrees with (16.3.4). Since the choice of vector field basis elements was arbitrary, the two $(k+1)$-forms must be equal. □

Example 16.3.3 (Forms on $\mathbb{R}^3$)**.** Suppose $M = \mathbb{R}^3$. We have already seen how to compute d on 0-forms using Proposition 15.1.3: if $\omega = f$, then

$$df = \frac{\partial f}{\partial x}\,dx + \frac{\partial f}{\partial y}\,dy + \frac{\partial f}{\partial z}\,dz,$$

which looks like the gradient. And we computed d on 1-forms in (15.4.3): if $\omega = u\,dx + v\,dy + w\,dz$, then

$$d\omega = \left(\frac{\partial w}{\partial y} - \frac{\partial v}{\partial z}\right) dy \wedge dz + \left(\frac{\partial u}{\partial z} - \frac{\partial w}{\partial x}\right) dz \wedge dx + \left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\right) dx \wedge dy,$$

which looks like the curl.

Let's see how to do this on 2-forms. First, the dimension of $\Omega^2(\mathbb{R}^3)$ is $\binom{3}{2} = 3$ at each tangent space, and it is spanned by $dy \wedge dz$, $dz \wedge dx$, and $dx \wedge dy$. So we can write an arbitrary 2-form ω as

$$\omega = p(x,y,z)\,dy \wedge dz + q(x,y,z)\,dz \wedge dx + r(x,y,z)\,dx \wedge dy,$$

and the coordinate formula (16.3.1) gives

$$d\omega = \frac{\partial p}{\partial x}\,dx \wedge dy \wedge dz + \frac{\partial q}{\partial y}\,dy \wedge dz \wedge dx + \frac{\partial r}{\partial z}\,dz \wedge dx \wedge dy.$$

Now by the antisymmetry of 1-dimensional wedge products, we have

$$dy \wedge dz \wedge dx = -dz \wedge dy \wedge dx = dz \wedge dx \wedge dy = -dx \wedge dz \wedge dy = dx \wedge dy \wedge dz.$$

Similarly $dz \wedge dx \wedge dy = dx \wedge dy \wedge dz$, so that

$$d\omega = \left(\frac{\partial p}{\partial x} + \frac{\partial q}{\partial y} + \frac{\partial r}{\partial z}\right) dx \wedge dy \wedge dz. \tag{16.3.5}$$

Note the similarity to the divergence operator (which also maps from a three-dimensional space into a one-dimensional space). ☺

We have seen that effectively, the "gradient" is d of a function, the "curl" is d of a 1-form, and the "divergence" is d of a 2-form in $\mathbb{R}^3$. (These analogies work best in Euclidean coordinates; in more general coordinate systems, the k-forms are much easier to work with than the vector fields of vector calculus.) Observe that d of any 3-form must be zero, since there are no 4-forms in $\mathbb{R}^3$. In vector calculus, we have the well-known formulas

$$\operatorname{curl}\operatorname{grad} f = 0 \qquad \text{and} \qquad \operatorname{div}\operatorname{curl} X = 0,$$

for any function f and any vector field X. Both these formulas are contained in the more general formula

$$d^2 = 0,$$

which we will prove in a moment. Applying $d^2 = 0$ to a 0-form we get $\operatorname{curl}\operatorname{grad} = 0$, and applying $d^2 = 0$ to a 1-form we get $\operatorname{div}\operatorname{curl} = 0$. (When we apply it to a 2-form, it doesn't say anything nontrivial, since as mentioned all 4-forms are automatically zero in $\mathbb{R}^3$.)

Example 16.3.4. We can easily check this in the cases computed above: for example if $u = \frac{\partial f}{\partial x}$, $v = \frac{\partial f}{\partial y}$, and $w = \frac{\partial f}{\partial z}$, then $d(u\,dx + v\,dy + w\,dz) = 0$. Similarly if $p = \frac{\partial w}{\partial y} - \frac{\partial v}{\partial z}$, $q = \frac{\partial u}{\partial z} - \frac{\partial w}{\partial x}$, and $r = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}$, then we can check that $d(p\,dy \wedge dz + q\,dz \wedge dx + r\,dx \wedge dy) = 0$. Note that *all* of these formulas work because of the commuting of mixed partial derivatives, Theorem 5.1.9. Note also that these

computations are basically the same in any coordinate chart; this is *not* the case if you try for example to prove that the curl of the gradient is zero in spherical coordinates: it's true, but not so trivial. ☺

Example 16.3.4 tells us that k-forms and the d operator are really the correct way to think of these differential operators, since they make the important fact $d^2 = 0$ easy to prove in all coordinate charts.

Proposition 16.3.5. *For any k-form ω, $d(d\omega) = 0$.*

Proof. We could prove this by using the Definition 16.2.2, but that would involve lots of sign-checking and such. It's easier to use Proposition 16.3.2 to do this in coordinates. By linearity of d, it is enough to do the computation in the special case where

$$\omega = f\, dx^{i_1} \wedge \cdots \wedge dx^{i_k}$$

for some smooth function f.

Then we have

$$d\omega = \sum_{j=1}^{n} \frac{\partial f}{\partial x^j}\, dx^j \wedge dx^{i_1} \wedge \cdots \wedge dx^{i_k},$$

and

$$d(d\omega) = \left(\sum_{\ell=1}^{n} \sum_{j=1}^{n} \frac{\partial^2 f}{\partial x^\ell \partial x^j}\, dx^\ell \wedge dx^j \right) \wedge (dx^{i_1} \wedge \cdots \wedge dx^{i_k}),$$

using associativity of the wedge product, Proposition 4.3.5.

Now we already have

$$\sum_{\ell=1}^{n} \sum_{j=1}^{n} \frac{\partial^2 f}{\partial x^\ell \partial x^j}\, dx^\ell \wedge dx^j = 0,$$

since the first term is symmetric (by Theorem 5.1.9) while the second term is antisymmetric (by definition of the wedge product). You can see this explicitly by interchanging the dummy indices j and ℓ everywhere and getting the negative of what you started with.

Thus the whole expression $d(d\omega)$ must be zero. □

We also have a product rule for the differential of forms.

Proposition 16.3.6. *If α is a k-form and β is an ℓ-form, then*

$$d(\alpha \wedge \beta) = (d\alpha) \wedge \beta + (-1)^k \alpha \wedge (d\beta). \tag{16.3.6}$$

Proof. Again it's easiest to prove this using the coordinate formula (16.3.1). Also, by linearity of both sides, it's enough to prove this for basic forms $\alpha = f\, dx^{i_1} \wedge \cdots \wedge dx^{i_k}$ and $\beta = g\, dx^{j_1} \wedge \cdots \wedge dx^{j_\ell}$. In this case, we have

$$\alpha \wedge \beta = fg\, dx^{i_1} \wedge \cdots \wedge dx^{i_k} \wedge dx^{j_1} \wedge \cdots \wedge dx^{j_\ell}$$

so that

$$\begin{aligned}
d(\alpha\wedge\beta) &= \sum_{m=1}^{n}\frac{\partial(fg)}{\partial x^m}\,dx^m\wedge dx^{i_1}\wedge\cdots\wedge dx^{i_k}\wedge dx^{j_1}\wedge\cdots\wedge dx^{j_\ell}\\
&= \left(\sum_{m=1}^{n}\frac{\partial f}{\partial x^m}\,dx^m\wedge dx^{i_1}\wedge\cdots\wedge dx^{i_k}\right)\wedge\left(g\,dx^{j_1}\wedge\cdots\wedge dx^{j_\ell}\right)\\
&\qquad+\left(\sum_{m=1}^{n}f dx^m\wedge(dx^{i_1}\wedge\cdots\wedge dx^{i_k}\right)\wedge\left(\frac{\partial g}{\partial x^m}\,dx^{j_1}\wedge\cdots\wedge dx^{j_\ell}\right)\\
&= (d\alpha)\wedge\beta\\
&\qquad+(-1)^k(f\,dx^{i_1}\wedge\cdots\wedge dx^{i_k})\wedge\left(\sum_{m=1}^{n}\frac{\partial g}{\partial x^m}\,dx^m\wedge dx^{j_1}\wedge\cdots\wedge dx^{j_\ell}\right)\\
&= (d\alpha)\wedge\beta+(-1)^k\alpha\wedge(d\beta).
\end{aligned}$$

□

16.4. **The pull-back on forms.** Suppose M is an m-dimensional manifold and N is an n-dimensional manifold, and suppose we have a smooth map $\eta\colon M\to N$. We constructed in Definition 11.1.1 the push-forward operation, which gives for any $p\in M$, an operator $\eta_*\colon T_pM\to T_{\eta(p)}N$ which pushes vectors from the domain to the range of η. Using this operation, we get an operation going backwards on the cotangent spaces, naturally called the pull-back operation, denoted by $\eta^*\colon T^*_{\eta(p)}N\to T^*_pM$ and defined exactly as in Definition 4.1.4: for any $\beta\in T^*_{\eta(p)}N$, the cotangent vector is $\eta^*\beta\in T^*_pM$ which acts on any tangent vector $v\in T_pM$ by $(\eta^*\beta)(v)=\beta(\eta_*v)$.

In the discussion before Definition 14.2.6, we tried to extend the push-forward from a map defined on spaces of vectors at a point to a map defined on the space of vector fields, but we saw that this generally doesn't work unless η is a diffeomorphism. However the pull-back operation *does* extend from an operation on each cotangent space to an operation on 1-forms (cotangent vector fields), regardless of what η is.

Definition 16.4.1. If M and N are manifolds and $\eta\colon M\to N$, then for any k-form ω defined on N, there is a k-form $\eta^{\#}\omega$ defined on M and called the *pull-back*[35] of ω.

We define it by the following operation on vectors at each T_pM:

$$\eta^{\#}\omega\big((X_1)_p,\cdots,(X_k)_p\big)\equiv\omega\big(\eta_*[(X_1)_p],\cdots,\eta_*[(X_k)_p]\big). \tag{16.4.1}$$

For 0-forms (i.e., functions $f\colon N\to\mathbb{R}$), we define the pull-back by

$$\eta^{\#}(p)=f\big(\eta(p)\big).$$

Notice that, point by point, this is exactly the same operation η^* we get from Definition 4.3.9, when we think of $\eta_*\colon T_pM\to T_{\eta(p)}N$ as being the linear transformation. Then for any $\omega\in\Omega^k(N)$, $(\eta^{\#}\omega)_p\in\Omega^k(T_pM)$ is precisely the operation

[35]Many authors use η^* to denote both the pull-back map in each cotangent space as well as the pull-back map on the space of k-forms. I have found this can be confusing initially, so I am using $\eta^{\#}$ instead of η^* for the operation on fields.

defined by (4.3.9). In case you were wondering, *this* is why we went to all that trouble of working out the linear algebra of $\Omega^k(V)$ for general vector spaces. Now everything is easy.

Observe that $\eta^\# \omega$ is defined at every point $p \in M$, since all we have to do is push forward vectors in T_pM to $T_{\eta(p)}N$, operate on them using $\omega_{\eta(p)}$, and obtain a real number. So a map η from M to N automatically gives a k-form on M for every k-form on N, and this is true whether η has maximal rank at some or all points, and indeed whether M and N have the same dimension or not. In particular we don't need η to be a diffeomorphism. So k-forms can easily be pulled back from one manifold to another, unlike vector fields which are hard to push forward: recall that the push-forward $\eta_\# X$ of a vector field X defined in Definition 14.2.6 only works when η is a diffeomorphism.

Now let's work out something explicit.

Example 16.4.2 (The pull-back of a 2-form)**.** Consider the map $\eta\colon \mathbb{R}^3 \to \mathbb{R}^2$ given in Cartesian coordinates by $(u,v) = \eta(x,y,z) = (z^2 - x^2, xy)$. (We will suppose (x,y,z) are the Cartesian coordinates on the domain and (u,v) are the Cartesian coordinates on the range.) Let $\beta = (3u+v)\,du \wedge dv$ be a 2-form on $\mathbb{R}^2$. What is the 2-form $\eta^\#\beta$ on $\mathbb{R}^3$? First we are going to compute it directly, but see Example 16.4.4 for a simpler computation.

To find it, first we compute $(\eta_*)_{(x,y,z)}$:

$$\begin{aligned}
\eta_*\left(\frac{\partial}{\partial x}\Big|_{(x,y,z)}\right) &= \frac{\partial u}{\partial x}\frac{\partial}{\partial u}\Big|_{(u,v)} + \frac{\partial v}{\partial x}\frac{\partial}{\partial v}\Big|_{(u,v)} \\
&= -2x\,\frac{\partial}{\partial u}\Big|_{(u,v)} + y\,\frac{\partial}{\partial v}\Big|_{(u,v)}, \\
\eta_*\left(\frac{\partial}{\partial y}\Big|_{(x,y,z)}\right) &= x\,\frac{\partial}{\partial v}\Big|_{(u,v)}, \\
\eta_*\left(\frac{\partial}{\partial z}\Big|_{(x,y,z)}\right) &= 2z\,\frac{\partial}{\partial u}\Big|_{(u,v)}.
\end{aligned}$$

From these formulas and the definition (16.4.1), we have

$$\begin{aligned}
(\eta^\#\beta)_{(x,y,z)}\left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right) &= \beta_{\eta(x,y,z)}\left[\eta_*\left(\frac{\partial}{\partial x}\right), \eta_*\left(\frac{\partial}{\partial y}\right)\right] \\
&= (3u+v)\,(du\wedge dv)\left(-2x\,\frac{\partial}{\partial u} + y\,\frac{\partial}{\partial v}, x\,\frac{\partial}{\partial v}\right) \\
&= (3z^2 - 3x^2 + xy)(-2x^2), \\
(\eta^\#\beta)_{(x,y,z)}\left(\frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right) &= (3z^2 - 3x^2 + xy)(-2xz), \\
(\eta^\#\beta)_{(x,y,z)}\left(\frac{\partial}{\partial z}, \frac{\partial}{\partial x}\right) &= (3z^2 - 3x^2 + xy)(2yz).
\end{aligned}$$

These three formulas completely determine the 2-form $\eta^\#\beta$ in $\mathbb{R}^3$, since $\Omega^2(T_p\mathbb{R}^3)$ is three-dimensional. Thus we have

$$(\eta^\#\beta)_{(x,y,z)} = (3z^2 - 3x^2 + xy)\big(-2x^2\,dx\wedge dy - 2xz\,dy\wedge dz + 2yz\,dz\wedge dx\big).$$

☺

The pull-back operation has a few remarkable properties. Intuitively they can be thought of as arising from coordinate-invariance: if η happens to be a diffeomorphism, then when we write coordinates charts $\mathbf{x}$ and $\mathbf{y}$ on the domain and range of η, we see that η is basically just a coordinate transformation, since the only thing we ever actually use in coordinate computations is the map $\mathbf{y} \circ \eta \circ \mathbf{x}^{-1}$. Hence we'd expect the formulas below to be true just based on the fact that the wedge product and differential of forms are coordinate-invariant notions. The nice thing is that they work even if η is not a diffeomorphism.

Proposition 16.4.3. *Suppose M and N are manifolds. If α is any j-form and β is any k-form on N, then for any smooth map $\eta\colon M \to N$, we have*

$$\eta^{\#}(\alpha \wedge \beta) = (\eta^{\#}\alpha) \wedge (\eta^{\#}\beta). \tag{16.4.2}$$

Furthermore,

$$d(\eta^{\#}\alpha) = \eta^{\#}(d\alpha). \tag{16.4.3}$$

Proof. To prove (16.4.2) when both j and k are positive, we just use formula (4.3.6). Let $X_1, \cdots, X_{j+k}$ be any $(j+k)$ vector fields on $\mathbb{R}^m$. For the left side, we have

$$\begin{aligned}
\big(\eta^{\#}(\alpha \wedge \beta)\big)_p(X_1, \cdots, X_{j+k}) &= (\alpha \wedge \beta)\big(\eta_*(X_1)_p, \cdots, \eta_*(X_{j+k})_p\big) \\
&= \sum_{\sigma \in S_{j+k}} \operatorname{sgn}(\sigma)\, \alpha\big(\eta_*(X_{\sigma(1)})_p, \cdots, \eta_*(X_{\sigma(j)})_p\big) \\
&\qquad \cdot \beta\big(\eta_*(X_{\sigma(j+1)})_p, \cdots, \eta_*(X_{\sigma(j+k)})_p\big) \\
&= \sum_{\sigma \in S_{j+k}} \operatorname{sgn}(\sigma)\, (\eta^{\#}\alpha)_p(X_{\sigma(1)}, \cdots, X_{\sigma(j)}) \\
&\qquad \cdot (\eta^{\#}\beta)_p(X_{\sigma(j+1)}, \cdots, X_{\sigma(j+k)}) \\
&= \big[(\eta^{\#}\alpha) \wedge (\eta^{\#}\beta)\big]_p\big(X_1, \cdots, X_{j+k}\big).
\end{aligned}$$

The other cases are when one of j or k is zero, and when both j and k are zero. If $j = 0$ and $k > 0$, then α is a function f, so that $\alpha \wedge \beta = f\beta$. Thus for any k vector fields $X_1, \cdots, X_k$, we have

$$\begin{aligned}
\eta^{\#}(f\beta)_p(X_1, \cdots, X_k) &= f\beta(\eta_*(X_1)_p, \cdots, \eta_*(X_k)_p) \\
&= f\big(\eta(p)\big)\beta(\eta_*(X_1)_p, \cdots, \eta_*(X_k)_p) = (\eta^{\#}f)(p)(\eta^{\#}\beta)_p(X_1, \cdots, X_k)
\end{aligned}$$

so that $\eta^{\#}(f \cdot \beta) = (\eta^{\#}f) \cdot (\eta^{\#}\beta)$.

If j and k are both zero, then obviously

$$\eta^{\#}(fg) = (fg) \circ \eta = (f \circ \eta)(g \circ \eta) = (\eta^{\#}f)(\eta^{\#}g).$$

Thus we have the product rule (16.4.2) in all cases.

Now we want to prove that $d(\eta^{\#}\alpha) = \eta^{\#}(d\alpha)$. We can do this in coordinates: choose charts $(\mathbf{x}, U)$ on the m-dimensional manifold M and $(\mathbf{y}, V)$ on the n-dimensional manifold N. write

$$\alpha = \sum_{1 \le i_1 < \cdots < i_k \le n} \alpha_{i_1 \cdots i_k}(\mathbf{y})\, dy^{i_1} \wedge \cdots \wedge dy^{i_k};$$

since both d and η^* are linear operators, it's enough to prove the formula when $\alpha = f(\mathbf{y})\, dy^{i_1} \wedge \cdots \wedge dy^{i_k}$. First, by the pull-back product rule (16.4.2), we have

$$\eta^\# \alpha = f\big(\eta(\mathbf{x})\big)\, (\eta^\# dy^{i_1}) \wedge \cdots \wedge (\eta^\# dy^{i_k}). \tag{16.4.4}$$

Let's first compute $\eta^\# dy^i$: we have for any index $j \in \{1, \cdots, m\}$ that

$$\eta^\# dy^i \left(\frac{\partial}{\partial x^j} \right) = dy^i \left(\eta_* \frac{\partial}{\partial x^j} \right) = dy^i \left(\sum_{m=1}^n \frac{\partial (y^m \circ \eta \circ \mathbf{x}^{-1})}{\partial x^j} \frac{\partial}{\partial y^m} \right) = \frac{\partial y^i \circ \eta \circ \mathbf{x}^{-1}}{\partial x^j}.$$

Therefore

$$\eta^\# dy^i = \sum_{j=1}^n \frac{\partial y^i \circ \eta \circ \mathbf{x}^{-1}}{\partial x^j}\, dx^j = d(y^i \circ \eta) = d(\eta^\# y^i).$$

So formula (16.4.4) becomes

$$\eta^\# \alpha = f\big(\eta(\mathbf{x})\big)\, d(\eta^\# y^{i_1}) \wedge \cdots \wedge d(\eta^\# y^{i_k}),$$

and by the product rule for differentials (16.3.6) (and the fact that $d^2 = 0$), we have

$$\begin{aligned}
d(\eta^\# \alpha) &= d\big(f \circ \eta(\mathbf{x})\big) \wedge d(\eta^\# y^{i_1}) \wedge \cdots \wedge d(\eta^\# y^{i_k}) \\
&= \sum_{j=1}^m \frac{\partial}{\partial x^j} (f \circ \eta)\, dx^j \wedge d(\eta^\# y^{i_1}) \wedge \cdots \wedge d(\eta^\# y^{i_k}) \\
&= \sum_{j=1}^m \sum_{l=1}^n \frac{\partial f}{\partial y^l}\big(\eta(\mathbf{x})\big) \frac{\partial y^l \circ \eta \circ \mathbf{x}^{-1}}{\partial x^j}\, dx^j \wedge d(\eta^\# y^{i_1}) \wedge \cdots \wedge d(\eta^\# y^{i_k}) \\
&= \sum_{l=1}^n \frac{\partial f}{\partial y^l}\big(\eta(\mathbf{x})\big)\, d(\eta^\# y^l) \wedge d(\eta^\# y^{i_1}) \wedge \cdots \wedge d(\eta^\# y^{i_k}) \\
&= \sum_{l=1}^n \eta^\# \left(\frac{\partial f}{\partial y^l} \right)\, \eta^\#(dy^l) \wedge \eta^\#(dy^{i_1}) \wedge \cdots \wedge \eta^\#(dy^{i_k}) \\
&= \eta^\# \left(\sum_{l=1}^n \frac{\partial f}{\partial y^l}\, dy^l \wedge dy^{i_1} \wedge \cdots \wedge dy^{i_k} \right) \\
&= \eta^\#(d\alpha).
\end{aligned}$$

□

We can also get a coordinate-independent proof of (16.4.3) using the definition 16.2.2, at least when η is a diffeomorphism, using the fact that $\eta_\# [X, Y] = [\eta_\# X, \eta_\# Y]$ in that case from Proposition 14.2.9, along with the fact that

$$(\eta^\# \omega)(X_1, \cdots, X_k) = \big(\omega(\eta_\# X_1, \cdots, \eta_\# X_k)\big) \circ \eta.$$

The difficulty in dealing with $\eta_\# X$ when η is not a diffeomorphism seem to me to make such a proof in the general case more trouble than it's worth.

Finally let's redo Example 16.4.2 using these shortcuts.

Example 16.4.4. Again suppose $\eta \colon \mathbb{R}^3 \to \mathbb{R}^2$ is given in Cartesian coordinates by $(u, v) = \eta(x, y, z) = (z^2 - x^2, xy)$, and let $\beta = (3u + v)\, du \wedge dv$ be a 2-form on $\mathbb{R}^2$. We want to compute $\eta^\# \beta$.

First notice that

$$\begin{aligned}\eta^{\#}\beta &= \eta^{\#}\big((3u+v)\,du\wedge dv\big)\\ &= \big[(3u+v)\circ\eta\big]\,\eta^{\#}(du)\,\wedge\,\eta^{\#}(dv)\\ &= \big[(3u+v)\circ\eta\big]\,d(u\circ\eta)\wedge d(v\circ\eta).\end{aligned}$$

Now $u\circ\eta = z^2 - x^2$ and $d(u\circ\eta) = -2x\,dx + 2z\,dz$, while $v\circ\eta = xy$ and $d(v\circ\eta) = y\,dx + x\,dy$. Therefore we have

$$\begin{aligned}\eta^{\#}\beta &= \big(3(z^2-x^2)+(xy)\big)\,(-2x\,dx+2z\,dz)\wedge(y\,dx+x\,dy)\\ &= (3z^2-3x^2+xy)\,(-2xz\,dy\wedge dz + 2yz\,dz\wedge dx - 2x^2\,dx\wedge dy).\end{aligned}$$

This gives the same answer much more quickly, especially on higher-dimensional manifolds. ☺

17. Integration and Stokes' Theorem

"Size matters not. Look at me. Judge me by my size, do you?"

In this Chapter, we will define integrals of differential forms over k-chains, which are a generalization of parametrized submanifolds. This allows us to discuss an easy version of Stokes' Theorem, which generalizes the Fundamental Theorem of Calculus. Then we discuss integration over (nonparametrized) manifolds using partitions of unity. Finally we define manifolds with boundary, which are the language in which the "real" version of Stokes' Theorem is written.

In multivariable calculus, we can integrate functions over volumes or surfaces or curves, as long as we multiply by the correct "volume element" or "area element" or "length element." For example, if we have a function $f\colon \mathbb{R}^3 \to \mathbb{R}$ given by $f(x,y,z) = x^2$, we can integrate it over the unit sphere by using spherical coordinates $\psi = (\theta, \phi)$, where $0 < \theta < \pi$ and $0 < \phi < 2\pi$, and the area element is $dA = \sin\theta\, d\theta\, d\phi$. The function in spherical coordinates is $f \circ \psi^{-1}(\theta,\phi) = \sin^2\theta \cos^2\phi$, and we obtain

$$\int_{S^2} f\, dA = \int_0^{2\pi} \int_0^{\pi} \sin^2\theta \cos^2\phi\, d\theta\, d\phi = \frac{\pi^2}{2}.$$

Now there are two difficulties here. One is that in order to perform the integration, I have to use a coordinate chart, but this chart does not cover the entire manifold. The set of points not covered is diffeomorphic to a segment, which should have "measure zero," and so we don't expect there to be any meaningful contribution to the integral from this portion; however it's certainly not a satisfactory definition if it relies on this particular coordinate chart. The second difficulty is why the area element should be $dA = \sin\theta\, d\theta\, d\phi$. This is natural in Euclidean geometry (where I assume the square $[0,1] \times [0,1]$ has unit area), but that won't make sense on a general manifold without a notion of length (which leads to area, volume, etc.).

It is fair to object that I shouldn't be trying to integrate functions if I don't have a notion of length. The way to resolve this is not by defining a notion of length, but rather by not trying to integrate functions. Instead we should be integrating differential forms. From a certain point of view, this is already intuitive, and the notation is suggestive. For example every 1-form ω on $\mathbb{R}$ is of the form $f(x)\, dx$ for some function f. Assume ω has compact support; then it makes sense to compute

$$I = \int_{\mathbb{R}} \omega = \int_{-\infty}^{\infty} f(x)\, dx$$

which is actually finite. If I change coordinates by $x = h(u)$, then the integral of course changes to $I = \int_{-\infty}^{\infty} f\big(h(u)\big) h'(u)\, du$, and so it looks like the function has changed to $f\big(h(u)\big)h'(u)$. However the 1-form has stayed exactly the same.

17.1. Line integrals and 1-forms.

More generally we can define the integral of a 1-form along a curve as follows.

Definition 17.1.1. Suppose ω is a 1-form field on a smooth manifold and that $\gamma\colon [a,b] \to \mathbb{R}$ is a smooth curve. Then we define $\int_\gamma \omega$, the *line integral of ω along*

γ, by the formula

$$\int_\gamma \omega = \int_a^b \gamma^\# \omega = \int_a^b \omega\big(\gamma'(t)\big)\, dt. \tag{17.1.1}$$

Notice that in the usual coordinate t on $\mathbb{R}$, we have by Proposition 11.1.1 that

$$\gamma_* \left(\frac{\partial}{\partial t} \right) = \frac{d}{dt}\gamma(t) = \gamma'(t).$$

This is a strange-looking formula, but it's a special case of the definition where we view $\mathbb{R}$ as a manifold and the identity map as a curve in that manifold, so that $\gamma\colon \mathbb{R} \to M$ is a map of manifolds which takes the identity map on $\mathbb{R}$ to the curve $t \mapsto \gamma(t)$. Thus by Definition 16.4.1 the 1-form $\gamma^\#\omega$ acts on the vector $\frac{\partial}{\partial t}$ by

$$(\gamma^\#\omega) \left(\frac{\partial}{\partial t} \right) = \omega \left(\gamma_* \left(\frac{\partial}{\partial t} \right) \right) = \omega\big(\gamma'(t)\big).$$

Thus $\gamma^\#\omega = \omega\big(\gamma'(t)\big)\, dt$, where dt is the 1-form on $\mathbb{R}$. Hence the second and third terms of (17.1.1) really are equal.

Example 17.1.2. Suppose $M \cong \mathbb{R}^2$ and ω is the 1-form given in coordinates by $\omega = x^3 y\, dx + x\, dy$, and let $\gamma\colon [0, 2\pi] \to M$ be given in coordinates by $\gamma(t) = (\cos t, \sin 2t)$. Then $\gamma'(t) = -\sin t \, \frac{\partial}{\partial x}\big|_{\gamma(t)} + 2\cos 2t \, \frac{\partial}{\partial y}\big|_{\gamma(t)}$, so that

$$\omega(\gamma'(t)) = (x^3 y)|_{\gamma(t)} \cdot (-\sin t) + x|_{\gamma(t)} \cdot (2\cos 2t) = -\cos^3 t \sin 2t \sin t + 2\cos t \cos 2t.$$

Thus the line integral is

$$\int_\gamma \omega = \int_0^{2\pi} \Big(-\cos^3 t \sin 2t \sin t + 2\cos t \cos 2t \Big)\, dt = -\frac{\pi}{4}.$$

☺

In some sense, the set of points forming the curve is more important than the parametrization of the curve. What we have in mind is integrating a 1-form over various 1-dimensional submanifolds, and although any connected 1-dimensional submanifold is equal to the image of some smooth curve, there may be many possible parametrizations of the same submanifold. Fortunately, the line integral does not depend on the choice of parametrization.

Proposition 17.1.3. *Suppose $h\colon [c, d] \to [a, b]$ is a smooth increasing diffeomorphism, let $\gamma\colon [a, b] \to M$ be a smooth curve, and let $\tilde{\gamma} = \gamma \circ h$ be a reparametrization of γ. Then for any smooth 1-form ω on M, we have $\int_{\tilde{\gamma}} \omega = \int_\gamma \omega$.*

Proof. Just use integration by substitution. Write $x = h(u)$ and $\tilde{\gamma}(u) = \gamma(h(u))$. We then have

$$\begin{aligned} \int_{\tilde{\gamma}} \omega &= \int_c^d (\gamma \circ h)^\# \omega = \int_c^d \omega_{\gamma(h(u))}\big((\gamma \circ h)'(u)\big)\, du \\ &= \int_c^d \omega_{\gamma(h(u))}\big(\gamma'(h(u)) \cdot h'(u)\big)\, du = \int_a^b \omega_x\big(\gamma'(x)\big)\, dx = \int_\gamma \omega. \end{aligned}$$

□

Remark 17.1.4. Proposition 17.1.3 is not quite as strong as it seems: notice that all our reparametrizations must have $h' > 0$ everywhere since we needed to assume that $h(c) = a$ and $h(d) = b$. In fact if we had reversed the direction of the curve, we would have flipped the sign of the line integral. For example, take $\gamma(t) = (t, 0)$ in $\mathbb{R}^2$ and suppose $\omega = 2x\,dx$. Then

$$\int_\gamma \omega = \int_0^1 2t\,dt = 1,$$

while if we take $\tilde{\gamma}(t) = (1 - t, 0)$, then we get

$$\int_{\tilde{\gamma}} \omega = \int_0^1 2(1-t)(-1)\,dt = -1.$$

We don't notice this when we're integrating functions over "curves" (segments) in $\mathbb{R}$, because we always implicitly decide that the curve goes from left to right. Of course for curves in higher dimensions, we have to decide how to traverse the curve: if it's diffeomorphic to a segment, we have to decide what the beginning and end are, and if it's diffeomorphic to a circle, we have to decide whether to traverse it clockwise or counterclockwise. This is the one-dimensional version of *orientation*, which we defined in Definition 8.2.14, corresponding to the choice that our basis vector on $\mathbb{R}$ points to the right rather than to the left. In higher dimensions we must make similar choices of orientation: for example in $\mathbb{R}^2$ that we will list the "standard" coordinates and vectors as "right-pointing" followed by "up-pointing," which then leads to "counterclockwise" as being the natural induced orientation on curves. These choices were made for you so long ago that you may have forgotten that they ever could have been any other way, but now we need to remember it.

Given any one-dimensional submanifold possibly with boundary (meaning either a closed curve or a set diffeomorphic to a closed interval) we can thus define the integral of a 1-form over this submanifold by choosing an orientation and then parametrizing it, and the value we get does not depend on the choice of parametrization (though it does depend on the choice of orientation).

Remark 17.1.5. The first time most students see line integrals is in physics, where the line integral $\int_a^b \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r}$ is defined to be the work done in transporting a particle from $\mathbf{r}(a)$ to $\mathbf{r}(b)$ which is subject to the force $\mathbf{F}$. This fits into our framework if we reinterpret force to be a 1-form rather than a vector field. But actually this is the natural way to do it: classical mechanics in the Hamiltonian approach uses the cotangent bundle (positions and momenta) rather than the tangent bundle (positions and velocities), where the derivative of the momentum 1-form is the force 1-form. (In elementary physics the momentum is just the mass times the velocity, but when one tries to generalize mechanics to incorporate constraints or deal with forces that may depend on both position and velocity, this is no longer true, and "conservation of momentum" only works if momentum is redefined.)

The following generalization of the Fundamental Theorem of Calculus is a simple version of Stokes' Theorem, which we will prove in Section 17.3.

Theorem 17.1.6. *Suppose $f\colon M \to \mathbb{R}$ is a smooth function and that $\omega = df$ is the 1-form defined by Proposition 15.2.7. Then for any curve $\gamma\colon [a, b] \to M$ we have*

$$\int_\gamma df = f\big(\gamma(b)\big) - f\big(\gamma(a)\big).$$

Proof. This is literally just the ordinary Fundamental Theorem of Calculus once we write it correctly. We have

$$df\big(\gamma'(t)\big) = \frac{d}{dt}(f \circ \gamma)(t)$$

by definition of df and $\gamma'(t)$, and therefore

$$\int_\gamma df = \int_a^b df\big(\gamma'(t)\big)\,dt = \int_a^b \frac{d}{dt}\big((f \circ \gamma)(t)\big)\,dt = (f \circ \gamma)(b) - (f \circ \gamma)(a).$$

□

However we can also do something interesting with line integrals on manifolds which is not interesting on $\mathbb{R}$. On $\mathbb{R}$, every 1-form is a differential of some function: we can write $\omega = g(x)\,dx$ and we know that if $f(x) = \int_a^x g(t)\,dt$ for any $a \in \mathbb{R}$, then $\omega = f'(x)\,dx = df$. Already on $\mathbb{R}^2$ we have found that most 1-forms are not differentials of functions: $\omega = h\,dx + j\,dy$ is equal to df for some f if and only if $h_y = j_x$, as in the Poincaré Lemma 15.2.10. In other words, on $M = \mathbb{R}^2$ we can determine whether a 1-form is the differential of a function by checking to see whether $d\omega = 0$. In the next Theorem we will see how to determine this without actually computing a derivative, which becomes useful when we want to talk about algebraic topology (which involves only continuous functions, not differentiable ones).

First we want to generalize the notion of integral along a curve to curves that are only piecewise smooth. In fact the notion of integral along a curve can be generalized much more, but the piecewise smooth notion is as much as we need. In fact with more work we could state and prove everything just for smooth curves, but we'd have the annoying technical issue of having to smooth everything out.

Definition 17.1.7. A *piecewise-smooth curve* $\gamma\colon [a,b] \to M$ is a continuous function such that there are times $\{t_k \mid 0 \le k \le m\}$, with $a = t_0 < t_1 < \cdots < t_m = b$, such that $\gamma|_{[t_{k-1},t_k]}$ is a smooth curve on each interval $[t_{k-1}, t_k]$.

The integral of a 1-form ω over a piecewise-smooth curve γ is

$$\int_\gamma \omega = \sum_{k=1}^m \int_{\gamma|_{[t_{k-1},t_k]}} \omega.$$

A piecewise curve is continuous but its tangent vector may have jump discontinuities at the break points t_k. There is a bit of concern over whether the integral definition is consistent: for example you could take a smooth curve on $[a,b]$, set $t_1 = \frac{a+b}{2}$, then decide that γ is a piecewise smooth curve with break point at t_1. Fortunately the integral is the same since in one-dimensional calculus you can always break up the integral of a continuous function $f(t)$ as

$$\int_a^b f(t)\,dt = \int_a^{t_1} f(t)\,dt + \int_{t_1}^b f(t)\,dt.$$

Now we show how to determine whether a 1-form ω is the differential of a function without computing $d\omega$.

Theorem 17.1.8. *Suppose M is a path-connected smooth manifold. Suppose ω is a 1-form on M and that for every piecewise-smooth curve $\gamma\colon [a,b] \to M$ with $\gamma(a) = \gamma(b)$ we have $\int_\gamma \omega = 0$. Then $\omega = df$ for some smooth function $f\colon M \to \mathbb{R}$; in particular we have $d\omega = 0$.*

Proof. Certainly if we already knew that $\omega = df$, then for any curve $\gamma\colon [0,1] \to M$ with $\gamma(0) = p$ and $\gamma(1) = q$ we would have $f(q) - f(p) = \int_\gamma \omega$. Furthermore we can assume without loss of generality that at some fixed point p we have $f(p) = 0$, since adding a constant to f does not change df. The idea is thus to use this formula to define $f(q)$, then check that we actually end up with $df = \omega$.

Fix a point $p \in M$. For any point $q \in M$, construct a piecewise-smooth curve from p to q as follows: first choose an arbitrary continuous curve $\eta\colon [0,1] \to M$ such that $\eta(0) = p$ and $\eta(1) = q$. Since $\eta[0,1]$ is compact, there are finitely many coordinate charts (ϕ_k, U_k) for $1 \leq k \leq m$ which cover $\eta[0,1]$. Rearrange the indices of U_k appropriately so that they appear in order along γ, and choose points $t_k \in [0,1]$ such that $\eta[t_{k-1}, t_k] \subset U_k$. Since the continuous curve $\eta[t_{k-1}, t_k]$ is completely contained in a coordinate chart, we can replace it with a smooth curve $\gamma\colon [t_{k-1}, t_k] \to M$ which has the same endpoints (for example, by taking the curve which is a straight line segment in coordinates). The union of these curves is then a piecewise-smooth curve γ with endpoints p and q.

Having constructed a curve γ with $\gamma(0) = p$ and $\gamma(1) = q$, define $f(q) = \int_\gamma \omega$. This definition does not depend on the choice of piecewise-smooth curve γ, since if γ_1 and γ_2 were two piecewise-smooth curves with $\gamma_i(0) = p$ and $\gamma_i(1) = q$, we could concatenate them to get

$$\gamma_3(t) = \begin{cases} \gamma_1(t) & 0 \leq t \leq 1 \\ \gamma_2(2-t) & 1 < t \leq 2. \end{cases}$$

Then γ_3 is a piecewise-smooth curve with $\gamma_3(0) = \gamma_3(2) = p$ which means we have $\int_{\gamma_3} \omega = 0$. However we also have $\int_{\gamma_3} \omega = \int_{\gamma_1} \omega - \int_{\gamma_2} \omega$, where the minus sign comes from the fact that γ_2 is running backwards (using a u-substitution). Thus $\int_{\gamma_1} \omega = \int_{\gamma_2} \omega$.

We now have a well-defined function f. It remains to check that $df = \omega$. This is easiest to do in coordinates, since it's just a local computation. In a coordinate chart, we can write $\omega = \sum_{k=1}^n \omega_k(x^1, \dots, x^n)\, dx^k$. To compute $\frac{\partial f}{\partial x^1}$, we note that

$$f(x^1, x^2, \dots, x^n) = f(0, x^2, \dots, x^n) + \int_\gamma \omega,$$

where γ in coordinates is given by $\gamma(t) = (tx^1, x^2, \dots, x^n)$. We then have $\gamma'(t) = x^1 \frac{\partial}{\partial x^1}\big|_{\gamma(t)}$, so that

$$\omega\big(\gamma'(t)\big) = x^1 \omega_1(tx^1, x^2, \dots, x^n).$$

Thus we have

$$f(x^1, x^2, \dots, x^n) = f(0, x^2, \dots, x^n) + x^1 \int_0^1 \omega_1(tx^1, x^2, \dots, x^n)\, dt.$$

Now change variables in the integral to $s = tx^1$; then we have

$$f(x^1, x^2, \dots, x^n) = f(0, x^2, \dots, x^n) + \int_0^{x^1} \omega_1(s, x^2, \dots, x^n)\, ds,$$

and it is now clear that

$$\frac{\partial f}{\partial x^1}(x^1, x^2, \dots, x^n) = \omega_1(x^1, x^2, \dots, x^n).$$

The same works for any other index, and we obtain $\frac{\partial f}{\partial x^k} = \omega_k$ for each k, so that $\omega = df$. In particular f itself is smooth in coordinates, and by Proposition 16.3.5 we have $d\omega = 0$. □

We have shown that if $\int_\gamma \omega = 0$ for any piecewise-smooth closed curve γ, then $\omega = df$ for some smooth function f, and the converse is true by Proposition 17.1.6 (which clearly applies as well to piecewise-smooth curves). Hence if $\int_\gamma \omega = 0$ for any piecewise-smooth closed curve, then $d\omega = 0$. The converse is not true however. There are 1-forms ω such that $d\omega = 0$ but ω is not df for any smooth function f, and the easiest way to prove this is to show that there is at least one piecewise-smooth closed curve γ such that $\int_\gamma \omega \neq 0$. Clearly this cannot happen if $M = \mathbb{R}^2$ by the Poincaré Lemma 15.2.10, but if $M = \mathbb{R}^2 \backslash \{\mathbf{0}\}$ it can.

Example 17.1.9. Suppose $M = \mathbb{R}^2 \backslash \{\mathbf{0}\}$, and let $\omega_{(x,y)} = h(x,y)\,dx + j(x,y)\,dy$, where

$$h(x,y) = -\frac{y}{x^2+y^2} \qquad \text{and} \qquad j(x,y) = \frac{x}{x^2+y^2}.$$

Then we have

$$\frac{\partial h}{\partial y} = \frac{y^2 - x^2}{(x^2+y^2)^2} = \frac{\partial j}{\partial x},$$

so that $d\omega = 0$. Although the condition in the Poincaré Lemma 15.2.10 is satisfied, ω cannot be df for any function $f\colon M \to \mathbb{R}$ (even if we exclude the origin). Essentially the reason for this is that $f(x,y)$ "wants to be" $\theta(x,y)$, given by formula (5.2.4) on the plane minus the negative x-axis; in other words $f(x,y) = \arctan(y/x)$ and the smooth extension of this. This function θ has a jump on the negative x-axis, where it goes from $-\pi$ below to $+\pi$ above, and hence it is not even continuous. The fact that $f - \theta$ is constant comes from the fact that $d(f-\theta) = 0$, so f must have the same difficulty.

A simpler yet more rigorous proof is to use the contrapositive of Theorem 17.1.6: if there is a curve $\gamma\colon [a,b] \to M$ with $\gamma(a) = \gamma(b)$ but $\int_\gamma \omega \neq 0$, then ω cannot be df for any smooth function f. Using the curve $\gamma(t) = (\cos t, \sin t)$ on $[0, 2\pi]$, we get

$$(\gamma^\# \omega)_t = \omega(\gamma'(t))\,dt = -\frac{\sin t}{\sin^2 t + \cos^2 t}\,d(\cos t) + \frac{\cos t}{\sin^2 t + \cos^2 t}\,d(\sin t) = dt.$$

Therefore we have

$$\int_\gamma \omega = \int_0^{2\pi} \omega(\gamma'(t))\,dt = \int_0^{2\pi} dt = 2\pi,$$

in spite of the fact that $\gamma(0) = \gamma(2\pi)$. So ω cannot be the differential of a function. ☺

This example is one of the most important and basic examples in the entire subject of differential forms, so you should be familiar with it. It's the first counterexample to the conjecture that $d\omega = 0$ implies that $\omega = df$, and quantifying the failure of this property ends up leading to de Rham cohomology.

17.2. **Integration of k-forms.** We now want to define the basic language in which Stokes' Theorem is stated. In the last Section, we saw that the Fundamental Theorem of Calculus on $\mathbb{R}$ generalizes to line integrals of a 1-form on a manifold. To proceed, we first want the correct notion of integration on $\mathbb{R}^k$, and then we will see how to integrate k-forms. First we have to handle the issue of orientation, which

arose already in the one-dimensional case (Remark 17.1.4). There we observed that integrals of 1-forms made sense on one-dimensional submanifolds as long as we were able to specify a forward direction; in the more general case we have to distinguish between things like the outside surface of a sphere and the inside surface. For right now, this is taken care of automatically because we are parametrizing everything by subsets of $\mathbb{R}^k$, so all we need is an orientation for $\mathbb{R}^k$ itself. Later we will generalize this to manifolds and connect it to the notion of orientability of a manifold from Definition 8.2.14.

Definition 17.2.1. On $\mathbb{R}^k$ or any subset of it, we define the *standard orientation* to be the k-form $\mu = dx^1 \wedge \cdots \wedge dx^k$ where $(x^1, \ldots, x^k)$ are the Cartesian coordinates on $\mathbb{R}^k$. A nowhere-zero k-form ω on $\mathbb{R}^k$ or a connected subset of it can always be written as $\omega = f\mu$ for some nowhere-zero function f; then ω is called *positively-oriented* if f is always positive and *negatively-oriented* if f is always negative.

Thus for example on $\mathbb{R}^3$ we have the standard orientation $dx \wedge dy \wedge dz$; the 3-form $dz \wedge dx \wedge dy$ is also positively-oriented while the 3-form $dy \wedge dx \wedge dz$ is negatively-oriented. Of course in general one needs to specify a "preferred" ordering of the coordinates that are not named in order to obtain an orientation.

We now define the integral of a k-form over a parametrized k-dimensional submanifold in the same way as we defined line integrals: pull back the k-form using the parametrization map and integrate over a region in $\mathbb{R}^k$. For simplicity we assume that all parametrizations are defined on cubes.

Definition 17.2.2. Let $I = [0, 1]$, and let M be a smooth n-dimensional manifold. A *singular k-cube* in M is the image $c(I^k)$ of a smooth map $c \colon I^k \subset \mathbb{R}^k \to M$.

The *integral of a k-form ω over a singular k-cube c* is defined to be

$$\int_c \omega = \int_{I^k} c^\# \omega;$$

here if $c = c(x^1, \cdots, x^k)$, then $c^\# \omega = f(x^1, \cdots, x^k)\, dx^1 \wedge \cdots \wedge dx^k$ for some function f, and we define

$$\int_{I^k} c^\# \omega = \int_{I^k} f(\mathbf{x})\, dx^1 \wedge \cdots \wedge dx^k = \int_{I^k} f(\mathbf{x})\, dx^1 \cdots dx^k \tag{17.2.1}$$

if $dx^1 \wedge \cdots \wedge dx^k$ is positively oriented.

If the form were negatively oriented, we'd have to insert a minus sign in the integral. The reason is that forms are antisymmetric under transpositions, while iterated integrals, by Fubini's Theorem 5.3.1, are symmetric under transpositions.

The name "singular" comes not from any lack of smoothness, but rather from the fact that we aren't requiring that c be an immersion; in fact it could map the entire cube to a single point. We sometimes just omit the adjective "singular" entirely.

It's easy to see that Definition 17.2.2 reduces to Definition 17.1.1 when $k = 1$. Let's compute an example when $k = 2$.

Example 17.2.3. Suppose $\omega = z\, dx \wedge dy + x\, dy \wedge dz$ on $\mathbb{R}^3$, and that we want to integrate over the unit 2-sphere. Parametrize the sphere on $[0, 1] \times [0, 1]$ by the formula

$$(x, y, z) = \eta(u, v) = \big(\sin(\pi u)\cos(2\pi v), \sin(\pi u)\sin(2\pi v), \cos(\pi u)\big).$$

Then $\eta^{\#}(\omega)$ is a 2-form, which we can compute as in Proposition 16.4.3 as

$$\eta^{\#}(\omega) = (z \circ \eta)\, d(x \circ \eta) \wedge d(y \circ \eta) + (x \circ \eta)\, d(y \circ \eta) \wedge d(z \circ \eta).$$

Computing one term at a time, we get

$$\begin{aligned}
d(x \circ \eta) &= d\big(\sin(\pi u)\cos(2\pi v)\big) = \pi\cos(\pi u)\cos(2\pi v)\, du - 2\pi\sin(\pi u)\sin(2\pi v)\, dv \\
d(y \circ \eta) &= d\big(\sin(\pi u)\sin(2\pi v)\big) = \pi\cos(\pi u)\sin(2\pi v)\, du + 2\pi\sin(\pi u)\cos(2\pi v)\, dv \\
d(z \circ \eta) &= -\pi\sin(\pi u)\, du
\end{aligned}$$

We therefore have

$$\eta^{\#}(dx \wedge dy) = d(x \circ \eta) \wedge d(y \circ \eta) = 2\pi^2 \sin(\pi u)\cos(\pi u)\, du \wedge dv$$

and

$$\eta^{\#}(dy \wedge dz) = d(y \circ \eta) \wedge d(z \circ \eta) = 2\pi^2 \sin^2(\pi u)\cos(2\pi v)\, du \wedge dv.$$

Combining, we thus get that

$$\eta^{\#}(\omega) = 2\pi^2\big[\sin(\pi u)\cos^2(\pi u) + \sin^3(\pi u)\cos^2(2\pi v)\big]\, du \wedge dv.$$

Now if we decide that $du \wedge dv$ is positively-oriented, the integral is then

$$\begin{aligned}
\int_\eta \omega &= \int_0^1 \int_0^1 \eta^{\#}(\omega) \\
&= 2\pi^2 \int_0^1 \int_0^1 \sin(\pi u)\cos^2(\pi u)\, du\, dv + 2\pi^2 \int_0^1 \int_0^1 \sin^3(\pi u)\cos^2(2\pi v)\, du\, dv \\
&= \frac{8\pi}{3}.
\end{aligned}$$

We could have decided that $dv \wedge du$ was positively-oriented, which would have given us the negative of this answer. The choice corresponds to computing the outward flux or the inward flux, and one way of determining this is to take the outward normal $N = x\,\frac{\partial}{\partial x} + y\,\frac{\partial}{\partial y} + z\,\frac{\partial}{\partial z}$, and then decide that the variable which comes first is the one for which

$$dx \wedge dy \wedge dz\left(\eta_*\Big(\frac{\partial}{\partial u}\Big), \eta_*\Big(\frac{\partial}{\partial v}\Big), N_{\eta(u,v)}\right) > 0,$$

in terms of the standard orientation $dx \wedge dy \wedge dz$ on $\mathbb{R}^3$. You can compute that $dx \wedge dy \wedge dz(\eta_*(\frac{\partial}{\partial u}), \eta_*(\frac{\partial}{\partial v}), N) = 2\pi^2 \sin(\pi u) > 0$, so that we made the correct choice by this standard. ☺

If we want to generalize the Fundamental Theorem of Calculus (for 1-forms) given by Theorem 17.1.6, we want to figure out what the boundary of a k-cube is. We already have a dilemma when $k = 1$: the boundary of a 1-cube is not a 0-cube (using the natural definition that a 0-cube is just a point in the manifold). Rather the boundary of a 1-cube is two 0-cubes. In fact since $\int_\gamma df = f(\gamma(b)) - f(\gamma(a))$, it is in some sense natural to think of the boundary of $\gamma\colon [a, b] \to M$ as being $\gamma(b) - \gamma(a)$, where the subtraction of points is shorthand for saying "to apply a function, apply to each point and then subtract."

Now suppose we take a 2-cube (a square). Obviously the boundary consists of four segments, which we can either write as a single piecewise-smooth curve or as the union of four disjoint 1-cubes. Again we will find it convenient to write this as a linear combination of k-cubes, as shown in the following proposition.

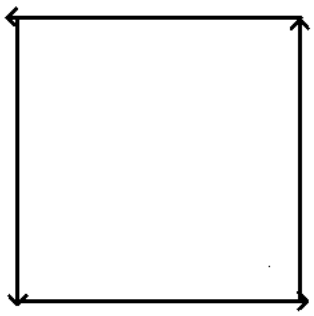

FIGURE 17.1. The boundary of I^2

Proposition 17.2.4. *Let ω be a 1-form on $\mathbb{R}^2$ and set $\beta = d\omega$. Then*

$$\int_{I^2} \beta = \int_{\iota_1^+} \omega - \int_{\iota_1^-} \omega - \int_{\iota_2^+} \omega + \int_{\iota_2^-} \omega, \tag{17.2.2}$$

where $\iota_1^+(t) = (1,t)$, $\iota_1^-(t) = (0,t)$, $\iota_2^+(t) = (t,1)$, and $\iota_2^-(t) = (t,0)$.

Proof. Any 1-form on $\mathbb{R}^2$ can be written as $\omega = f(x,y)\,dx + g(x,y)\,dy$ for some functions f and g, and we have

$$\beta = d\omega = \left(\frac{\partial g}{\partial x}(x,y) - \frac{\partial f}{\partial y}(x,y)\right) dx \wedge dy.$$

Using the standard orientation on $\mathbb{R}^2$, we write the integral over the square as

$$\begin{aligned}
\int_{I^2} \beta &= \int_0^1 \int_0^1 \left(\frac{\partial g}{\partial x}(x,y) - \frac{\partial f}{\partial y}(x,y)\right) dx\,dy \\
&= \int_0^1 \int_0^1 \frac{\partial g}{\partial x}(x,y)\,dx\,dy - \int_0^1 \int_0^1 \frac{\partial f}{\partial y}(x,y)\,dy\,dx \\
&= \int_0^1 \big[g(1,y) - g(0,y)\big]\,dy - \int_0^1 \big[f(x,1) - f(x,0)\big]\,dx \\
&= \int_0^1 (\iota_1^+)^\# \omega - \int_0^1 (\iota_1^-)^\# \omega - \int_0^1 (\iota_2^+)^\# \omega + \int_0^1 (\iota_2^-)^\# \omega,
\end{aligned}$$

which is (17.2.2) by Definition 17.1.1. □

We thus define the boundary of the square to be $\partial c = \iota_1^+ - \iota_1^- - \iota_2^+ + \iota_2^-$, which makes sense since the only thing we're going to do with these sums of curves is to compute line integrals on them (by integrating along each curve and then adding up the integrals). See Figure 17.1. The reason to do this is that formula (17.2.2) becomes $\int_c d\omega = \int_{\partial c} \omega$. This is already a neat generalization of the Fundamental Theorem of Calculus, but what's even nicer is that just like Theorem 17.1.6, it generalizes to a formula for integration of arbitrary 2-forms over arbitrary parametrized 2-dimensional surfaces in a manifold of any dimension. And as you might be ready to guess, the entire process generalizes to integrals of k-forms over singular k-cubes. First let's make sense of the boundary of a general k-cube.

Definition 17.2.5. A *k-chain* in M denotes the formal sum of singular k-cubes: $c = i_1 c_1 + \cdots + i_m c_m$, where each $i_j \in \mathbb{Z}$ and each c_j is a singular k-cube in the sense of Definition 17.2.2.

The *boundary* of a singular k-cube is denoted by $\partial c(I^k)$. It consists of the $(k-1)$-chain composed of the sum of the signed images of the $2k$ faces of the k-cube I^k; the faces are parametrized by the maps ι_i^+ and ι_i^- for $1 \le i \le k$, where

$$\begin{aligned} \iota_i^+(u^1, \cdots, u^{k-1}) &= (u^1, \cdots, u^{i-1}, 1, u^i, \cdots, u^{k-1}), \\ \iota_i^-(u^1, \cdots, u^{k-1}) &= (u^1, \cdots, u^{i-1}, 0, u^i, \cdots, u^{k-1}). \end{aligned} \tag{17.2.3}$$

Then we set ∂c to be the $(k-1)$-chain

$$\partial c = \sum_{i=1}^{k} (-1)^{i+1} c \circ \iota_i^+ + (-1)^i c \circ \iota_i^-. \tag{17.2.4}$$

We extend this boundary operator to general k-chains by linearity: $\partial(\sum_j i_j c_j) = \sum_j i_j \partial c_j$.

Finally, the integral of a k-form ω over a singular k-chain $c = \sum_{j=1}^m i_j c_j$ is defined to be

$$\int_c \omega = \sum_{j=1}^{m} i_j \int_{c_j} \omega. \tag{17.2.5}$$

The notion of a chain, while at first rather bizarre, is actually fairly natural, in the sense that we frequently break up integrals over large sets into sums of integrals over subsets. So the notion of chain just formalizes this idea that a set is the sum of its disjoint subsets. The weird part is the subtraction, but this is also natural since cubes *do* have a natural sign on their faces. The two parallel faces, in each dimension, should have the opposite sign, since no matter how one orients the cube, the "outward" facing direction is opposite on those two faces. For this reason, it makes sense to both add and subtract subsets when doing integrations. This is exactly what happened in the 1-dimensional case of Theorem 17.1.6 and the 2-dimensional case of Proposition 17.2.4.

Example 17.2.6. Let's see how the boundary operator of a familiar surface like the disc looks. Intuitively the boundary of a disc is a single curve (the boundary circle) which is already parametrized by a smooth 1-cube, so what happens to the rest of the square?

We can express the disc as a 2-chain using polar coordinates: write the 2-chain $c_2 \colon [0,1]^2 \to \mathbb{R}^2$ as $c_2(s,t) = \big(s\cos(2\pi t), s\sin(2\pi t)\big)$. First let us compute the boundary of the disc, ∂c_2. We have from formula (17.2.4) that

$$\begin{aligned} \partial c_2(t) &= \sum_{i=1}^{2} (-1)^{i+1} c_2 \circ \iota_i^+(t) + (-1)^i c_2 \circ \iota_i^-(t) \\ &= c_2(1,t) - c_2(t,1) - c_2(0,t) + c_2(t,0) \\ &= \big(\cos(2\pi t), \sin(2\pi t)\big) - \big(t,0\big) - \big(0,0\big) + \big(t,0\big). \end{aligned}$$

Observe that the first term goes counterclockwise around the unit circle, starting and ending at $(1,0)$; the second term goes along the horizontal segment to the origin; the third term is fixed at the origin; and the fourth term goes back along the same horizontal segment from the origin to $(1,0)$.

Write $c_1(t) = \big(\cos(2\pi t), \sin(2\pi t)\big)$. It is tempting to say that $\partial c_2 = c_1$, i.e., that the boundary of the disc is just the outer circle. To make this precise, we can define two k-chains η_1 and η_2 to be *equivalent* if $\int_{\eta_1} \omega = \int_{\eta_2} \omega$ for every k-form ω on M. In the present case, the integral of any 1-form over the "curve" $c_2 \circ \iota_2^-$ which sends everything to the origin must be zero. And the integrals of any 1-form over the two horizontal segments $c_2 \circ \iota_2^- - c_2 \circ \iota_1^+$ cancel each other out. Hence ∂c_2 is equivalent to c_1, in the sense that $\int_{\partial c_2} \omega = \int_{c_1} \omega$ for any 1-form ω. Thus the boundary of the disc is the (counterclockwise-traversed) unit circle, which is just what we'd expect. ☺

Remark 17.2.7. It turns out that the boundary of a boundary is always zero, i.e., $\partial(\partial c) = 0$ for any k-chain c. We will prove this in general in Chapter 18, but for now we will just describe the proof for 2-chains. Since the boundary operator ∂ is linear, it is sufficient to prove it for 2-cubes. We have

$$\partial c(t) = c(1,t) - c(t,1) - c(0,t) + c(t,0),$$

a sum of four curves. Furthermore the boundary of a curve $\gamma(t)$ is $\partial\gamma(t) = \gamma(1) - \gamma(0)$, and thus we have

$$\partial\partial c = \big[c(1,1) - c(1,0)\big] - \big[c(1,1) - c(0,1)\big] - \big[c(0,1) - c(0,0)\big] + \big[c(1,0) - c(0,0)\big].$$

Clearly these eight terms cancel out pairwise, so $\partial\partial c$ is actually equal to zero (not just equivalent to zero). Intuitively we start with a solid square, take the boundary to get a hollow square, and the hollow square has no endpoints and thus no boundary.

We can ask the same kinds of questions about boundary operators as we ask about the differentials of forms: for example, if γ is a 1-chain with $\partial\gamma = 0$, is there a 2-chain η such that $\partial\eta = \gamma$? (We allow for equivalence rather than equality to make things easier.)

We will discuss this in more detail later, in Chapter 18, but for now we will just say that if $M = \mathbb{R}^2$, then every 1-chain with empty boundary must be the boundary of some 2-chain: the idea is to write the 1-chain as a linear combination of closed curves, and fill in each closed curve by drawing a segment from each point of the curve to the origin. On the other hand, if $M = \mathbb{R}^2 \backslash \{(0,0)\}$, then the unit circle has empty boundary, but it is not the boundary of any parametrized surface. These results do not rely on the fact that our k-chains are *smooth*; in fact the same things happen if the k-chains are merely continuous, and these results can be proved using general methods of algebraic topology. Thus we will not get into them here.

We should check that this Definition actually is independent of coordinates, since we explicitly introduced coordinates to define (17.2.1). The following Proposition 17.2.9 generalizes the Proposition 17.1.3 on reparametrizing curves in a line integral. First we need a Lemma to show us how to change variables in a top-dimensional form.

Lemma 17.2.8. *Suppose ω is a k-form on a k-dimensional manifold, with two coordinate charts $\mathbf{x}$ and $\mathbf{y}$. If we express*

$$\omega = f(x^1, \ldots, x^k)\, dx^1 \wedge \cdots \wedge dx^k = g(y^1, \ldots, y^k)\, dy^1 \wedge \cdots \wedge dy^k,$$

then

$$f(x^1, \ldots, x^k) \det\left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right) = g(y^1, \ldots, y^k), \tag{17.2.6}$$

where we interpret the left side as a function of $(y^1, \dots, y^k)$ *using the transition function* $\mathbf{x} \circ \mathbf{y}^{-1}$.

Proof. The fact that any k-form on a k-dimensional manifold must be a function multiplied by a basic k-form is a consequence of the fact that the dimension of $\Omega^k(T_pM)$ is $\binom{k}{k} = 1$. The only other thing to verify is the formula (17.2.6), which follows directly from Proposition 4.3.10. We could also prove this directly, using basically the same technique as in Proposition 3.3.4. □

Proposition 17.2.9. *Suppose* $c_1 \colon [0,1]^k \to M$ *and* $c_2 \colon [0,1]^k \to M$ *have the same image, and that* $c_2 \circ c_1^{-1}$ *is an orientation-preserving diffeomorphism (that is,* $\det D(c_2 \circ c_1^{-1}) > 0$*). Then for any* k*-form* ω *on* M*, we have*

$$\int_{c_1} \omega = \int_{c_2} \omega.$$

Proof. Just combine the Change of Variables Theorem 5.3.2 with Lemma 17.2.8. If $c_1 = c_1(x^1, \cdots, x^k)$ and $c_2 = c_2(y^1, \cdots, y^k)$ for some coordinate charts $\mathbf{x}$ and $\mathbf{y}$ on $\mathbb{R}^k$, write $(c_1)^\# \omega = f(x^1, \dots, x^k)\, dx^1 \wedge \cdots \wedge dx^k$ and $(c_2)^\# \omega = g(y^1, \dots, y^k)\, dy^1 \wedge \cdots \wedge dy^k$. Then $\zeta = c_2^{-1} \circ c_1$ is just a coordinate change which preserves the unit cube $[0,1]^k$, giving $\mathbf{y}$ in terms of $\mathbf{x}$. We see that

$$\begin{aligned}\int_{c_2} \omega = \int_{I^k} c_2^\# \omega = \int_{I^k} g(\mathbf{y}) dy^1 \cdots dy^k = \int_{I^k} g\big(\zeta(\mathbf{x})\big)\, \det\left(\tfrac{\partial \mathbf{y}}{\partial \mathbf{x}}\right) dx^1 \cdots dx^k \\ = \int_{I^k} f(\mathbf{x}) dx^1 \cdots dx^k = \int_{I^k} c_1^\# \omega = \int_{c_1} \omega.\end{aligned}$$

□

Note that if c fails to be a diffeomorphism at a single point or more generally on a set of measure zero, this doesn't affect the result, so we could lighten the hypotheses in Proposition 17.2.9 somewhat. However we can't go too far: for example traversing the set more than once or traversing it backwards will change the result.

We have tried to justify the definitions of these things as they came up: a k-form is defined to be an antisymmetric tensor because that's what the derivative of a $(k-1)$-form wants to be. The differential d of a k-form is defined so that it is coordinate-independent (which results in $d\eta^\# \omega = \eta^\# d\omega$). Integrals of k-forms are defined by just changing the basic k-form $dx^1 \wedge \cdots \wedge dx^k$ into the k-dimensional volume element $dx^1 \cdots dx^k$; the transformation formula for k-forms on $\mathbb{R}^k$ ends up coinciding exactly with the transformation formula for the corresponding integrals, and this is why k-forms should be thought of as the "correct" version of the k-dimensional volume form (which otherwise has no meaning except attached to an integral). Finally we defined chains because we wanted to consider boundaries of cubes, which are obviously composed of $2k$ pieces of $(k-1)$-dimensional cubes (with positive and negative orientations). Whether or not you thought it was natural so far, you'll hopefully appreciate the punch line.

17.3. **Stokes' Theorem on chains.**

Theorem 17.3.1. *[Stokes' Theorem] Suppose $k \in \mathbb{N}$. If c is any k-chain on M, and ω is any $(k-1)$-form on M, then*

$$\int_c d\omega = \int_{\partial c} \omega. \tag{17.3.1}$$

Proof. By definition of integration over chains, we know that

$$\int_c d\omega = \int_{I^k} c^\# d\omega = \int_{I^k} d(c^\# \omega).$$

Furthermore we know that

$$\int_{\partial c} \omega = \int_{\partial I^k} c^\# \omega.$$

Hence regardless of what's happening on the manifold M, this really is just a statement about k-forms on $I^k \subset \mathbb{R}^k$. So we lose nothing by working in Cartesian coordinates there.

Now regardless of what ω actually is, $c^\# \omega$ is a $(k-1)$-form on $\mathbb{R}^k$, so it must be (in coordinates)

$$c^\# \omega = \sum_{j=1}^k f_j(x^1, \cdots, x^k)\, dx^1 \wedge \cdots dx^{j-1} \wedge dx^{j+1} \wedge \cdots \wedge dx^k$$

for some k functions $f_j \colon I^k \to \mathbb{R}$. Thus obviously

$$d(c^\# \omega) = \sum_{j=1}^k \sum_{i=1}^k \frac{\partial f_j}{\partial x^i}\, dx^i \wedge dx^1 \wedge dx^{j-1} \wedge dx^{j+1} \wedge \cdots \wedge dx^k.$$

Now this k-form must be zero unless $i = j$ (because otherwise two of the basic 1-forms must be identical), and if $i = j$, then we must have

$$dx^i \wedge dx^1 \wedge dx^{j-1} \wedge dx^{j+1} \wedge \cdots \wedge dx^k = (-1)^{i-1} dx^1 \wedge \cdots \wedge dx^k.$$

Thus the formula for $d(c^\# \omega)$ is

$$d(c^\# \omega) = \sum_{j=1}^k (-1)^{j-1} \frac{\partial f_j}{\partial x^j}\, dx^1 \wedge \cdots \wedge dx^n.$$

Having computed these formulas, we just have to compute the integrals, using the fundamental theorem of calculus.

$$\begin{aligned}
\int_c d\omega &= \int_0^1 \cdots \int_0^1 \sum_{j=1}^k (-1)^{j-1} \frac{\partial f_j}{\partial x^j}\, dx^1 \cdots dx^n \\
&= \sum_{j=1}^k \int_0^1 \cdots \int_0^1 (-1)^{j-1} \left(\int_0^1 \frac{\partial f_j}{\partial x^j}\, dx^j \right) dx^1 \cdots dx^{j-1} dx^{j+1} \cdots dx^n \\
&= \sum_{j=1}^k \int_0^1 \cdots \int_0^1 (-1)^{j-1} \Big[f_j(x^1, \cdots, x^{j-1}, 1, x^{j+1}, \cdots, x^n) \\
&\qquad + (-1)^j f_j(x^1, \cdots, x^{j-1}, 0, x^{j+1}, \cdots, x^n) \Big] dx^1 \cdots dx^{j-1} dx^{j+1} \cdots dx^n
\end{aligned}$$

Now we just compute the other side from the definition (17.2.4) of ∂c. We first see what $(c \circ \iota_i^+)^\# \omega$ is. Since

$$(\iota_i^+)^\# dx^m = \begin{cases} du^m & m < i, \\ 0 & m = i, \\ du^{m-1} & m > i, \end{cases}$$

we know that

$$(\iota_i^+)^\#(dx^1 \wedge \cdots dx^{j-1} \wedge dx^{j+1} \wedge \cdots \wedge dx^k) = \begin{cases} du^1 \wedge \cdots \wedge du^{k-1} & i = j \\ 0 & i \neq j. \end{cases}.$$

Therefore

$$\begin{aligned}(c \circ \iota_i^+)^\# \omega &= (\iota_i^+)^\# c^\# \omega \\ &= (\iota_i^+)^\# \sum_{j=1}^k f_j(x^1, \cdots, x^k)\, dx^1 \wedge \cdots dx^{j-1} \wedge dx^{j+1} \wedge \cdots \wedge dx^k \\ &= f_i(u^1, \cdots, u^{i-1}, 1, u^i, \cdots, u^{k-1})\, du^1 \wedge \cdots \wedge du^{k-1}.\end{aligned}$$

Similarly we get

$$(c \circ \iota_i^-)^\# \omega = f_i(u^1, \cdots, u^{i-1}, 0, u^i, \cdots, u^{k-1})\, du^1 \wedge \cdots \wedge du^{k-1}.$$

Thus we have

$$\begin{aligned}\int_{\partial c} \omega &= \sum_{i=1}^k (-1)^{i+1} \int_{I^{k-1}} (\iota_i^+)^\# c^\# \omega + \int_{I^{k-1}} (-1)^i (\iota_i^-)^\# c^\# \omega \\ &= \sum_{i=1}^k \int_0^1 \cdots \int_0^1 \Big[(-1)^{i+1} f(u^1, \cdots, u^{i-1}, 1, u^i, \cdots, u^{k-1}) \\ &\qquad + (-1)^i f(u^1, \cdots, u^{i-1}, 0, u^i, \cdots, u^{k-1}) \Big]\, du^1 \cdots du^{k-1}.\end{aligned}$$

Clearly this is the same as what we computed for $\int_c d\omega$. □

Although this version of the theorem seems less widely applicable than the version known from vector calculus (since the shape we integrate over has to be the image of a cube), it's really not all that different, since the shape doesn't have to be the *diffeomorphic* image of a cube. For example the unit disc is the image of the cube $[0,1]^2$ using the formula $c(u,v) = \big(u\cos(2\pi v), v\sin(2\pi v)\big)$, as in Example 17.2.6. Of course this map isn't a diffeomorphism, but we don't really demand that. More generally it's quite easy to get any shape you'd typically want as some union of a finite image of cubes.

Example 17.3.2 (Stokes' Theorem for 2-forms in $\mathbb{R}^3$)**.** Let's see how the Divergence Theorem on the unit ball,

$$\int_{B(0,1)} \operatorname{div} X\, dV = \int_{S(0,1)} X \cdot \mathbf{n}\, dS,$$

is a consequence of this general Stokes' Theorem. (Here all the terms are the standard ones in Euclidean coordinates, i.e., $\mathbf{n}$ is the unit normal vector and $\cdot$ is

the Euclidean dot product.) We first get the unit ball $B(0,1) \subset \mathbb{R}^3$ using the following parametrization:

$$c(p,q,r) = \big(p\sin(\pi q)\cos(2\pi r), p\sin(\pi q)\sin(2\pi r), p\cos(\pi q)\big).$$

(Clearly these are just spherical coordinates.)

Now take a vector field $X = u(x,y,z)\,\frac{\partial}{\partial x} + v(x,y,z)\,\frac{\partial}{\partial y} + w(x,y,z)\,\frac{\partial}{\partial z}$ and rewrite it as a 2-form

$$\omega = u(x,y,z)\,dy\wedge dz + v(x,y,z)\,dz\wedge dx + w(x,y,z)\,dx\wedge dy,$$

so that

$$d\omega = \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}\right) dx\wedge dy\wedge dz.$$

Now the left side of Stokes' formula is $\int_c d\omega$, which is obviously

$$\int_c d\omega = \int_{B(0,1)} \operatorname{div} X\, dxdydz,$$

the left side of the Divergence Theorem. To evaluate the integral explicitly, we just have to compute $c^\# d\omega$, which by equation (17.2.6) is

$$c^\#(dx\wedge dy\wedge dz) = 2\pi^2 p^2 \sin(\pi q)\, dp\wedge dq\wedge dr.$$

(This is obviously how we would compute a volume integral over the ball: just change to spherical coordinates and use the spherical volume element $\rho^2\sin\theta$; the extra factor of $2\pi^2$ comes from the fact that we forced our integral to be over the unit cube, rather than a general rectangle, so we had to rescale the angles.)

The other part is more interesting. Just as in Example 17.2.6, where the boundary of the disc had four terms but three of them canceled out, here we will have a number of boundary terms but most will cancel out. The general formula for ∂c in (17.2.4) means we don't actually have to figure out how to integrate over a particular boundary (and we don't have to figure out the surface area element either). Just use the formula to get an automatic parametrization. We have

$$\begin{aligned}
c^\# dx &= d(x\circ c) = \frac{\partial x}{\partial p}\,dp + \frac{\partial x}{\partial q}\,dq + \frac{\partial x}{\partial r}\,dr \\
&= \sin(\pi q)\cos(2\pi r)\,dp + \pi p\cos(\pi q)\cos(2\pi r)\,dq - 2\pi p\sin(\pi q)\sin(2\pi r)\,dr, \\
c^\# dy &= \sin(\pi q)\sin(2\pi r)\,dp + \pi p\cos(\pi q)\sin(2\pi r)\,dq + 2\pi p\sin(\pi q)\cos(2\pi r)\,dr, \\
c^\# dz &= \cos(\pi q)\,dp - \pi p\sin(\pi q)\,dq.
\end{aligned}$$

Therefore we get

$$\begin{aligned}
c^\#(dy\wedge dz) &= c^\# dy \wedge c^\# dz \\
&= 2\pi^2 p^2 \sin^2(\pi q)\cos(2\pi r)\,dq\wedge dr - \pi p\sin(2\pi r)\,dp\wedge dq \\
&\qquad + 2\pi p\sin(\pi q)\cos(\pi q)\cos(2\pi r)\,dr\wedge dp \\
c^\#(dz\wedge dx) &= c^\# dz \wedge c^\# dx \\
&= 2\pi^2 p^2 \sin^2(\pi q)\sin(2\pi r)\,dq\wedge dr + \pi p\cos(2\pi r)\,dp\wedge dq \\
&\qquad + 2\pi p\sin(\pi q)\cos(\pi q)\sin(2\pi r)\,dr\wedge dp \\
c^\#(dx\wedge dy) &= c^\# dx \wedge c^\# dy \\
&= 2\pi^2 p^2 \sin(\pi q)\cos(\pi q)\,dq\wedge dr - 2\pi p\sin^2(\pi q)\,dr\wedge dp.
\end{aligned}$$

Now finally we want to compute the operations $(\iota_i^+)^\# c^\#$ based on these computations. We have for example that $\iota_2^+(s,t) = (s,1,t)$, so that

$$(\iota_2^+)^\#(dp \wedge dq) = 0, \quad (\iota_2^+)^\#(dq \wedge dr) = 0, \quad \text{and} \quad (\iota_2^+)^\#(dr \wedge dp) = -ds \wedge dt.$$

Therefore

$$\begin{aligned}
(\iota_2^+)^\# c^\#(dy \wedge dz) &= (\iota_2^+)^\#\big(2\pi p \sin(\pi q)\cos(\pi q)\cos(2\pi r)\, dr \wedge dp\big) \\
&= -\big(2\pi s \sin\pi \cos\pi \cos(2\pi t)\big)\, ds \wedge dt \\
&= 0, \\
(\iota_2^+)^\# c^\#(dz \wedge dx) &= (\iota_2^+)^\#\big(2\pi p \sin(\pi q)\cos(\pi q)\sin(2\pi r)\, dr \wedge dp\big) \\
&= -\big(2\pi s \sin\pi \cos\pi \sin(2\pi t)\big)\, ds \wedge dt \\
&= 0 \\
(\iota_2^+)^\# c^\#(dx \wedge dy) &= (\iota_2^+)^\#\big(-2\pi p \sin^2(\pi q)\, dr \wedge dp\big) \\
&= 2\pi s \sin^2\pi\, ds \wedge dt \\
&= 0.
\end{aligned}$$

Thus we get $(\iota_2^+)^\# c^\# \equiv 0$. Similarly we have $(\iota_2^-)^\# c^\# \equiv 0$.

For $\iota_3^\pm$, we have $\iota_3^\pm(s,t) = (s,t,\epsilon)$, where ϵ is either 0 or 1. Thus

$$(\iota_3^\pm)^\#(dp \wedge dq) = ds \wedge dt, \quad (\iota_2^\pm)^\#(dq \wedge dr) = 0, \quad \text{and} \quad (\iota_2^\pm)^\#(dr \wedge dp) = 0.$$

Thus

$$\begin{aligned}
(\iota_3^\pm)^\# c^\#(dy \wedge dz) &= 0, \\
(\iota_3^\pm)^\# c^\#(dz \wedge dx) &= \pi s\, ds \wedge dt, \\
(\iota_3^\pm)^\# c^\#(dx \wedge dy) &= 0.
\end{aligned}$$

As a result we get $(\iota_3^+)^\# - (\iota_3^-)^\# \equiv 0$, so that there is no contribution to the ∂c integral from ι_3.

Finally, it is easy to see that $(\iota_1^-)^\# \equiv 0$ since we are setting $p = 0$ everywhere, while

$$\begin{aligned}
(\iota_1^+)^\# c^\#(dy \wedge dz) &= 2\pi^2 \sin^2(\pi s)\cos(2\pi t)\, ds \wedge dt \\
(\iota_1^+)^\# c^\#(dz \wedge dx) &= 2\pi^2 \sin^2(\pi s)\sin(2\pi t)\, ds \wedge dt \\
(\iota_1^+)^\# c^\#(dx \wedge dy) &= 2\pi^2 \sin(\pi s)\cos(\pi s)\, ds \wedge dt.
\end{aligned}$$

Thus we have

$$\begin{aligned}
(\iota_1^+)^\# c^\#(\omega) = 2\pi^2 \sin(\pi s) \Big(&\sin(\pi s)\cos(2\pi t) u\big(c(1,s,t)\big) \\
&+ \sin(\pi s)\sin(2\pi t) v\big(c(1,s,t)\big) + \cos(\pi s) w\big(c(1,s,t)\big)\Big)\, ds \wedge dt \\
= 2\pi^2 \sin(\pi s)&(X \cdot \mathbf{n})\big(c(1,s,t)\big).
\end{aligned}$$

At long last, then we have

$$\int_{\partial c} \omega = \int_0^1 \int_0^1 2\pi^2 \sin(\pi s)(X \cdot \mathbf{n})\big(c(1,s,t)\big)\, ds dt = \int_{S(0,1)} (X \cdot \mathbf{n})\, dS.$$

Matching up the terms of Stokes' Theorem, we get the divergence theorem. ☺

17.4. **Stokes' Theorem in general.** Everything we have done so far relies on parametrizations of submanifolds, and although we have shown that the results we get don't actually depend on the choice of parametrization (as long as an orientation is respected), it is aesthetically less pleasing than simply being able to integrate a k-form over an arbitrary (oriented) k-dimensional submanifold. We were able to do such computations over spheres, for example, but only by parametrizing by squares (which ended up being singular although on a small set). It's not totally clear that we can parametrize arbitrary k-dimensional submanifolds by singular k-cubes which are diffeomorphisms except on a set of measure zero and fill up the entire submanifold—yet this is what we relied on to do all our practical computations.[36] We would like a more general definition, but this will rely on generalizing all our notions.

First is the question of what it means to integrate an n-form over an n-dimensional manifold. Immediately we run into a problem: in the previous case we parametrized our surfaces by cubes, which are compact, and hence all integrals were well-defined. On the other hand we can't in general even expect the integral of a general 1-form on $\mathbb{R}$ to be finite or well-defined. So we can't hope to make this work on every possible manifold. However on compact manifolds this is not a concern.

The one thing we are worried about is orientation. Without a fixed parametrization, we have an ambiguity when we try to integrate an n-form on an n-dimensional manifold: if the n-form is expressed in a coordinate chart as

$$\omega = f(x^1, \dots, x^n)\, dx^1 \wedge \cdots \wedge dx^n,$$

it is not clear whether the appropriate integral is $\int_I f\, dx^1 \cdots dx^n$ or $-\int_I f\, dx^1 \cdots dx^n$. Obviously one can make an arbitrary choice in each chart, but one wants to know that the manifold can actually be covered with charts for which the choices are the same. This relates to Definition 8.2.14.

Theorem 17.4.1. *Suppose M is a smooth compact n-dimensional manifold. Then the bundle $\Omega^n(M)$ is trivial if and only if M is orientable in the sense of Definition 8.2.14. A nowhere-zero section μ of $\Omega^n(M)$ is called an* orientation *of M or a* volume form*, and any two volume forms μ_1 and μ_2 are related by $\mu_2 = f\,\mu_1$ for some smooth nowhere-zero function f. Hence on a connected manifold, there are only two equivalence classes of orientation.*

Proof. The bundle is constructed in the same standard way we used for the tangent bundle TM in Definition 12.1.4 and the cotangent bundle T^*M in Definition 15.2.1. Namely, given a coordinate chart $(\phi = \mathbf{x}, U)$ on M, we observe that any $\mu_p \in \Omega^n(T_pM)$ can be written as $\mu_p = a\, dx^1|_p \wedge \cdots \wedge dx^n|_p$ for some $a \in \mathbb{R}$ since $\Omega^n(T_pM)$ is 1-dimensional by Proposition 4.3.8. The coordinate chart induced on $\Omega^n(M)$ is then

$$\overline{\Phi}\colon \mu \in \Omega^n(T_pM) \mapsto (x^1, \dots, x^n, c).$$

By the transition formula Proposition 4.3.10 (or just by computing directly using antisymmetry of the wedge product and the definition of the determinant), we know if there are two coordinate charts $(\phi = \mathbf{x}, U)$ and $(\psi = \mathbf{y}, V)$ near p, and if

$$\mu_p = a\, dx^1|_p \wedge \cdots \wedge dx^n|_p = b\, dy^1|_p \wedge \cdots \wedge dy^n|_p,$$

[36] You really can prove that you can do all necessary computations using k-chains, but that actually requires more work than what we're going to do.

then we must have $b = a \cdot \det\left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right)\Big|_{\mathbf{y}(p)}$. We conclude that coordinate transitions are smooth, and defining open sets on $\Omega^n(M)$ to be those which get mapped to open subsets of $\mathbb{R}^{n+1}$ by every coordinate chart, we get a topology which makes $\Omega^n(M)$ a smooth manifold and smooth bundle over M.

Suppose this bundle is trivial. Then there is a nowhere-zero section, which we call μ. Given a coordinate chart $(\mathbf{x}, U)$, we have $\mu|_U = f(x^1, \ldots, x^n)\, dx^1 \wedge \cdots \wedge dx^n$ for some function f which is nowhere-zero on U, and hence it is either always positive or always negative. Define the chart to be positively-oriented if $f > 0$ everywhere on U. Clearly if the chart is not positively-oriented, we can make a trivial modification (such as switching x^1 and x^2, or replacing x^1 with $-x^1$) which will make it positively-oriented. So there is an atlas of positively-oriented charts. Now for any two charts $\mathbf{x}$ and $\mathbf{y}$ in this atlas, we have $\mu = f(\mathbf{x})\, dx^1 \wedge \cdots \wedge dx^n$ and $\mu = g(\mathbf{y})\, dy^1 \wedge \cdots \wedge dy^n$ where both f and g are positive. Since $f(\mathbf{x}) = g(\mathbf{y}) \det\left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right)$ we conclude that the determinant of the coordinate transition matrix is positive, which is exactly Definition 8.2.14.

Conversely suppose M is orientable, and pick finitely-many charts (ϕ_i, U_i) which cover M and are compatible in the sense that the determinant $\det D(\phi_i \circ \phi_j^{-1}) > 0$ for all i and j. Choose a partition of unity ξ_i with $\operatorname{supp} \xi_i \subset U_i$ for each i. Let $\mu_i = \xi_i \cdot dx^1 \wedge \cdots \wedge dx^n$, defined a priori on U_i but extended smoothly to all of M. Define $\mu = \sum_i \mu_i$. This is a smooth section of $\Omega^n(M)$, and we want to prove it's nowhere-zero. To do this, pick any point $p \in M$ and some i such that $\xi(p) > 0$. Then $\mu_i(p) = dx_i^1|_p \wedge \cdots dx_i^n|_p$, and for any other j with $\xi_j(p) > 0$ we have

$$\mu_j(p) = dx_j^1|_p \wedge \cdots \wedge dx_j^n|_p = \det\left(\frac{\partial \mathbf{x}_j}{\partial \mathbf{x}_i}\right) dx_i^1|_p \wedge \cdots \wedge dx_i^n|_p.$$

Thus

$$\mu(p) = \sum_j \xi_j(p) \det\left(\frac{\partial \mathbf{x}_j}{\partial \mathbf{x}_i}\right) dx_i^1|_p \wedge \cdots \wedge dx_i^n|_p,$$

where each term is interpreted as zero if $\xi_j(p) = 0$. Since all the determinants are positive and $\xi_j(p) \geq 0$ for all j, the coefficient is a sum of nonnegative terms with at least one positive term, and hence positive. In particular it is not zero at p.

The remaining facts are obvious from the fact that the space of n-forms at each point is 1-dimensional, and hence any volume form spans at each point. The fact that the ratio of two volume forms is smooth follows from the quotient rule applied in any coordinate chart. □

Given a choice of orientation on an n-dimensional manifold M, we can specify preferred parametrizations, and once we do that, we can integrate any n-form which is supported in a single coordinate chart in a unique way: we just use Definition 17.2.2 to define the integral (using the inverse of the coordinate chart as a parametrization), and we know it does not depend on the choice of coordinate chart or parametrization by Lemma 17.2.8. Naturally, to define the integral globally on the entire manifold, we just use a partition of unity; the fact that the choice of coordinates does not matter suggests the choice of partition of unity does not matter either.

Definition 17.4.2. Suppose M is a compact orientable n-dimensional manifold with a specified choice of orientation, and let μ be any smooth n-form on M. Let (ϕ_j, U_j) be a covering of M by finitely many coordinate charts for $1 \leq j \leq m$

which are positively-oriented with respect to the chosen orientation. Let $\{\xi_j\}$ be a partition of unity subordinate to $\{U_j\}$, so that $D_j := \operatorname{supp}\xi_j \subset U_j$. Then each $\phi_j[D_j]$ is a compact subset of $\mathbb{R}^n$ contained in a cube of some large radius R_j, and by rescaling the coordinate chart we can assume $\phi_j[D_j]$ is contained in the unit cube I^n.

Let $\eta_j \colon \mathbb{R}^n \to M$ be the inverse of ϕ_j, and define

$$\int_M \mu = \sum_{j=1}^m \int_{I^n} (\xi_j \circ \eta_j)(\eta_j)^\# \mu, \tag{17.4.1}$$

where the integral over the n-cube is defined as in Definition 17.2.2.

Obviously for this to give us what we want, we need to check the invariance property.

Proposition 17.4.3. *If M and μ are as in Definition 17.4.2, then the number $\int_M \mu$ does not depend on the choice of positively-oriented coordinate charts or the partition of unity.*

Proof. Each term in the sum (17.4.1) is $\int_{\eta_j} \xi_j\mu$ by Definition 17.2.2. Consider another collection of charts (ψ_i, V_i) for $1 \leq i \leq m'$ with corresponding parametrizations $\zeta_i \colon I^n \to V_i \subset M$ given by $\zeta_i = \psi_i^{-1}$ and a partition of unity χ_i with $\operatorname{supp}\chi_i \subset \xi_i[I^n]$. Write

$$\int_M \mu = \sum_{j=1}^m \int_{\eta_j} \xi_j\mu = \sum_{j=1}^m \sum_{i=1}^{m'} \int_{\eta_j} \chi_i \cdot \xi_j \cdot \mu.$$

Suppose for some i and j that $\operatorname{supp}(\xi_j) \cap \operatorname{supp}(\chi_i) \neq \emptyset$. Then the corresponding coordinate charts overlap: let $\upsilon = \phi_j \circ \psi_i^{-1}$ be the smooth transition map. Since $\zeta_i = \psi_i^{-1}$ and $\eta_j = \phi_j^{-1}$, we have $\eta_j = \zeta_i \circ \upsilon^{-1}$. Let $D_1 = \phi_j\big[\operatorname{supp}(\xi_j) \cap \operatorname{supp}(\chi_i)\big]$ and let $D_2 = \psi_i\big[\operatorname{supp}(\xi_j) \cap \operatorname{supp}(\chi_i)\big]$; then both D_1 and D_2 are subsets of I^n and are the closure of some open set. We can write

$$\int_{\eta_j} \chi_i{\cdot}\xi_j{\cdot}\mu = \int_{D_1} \eta_j^\#\big(\chi_i{\cdot}\xi_j{\cdot}\mu\big) = \int_{D_1} (\upsilon^{-1})^\# \zeta_i^\#\big(\chi_i{\cdot}\xi_j{\cdot}\mu\big) = \int_{D_2} \zeta_i^\#\big(\chi_i{\cdot}\xi_j{\cdot}\mu\big) = \int_{\zeta_i} \chi_i\xi_j\mu,$$

where we used the fact that the υ^{-1} is a diffeomorphism from D_1 to D_2 to change variables in the integral. We thus have

$$\int_M \mu = \sum_i \sum_j \int_{\eta_j} \chi_i \cdot \xi_j \cdot \mu = \sum_i \sum_j \int_{\zeta_i} \chi_i \cdot \xi_j \cdot \mu = \sum_i \int_{\zeta_i} \chi_i \cdot \mu.$$

□

Now that we know how to integrate n-forms on n-dimensional manifolds in terms of integrations of n-cubes as in Definition 17.2.2, we can prove Stokes' Theorem on a smooth manifold using Stokes' Theorem 17.3.1 for n-chains.

Theorem 17.4.4. *Suppose M is a smooth oriented compact n-dimensional manifold and that ω is a smooth $(n-1)$-form on M. Then*

$$\int_M d\omega = 0. \tag{17.4.2}$$

Proof. Take a covering by finitely many charts with corresponding partition of unity as in Definition 17.4.2. Then we can write $\int_M d\omega = \sum_j \int_M \xi_j d\omega$. However by Proposition 16.3.6, the product rule for differential forms, we have $\xi_j d\omega = d(\xi_j\omega) - d\xi_j \wedge \omega$, so that

$$\int_M d\omega = \sum_j \int_M d(\xi_j\omega) - \int_M (\sum_j d\xi_j) \wedge \omega.$$

However since $\sum_j \xi_j \equiv 1$, we know $\sum_j d\xi_j \equiv 0$. Furthermore we have

$$\int_M d(\xi_j\omega) = \int_{\eta_j} d(\xi_j\omega) = \int_{\partial\eta_j} \xi_j\omega$$

by Stokes' Theorem 17.3.1 for n-chains. But since the support of $\xi_j \circ \eta_j$ is in contained in the unit cube, we know ξ_j is zero on every portion of the boundary $\partial\eta_j$. Thus $\int_{\partial\eta_j} \xi_j\omega = 0$. We conclude $\int_M d\omega = 0$. □

This version of Stokes' Theorem for smooth manifolds is the one most commonly used in applications. The idea is that a smooth manifold has no boundary, and thus Stokes' Theorem in the form $\int_M d\omega = \int_{\partial M} \omega$ would imply the right side is zero since $\partial M = \emptyset$. However for some applications we actually want to allow the manifold to have a nonempty boundary ∂M, which means we have to define what this means. We will also have to figure out the relationship between the orientation on M and the orientation on ∂M, since of course we cannot integrate forms without an orientation. We won't get too far into the details of this construction, but we will sketch the basic ideas.

Definition 17.4.5. A topological space M is called an *n-dimensional topological manifold with boundary* if for every $p \in M$ there is a neighborhood U of p and a map $\phi\colon U \to \mathbb{R}^n$ such that either ϕ is a homeomorphism onto $\mathbb{R}^n$ or ϕ is a homeomorphism onto the closed half-plane $H^n = \mathbb{R}^{n-1} \times [0, \infty)$.

Points of the first type are called *interior points*; those which are not of the first type are called *boundary points*, and the set of all boundary points is called the *boundary of M* and denoted by ∂M.

Proposition 17.4.6. *If M is a topological manifold with boundary, then the boundary ∂M is a topological manifold in the subspace topology.*

Proof. Suppose $p \in M$ is a boundary point, and let (ϕ, U) be a coordinate chart with $\phi\colon U \to H^n$ a homeomorphism. Since p is not an interior point, $\phi(p)$ must actually be on the boundary $\mathbb{R}^{n-1}\times\{0\}$, because if the last component were positive, there would be a neighborhood $V = \phi^{-1}\big(\mathbb{R}^{n-1} \times (0, \infty)\big)$ containing p, and since $\mathbb{R}^{n-1} \times (0, \infty)$ is homeomorphic to $\mathbb{R}^n$, we could find a homeomorphism from V to $\mathbb{R}^n$, which makes p an interior point.

I claim that if p is a boundary point, then for any chart (ϕ, U) with $p \in U$, we must have $\phi|_{\partial M} \subset \mathbb{R}^{n-1} \times \{0\}$. Suppose not; then there is some $q \in \partial M$ with $\phi(q) \in \mathbb{R}^{n-1} \times (0, \infty)$, and as before we can obtain a chart around q making q an interior point, a contradiction. Conversely if $\phi(q) \in \mathbb{R}^{n-1} \times \{0\}$ for some $q \in M$, I claim that $q \in \partial M$. If there were some other chart (ψ, V) with $q \in V$ and ψ a homeomorphism from V to $\mathbb{R}^n$, then $\phi \circ \psi^{-1}$ would be a homeomorphism from some open subset of $\mathbb{R}^n$ to some open subset of H^n containing the boundary. The contradiction here is a bit more complicated, but essentially relies on the fact from

algebraic topology that $\mathbb{R}^n$ and H^n are not homeomorphic: deleting the origin from the half-plane leaves a space that is contractible, while $\mathbb{R}^n$ minus a point is not contractible.

Hence for any point $p \in \partial M$, with chart (ϕ, U) on M, there is a chart $(\tilde{\phi}, \tilde{U})$ on ∂M given by $\tilde{U} = U \cap \partial M$ and $\tilde{\phi} = \phi|_{\partial M}$. The map $\tilde{\phi}$ is a bijection from $U \cap \partial M$ to $\mathbb{R}^{n-1} \times \{0\}$, and since ϕ is a homeomorphism, so is $\tilde{\phi}$. Identifying $\mathbb{R}^{n-1} \times \{0\}$ with $\mathbb{R}^{n-1}$, these are all coordinate charts on ∂M. □

Proposition 17.4.6 is often summarized as "the boundary of a boundary is zero," and this fact is analogous to the fact alluded to in Remark 17.2.7, that $\partial(\partial c) = 0$ for any k-chain c. If M happens to be the parametrized image of some n-cube, these are essentially the same thing.

We define a smooth manifold with boundary by the condition that all the transition maps are C^∞. All our transition maps will be defined either on a subset of H^n or $\mathbb{R}^n$, and map into either a subset of H^n or of $\mathbb{R}^n$. We already know what to do if the maps go from a subset of $\mathbb{R}^n$ to $\mathbb{R}^n$. If a transition map is defined only on an open subset of H^n, we say it is smooth if we can extend it to a smooth map on an open set containing H^n. This is not the only possible definition, but what's nice is that it allows us to take derivatives in all directions even up to the boundary (so that for example we can consider tangent vectors that stick out of the boundary). Most of the concepts we have already studied can be extended without difficulty to manifolds with boundary.

For example, as in Definition 9.1.9 a k-dimensional *smooth submanifold* N of a smooth manifold with boundary M is a subset which is a manifold in the subspace topology and such that for every $p \in N$, there is a chart (ϕ, U) on M such that $U \cap N = \phi^{-1}\big[\mathbb{R}^k \times \{0, \ldots, 0\}\big]$. It is easy to see from this definition that ∂M is an $(n-1)$-dimensional smooth submanifold of M, directly from the definition of the coordinate charts on a manifold with boundary.

We can also extend the notion of orientabilty from Definition 8.2.14: a smooth manifold with boundary is orientable if there is an atlas of charts $\{(\phi_i, U_i)\}$ such that $D(\phi_i \circ \phi_j^{-1}) > 0$ everywhere (where this map is defined either on $\mathbb{R}^n$ or on H^n). We now want to see how to obtain an orientation on the boundary of a manifold.

Definition 17.4.7. Let M be an n-dimensional smooth manifold with boundary. Let $p \in \partial M$, and let $v \in T_pM$. The vector v is called *tangent to the boundary* if there is a smooth curve $\gamma\colon (-\varepsilon, \varepsilon) \to M$ such that $\gamma(t) \in \partial M$ for all t, with $\gamma(0) = p$ and $\gamma'(0) = v$. The vector v is *inward-pointing* if it is not tangent to the boundary and there is a smooth curve $\gamma\colon [0, \varepsilon) \to M$ such that $\gamma(0) = p$ and $\gamma'(0) = v$. v is *outward-pointing* if $-v$ is inward-pointing.

In a coordinate neighborhood (ϕ, U) of $p \in \partial M$ such that $\phi[U] = \mathbb{R}^{n-1} \times [0, \infty)$, it is easy to see that a vector $v = \sum_{k=1}^n a^k \frac{\partial}{\partial x^k}\big|_p$ is tangent to the boundary iff $a^n = 0$, inward-pointing iff $a^n > 0$, and outward-pointing iff $a^n < 0$.

Proposition 17.4.8. *If M is a smooth manifold with boundary, then there is a smooth map $X\colon \partial M \to TM$ such that $X_p \in T_pM$ for every $p \in \partial M$ and such that X_p is outward-pointing at every p. We say that X is an* outward-pointing vector field along the boundary*.*

Proof. Since ∂M is a smooth submanifold of M, it is itself a smooth manifold. Hence it has a smooth partition of unity $\{\xi_i\}$ subordinate to some coordinate charts

$\{(\tilde{\phi}_i, \tilde{U}_i)\}$ on ∂M, each of which comes from a coordinate chart (ϕ_i, U_i) on M. In each coordinate chart U_i, define $X_i = -\xi_i \cdot \frac{\partial}{\partial x^n}$, and define $X = \sum_i X_i$. Since each ξ_i is nonnegative and each X_i is outward-pointing everywhere, we know that X is either outward-pointing or tangent to the boundary everywhere. Since for any point at least one ξ_i is strictly positive, we know X is actually outward-pointing at every point. □

Given a volume form μ on a smooth manifold with boundary M as in Theorem 17.4.1 and an outward-pointing vector field X on ∂M, we define the *induced orientation on* ∂M to be the $(n-1)$-form $\iota_X \mu$ given at each $p \in \partial M$ by

$$\iota_X \mu(v_1, \ldots, v_{n-1}) = \mu(X_p, v_1, \ldots, v_{n-1})$$

for any vectors $v_1, \ldots, v_{n-1} \in T_p \partial M$. In coordinates X_p has nonzero n^{th} component while all of the vectors v_k have zero n^{th} component, and thus it is easy to see that $\iota_X \mu$ is a nowhere-zero $(n-1)$-form on ∂M. As a consequence, the boundary of any orientable manifold with boundary is an orientable manifold. For example, this shows that any smooth compact surface which embeds in $\mathbb{R}^3$ must be orientable, and explains why the projective plane and Klein bottle cannot be embedded in $\mathbb{R}^3$.

We need one more thing: partitions of unity. We obtain them on manifolds with boundary just as we did for manifolds in Theorem 13.3.2: take bump functions on $\mathbb{R}^n$ and restrict them to H^n for those coordinate charts that include boundary points. Notice that a partition of unity on M restricts to a partition of unity on ∂M which is subordinate to the corresponding charts on ∂M.

Having set all this up, it now makes sense to ask whether $\int_M d\omega = \int_{\partial M} \omega$ for an $(n-1)$-form ω defined on a smooth orientable manifold with boundary M: we interpret the integral on the right as the integral of the pull-back $\iota^{\#}\omega$, where $\iota \colon \partial M \to M$ is the inclusion map, which is smooth since ∂M is a smooth submanifold. Making sense of all this is the hard part; actually proving the theorem is easy.

Theorem 17.4.9. *Suppose M is a smooth n-dimensional orientable compact manifold with boundary, and let ω be a smooth $(n-1)$-form on M. Then $\int_M d\omega = \int_{\partial M} \omega$.*

Proof. Take finitely many charts (ϕ_i, U_i) on M and construct a partition of unity ξ_i subordinate to them. Then we have

$$\int_M d\omega = \sum_i \int_M \xi_i d\omega = \sum_i \int_M d(\xi_i \omega) - \int_M (\sum_i d\xi_i) \wedge \omega.$$

Just as in Theorem 17.4.4, we have $\sum_i d\xi_i = 0$ so this last integral vanishes. Furthermore if $\eta_i = \phi_i^{-1}$ is the induced parametrization, we have

$$\int_M d(\xi_i \omega) = \int_{\eta_i} d(\xi_i \omega) = \int_{\partial \eta_i} \xi_i \omega.$$

Now η_i maps the unit cube to the manifold M. If it maps into the interior of the manifold, then $\int_{\eta_i} d(\xi_i \omega) = 0$ exactly as in the proof of Theorem 17.4.4; however if it comes from a boundary chart, then since the charts ϕ_i map $U_i \cap \partial M$ to $\mathbb{R}^{n-1} \times \{0\}$, we know that $\eta_i = \phi_i^{-1}$ maps one face of the unit cube to the boundary (and all other faces to the interior of the manifold). Thus the integral $\int_{\partial \eta_i} \xi_i \omega$ is equal to

$\int_{\tilde{\eta}_i} \tilde{\xi}_i \omega$ where $\tilde{\xi}_i$ is the restricted partition of unity function and $\tilde{\eta}_i$ is the inverse of the restricted coordinate chart $\tilde{\phi}_i$. We thus obtain

$$\int_M d\omega = \sum_i \int_{\eta_i} d(\xi_i \omega) = \sum_i \int_{\tilde{\eta}_i} \tilde{\xi}_i \omega = \int_{\partial M} \omega.$$

□

Stokes' Theorem is a beautiful theorem in its own right, and has a variety of applications (as you have probably seen in vector calculus or physics), but we will most often use it to relate the boundary operator ∂ to the exterior derivative operator d, and thus to relate homology to cohomology. See Chapter 18 for details.

18. De Rham Cohomology

"Around the survivors a perimeter create!"

Let's review the story so far. We defined vectors $v \in T_pM$ in terms of the operation $f \mapsto v(f) \in \mathbb{R}$ on smooth functions f defined near p. From there we assembled all the tangent spaces to build vector fields $X \in \chi(M)$ which differentiate functions globally, giving an operation $f \mapsto X(f) \in \mathcal{F}(M)$ for each $f \in \mathcal{F}(M)$. Then we changed perspective and viewed $X(f) = df(X)$ not as a map on $\mathcal{F}(M)$ but as a map on $\chi(M)$ given by $X \mapsto df(X)$. The object df became a section of the cotangent bundle or a 1-form, but we saw there were many 1-forms $\omega \in \Omega^1(M)$ which were not df for any function f.

Trying to determine whether a given ω was equal to df led to the construction of $d\omega$; we saw that if $\omega = df$ then $d\omega = 0$ by Proposition 16.3.5, a consequence of commuting mixed partials and antisymmetry of 2-forms. We then asked about the converse: if $d\omega = 0$, is there a function f such that $\omega = df$? If $M = \mathbb{R}^2$ we saw in Proposition 15.2.10 (the Poincaré Lemma) that the converse is true; on the other hand if $M = \mathbb{R}^2 \backslash \{\mathbf{0}\}$, we saw in Example 17.1.9 that there was at least one 1-form ω such that $d\omega = 0$ but ω is not df for any function f. We extended the exterior derivative d to k-forms of any order, and Proposition 16.3.5 shows that a *necessary* condition to have a given k-form ω be the exterior derivative of something is that $d\omega = 0$. However we did not yet get a *sufficient* condition.

We also reinterpreted k-forms: although we started with k-forms as certain tensorial operators, we saw by combining the Change of Variables Theorem 5.3.2 with Lemma 17.2.8 that k-forms change coordinates in exactly the same way as k-dimensional integrals (up to a possible change of sign), and thus it made sense to define the integral of a k-form over a parametrized k-dimensional surface (a k-cube) as in Definition 17.2.2. This quantity depends only on the surface and its orientation, but not on the choice of parametrization, and thus we were able to define a notion of integration of an n-form on an n-dimensional manifold in Definition 17.4.2. We extended this to manifolds with boundary as well.

Finally we realized that it was possible to define a boundary ∂c of a parametrized k-dimensional manifold $c[I^k]$ in such a way that we would have Stokes' Theorem $\int_c d\omega = \int_{\partial c} \omega$ for any $(k-1)$-form ω. We showed in Example 17.2.7 that for a 2-cube that $\partial(\partial c) = 0$ and hinted that this was true in general. We generalized this notion to define the boundary of an unparametrized manifold in Definition 17.4.5, and we showed in Theorem 17.4.9 that $\int_M d\omega = \int_{\partial M} \omega$ for any $(n-1)$-form ω. Hence even if we didn't know $\partial \circ \partial = 0$, for every k-cube c we would have

$$\int_{\partial(\partial c)} \omega = \int_{\partial c} d\omega = \int_c d(d\omega) = 0$$

for any $(k-2)$-form ω, which implies $\partial(\partial c)$ is at least equivalent to zero. We will prove that it is actually equal to zero, and then we can ask the question of whether $\partial c = 0$ for a k-chain c implies that $c = \partial b$ for some $(k-1)$-chain b.

We thus have two questions: is $\operatorname{im} d = \ker d$, and is $\operatorname{im} \partial = \ker \partial$? The first question is smooth cohomology, while the second question is smooth homology, and Stokes' Theorem implies the duality between these concepts (hence the "co"). Now homology and cohomology will turn out to be diffeomorphism invariants basically

because of Stokes' Theorem and the formula $\eta^\# \circ d = d \circ \eta^\#$. This is useful since as we have seen, there may be many very different ways of representing the same manifold, and determining whether two manifolds are diffeomorphic is not necessarily easy. Computing the homology and cohomology is one way of doing this. It will actually turn out that homology and cohomology are in fact smooth homotopy invariants as well (so that for example the cohomology of the sphere minus two points is the same as the cohomology of the circle). And from this it will even turn out that cohomology is a homeomorphism invariant, which is de Rham's very surprising theorem. It implies that in spite of needing derivatives and *smooth* functions to define the exterior derivative, the cohomology really only depends on the topological structure and the *continuous* functions.

For some background on the history of this topic, I highly recommend the following:

- `http://www.maths.ed.ac.uk/~aar/papers/poincare2009.pdf` John Stillwell's translation of the paper "Analysis situs" by Henri Poincaré; the first 12 pages include a sweeping tour of the history of homology before and after Poincaré invented it.
- `http://books.google.com/books?id=7iRijkz0rrUC` Victor Katz's article on "Differential forms" in *History of Topology*, which traces cohomology all the way back to Cauchy's integral formula.
- `http://www.cs.sjsu.edu/faculty/beeson/courses/Ma213/HistoryOfForms.pdf` Hans Samelson's history of differential forms up to de Rham's theorem, in American Mathematical Monthly.

18.1. **The basic cohomology groups.** For any $k \geq 1$, the exterior derivative d takes k-forms on a smooth manifold to $(k+1)$-forms. In this Chapter we will sometimes denote this operation by $d_k \colon \Omega^k(M) \to \Omega^{k-1}(M)$; keeping track of the index k helps to avoid confusion since we will be dealing with $(k+1)$-forms, k-forms, and $(k-1)$-forms simultaneously.

Definition 18.1.1. A k-form ω is called *closed* if $d\omega = 0$. A k-form ω is called *exact* if $\omega = d\alpha$ for some $(k-1)$-form α. Since d is linear, the spaces of closed and exact forms are vector subspaces. We denote the closed k-forms by $Z^k(M) \subset \Omega^k(M)$ and the exact k-forms by $B^k(M) \subset \Omega^k(M)$. We have $Z^k(M) = \ker d_k$ and $B^k(M) = \operatorname{im} d_{k-1}$.

Since $d_k \circ d_{k-1} = 0$, we know that $B^k(M)$ is a subspace of $Z^k(M)$, and the quotient vector space is denoted by $H^k(M) = Z^k(M)/B^k(M)$, and called the k-dimensional de Rham cohomology space. If $k = 0$ then we say $B^0(M)$ is the trivial vector space and $Z^0(M) = H^0(M)$.

We have already done a couple of examples of cohomology without saying so: now let's say so.

Example 18.1.2. If $M = \mathbb{R}^2$, then we can write any 1-form as $\omega = h(x,y)\,dx + j(x,y)\,dy$. Then $d\omega = \left(\dfrac{\partial j}{\partial x} - \dfrac{\partial h}{\partial y}\right) dx \wedge dy$, and if $d\omega = 0$ then $j_x = h_y$. By Proposition 15.2.10, we know that if $d\omega = 0$ then there is a function f such that $\omega = df$. In fact we can compute f explicitly, for example using the formula

$$f(x,y) = \int_0^x h(\sigma, y)\,d\sigma + \int_0^y j(0,\tau)\,d\tau. \tag{18.1.1}$$

Hence $Z^1(\mathbb{R}^2) = B^1(\mathbb{R}^2)$ and $H^1(\mathbb{R}^2) = \{0\}$.

If $M = \mathbb{R}^2 \backslash \{0\}$ then the 1-form $\omega = -\dfrac{y}{x^2+y^2}\,dx + \dfrac{x}{x^2+y^2}\,dy$ satisfies $d\omega = 0$, but we saw in Example 17.1.9 that ω cannot be df for any function f. The reason is that if $\gamma\colon [0,1] \to M$ is given by $\gamma(t) = \big(\cos(2\pi t), \sin(2\pi t)\big)$, then $\int_\gamma df = 0$ for any function f by the one-dimensional Stokes' Theorem Proposition 17.1.6, while $\int_\gamma \omega = 2\pi$. We thus have at least one element of $Z^1(M)$ which is not an element of $B^1(M)$, and therefore the de Rham cohomology $H^1(M)$ is a vector space of dimension at least one. (Later we will show that the dimension is exactly one.) ☺

As we see in Example 18.1.2, clever use of Stokes' Theorem can often make cohomology proofs simpler. Notice that I had to prove nonexistence of any function f such that $\omega = df$ (which is hard), and I did it by proving existence of a single curve γ with $\int_\gamma \omega \neq 0$ (which is easy). There are a few cases where one can easily compute the cohomology without any additional tools. The easiest cohomology spaces are the lowest- and highest-dimensional spaces.

Theorem 18.1.3. *Suppose M is a smooth n-dimensional manifold with q connected components. Then $H^0(M) \cong \mathbb{R}^q$.*

Proof. By definition we have $H^0(M) = Z^0(M)$, the space of all closed 0-forms on M. Since 0-forms are functions, we want to characterize smooth functions $f\colon M \to \mathbb{R}$ such that $df = 0$. In any coordinate chart we have $\frac{\partial f}{\partial x^i}\,dx^i = 0$, so that $\partial f/\partial x^i = 0$ for every i on $\mathbb{R}^n$. We conclude that f must be constant in any chart, and on overlapping charts it must be the *same* constant. Hence f is constant on each component. However on different components it may be different constants. Writing $M = \cup_{j=1}^q M_j$ where each M_q is connected, we can construct a basis $\{f_j\}$ of $H^0(M)$ where

$$f_j(p) = \begin{cases} 1 & p \in M_j, \\ 0 & p \notin M_j. \end{cases}$$

□

Alternatively we could have used the fact that smooth connected manifolds are also path-connected via piecewise-smooth paths to show that $f(p_2) - f(p_1) = \int_\gamma df = 0$ for any points p_1 and p_2 in the same component, in order to get a coordinate-independent proof.

To compute the top-dimensional cohomology we use Stokes' Theorem. Hence we require compactness and orientability. For the moment we can only show its dimension is positive; later we will show it's actually one-dimensional.

Theorem 18.1.4. *Suppose M is a compact n-dimensional orientable manifold. Then $\dim H^n(M) > 0$.*

Proof. Let μ be a volume form as in Theorem 17.4.1. Then since μ is an n-dimensional form and every $(n+1)$-dimensional form on an n-dimensional manifold must be zero, we know that $d\mu = 0$ so that $\mu \in Z^n(M)$.

If $\mu = d\omega$ for some $(n-1)$-form ω, then by Stokes' Theorem 17.4.4 we would have $\int_M \mu = \int_M d\omega = 0$. However by Definition 17.4.2, the integral is a finite sum of integrals over n-cubes of the form $\int_{\eta_j} \xi_j \mu$ where ξ_j is a partition of unity and each η_j is the inverse of a coordinate chart which is compatible with the orientation μ. Hence in each chart we have $\mu = f(x^1, \ldots, x^n)\,dx^1 \wedge \cdots \wedge dx^n$ where f is a

positive function, and thus each $\int_{\eta_j} \xi_j \mu$ is a positive number because ξ_j is positive on some open set. So $\int_M \mu \neq 0$, and thus μ cannot be $d\omega$ for any $(n-1)$-form ω. So $\mu \notin B^n(M)$. □

We also have the following trivial result on triviality.

Proposition 18.1.5. *If M is an n-dimensional manifold and $k > n$, then $H^k(M)$ is trivial.*

Proof. Every k-form is zero if $k > n$, and hence $\Omega^k(M)$ is already trivial. Thus so are its subspaces $Z^k(M)$ and $B^k(M)$, and thus so is the quotient $H^k(M)$. □

We can compute cohomology spaces in $\mathbb{R}^n$. In general we can prove that $H^k(\mathbb{R}^n)$ for any $k > 0$, but the general case is a mess in coordinates mainly because of the notation for the general k-form $dx^{i_1} \wedge \cdots \wedge dx^{i_k}$ where $i_1 < \cdots < i_k$. We'll do the easiest nontrivial case to try and illuminate how the general case works.

Theorem 18.1.6. *The cohomology space $H^1(\mathbb{R}^n)$ is trivial for any $n \geq 2$.*

Proof. Suppose ω is a closed 1-form. We will construct a formula for a function f such that $df = \omega$; the method is exactly the same as the special case (18.1.1) for $n = 2$. Write $\omega = \sum_{k=1}^n \omega_k(x^1, \ldots, x^n)\, dx^k$; then $d\omega = 0$ implies that

$$\frac{\partial \omega_k}{\partial x^j} = \frac{\partial \omega_j}{\partial x^k}$$

by formula (15.4.2). Define

$$f(x^1, \ldots, x^n) = \int_0^{x^1} \omega_1(t, x^2, \ldots, x^n)\, dt + \int_0^{x^2} \omega_2(0, t, x^3, \ldots, x^n)\, dt + \cdots + \int_0^{x^n} \omega_n(0, 0, \ldots, t)\, dt.$$

Clearly $\frac{\partial f}{\partial x^1} = \omega_1(x^1, \dots, x^n)$, and we have

$$\begin{aligned}
\frac{\partial f}{\partial x^k}(x^1, \dots, x^n) &= \sum_{j=1}^{k-1} \frac{\partial}{\partial x^k} \int_0^{x^j} \omega_j(0, \dots, 0, t, x^{j+1}, \dots, x^k, \dots, x^n)\, dt \\
&\qquad + \frac{\partial}{\partial x^k} \int_0^{x^k} \omega_k(0, \dots, 0, t, x^{k+1}, \dots, x^n)\, dt \\
&= \sum_{j=1}^{k-1} \int_0^{x^j} \frac{\partial \omega_j}{\partial x^k}(0, \dots, 0, t, x^{j+1}, \dots, x^k, \dots, x^n)\, dt \\
&\qquad + \omega_k(0, \dots, 0, x^k, x^{k+1}, \dots, x^n) \\
&= \sum_{j=1}^{k-1} \int_0^{x^j} \frac{\partial \omega_k}{\partial t}(0, \dots, 0, t, x^{j+1}, \dots, x^k, \dots, x^n)\, dt \\
&\qquad + \omega_k(0, \dots, 0, x^k, x^{k+1}, \dots, x^n) \\
&= \sum_{j=1}^{k-1} \Big(\omega_k(0, \dots, 0, x^j, x^{j+1}, \dots, x^n) - \omega_k(0, \dots, 0, 0, x^{j+1}, \dots, x^n)\Big) \\
&\qquad + \omega_k(0, \dots, 0, x^k, \dots, x^n) \\
&= \omega_k(x^1, \dots, x^n).
\end{aligned}$$

Putting it together, we see that $df = \omega$, as desired. □

We know the cohomology of the simplest 1-dimensional manifold: $H^k(\mathbb{R})$ is either $\mathbb{R}$ if $k = 0$ or $\{0\}$ if $k > 0$ by Theorem 18.1.6. Next let's compute the cohomology of S^1.

Theorem 18.1.7. *The de Rham cohomology of the circle is*

$$H^k(S^1) \cong \begin{cases} \mathbb{R} & k = 0, 1 \\ \{0\} & k \geq 2. \end{cases}$$

Proof. We already know $H^0(S^1) = \mathbb{R}$ since S^1 is connected, and that $H^k(S^1)$ is trivial for $k \geq 2$ since all k-forms are trivial on a 1-dimensional manifold. We just have to prove that $H^1(S^1)$ is one-dimensional.

Now the circle is diffeomorphic to $\mathbb{R}/\mathbb{Z}$. Let $\pi\colon \mathbb{R} \to S^1$ denote the quotient map. For any closed 1-form $\alpha \in Z^1(S^1)$, we know $\pi^\# \alpha$ is some 1-form on $\mathbb{R}$, and therefore it must be $\pi^\# \alpha = f(x)\, dx$ for some function $f\colon \mathbb{R} \to \mathbb{R}$. Since the projection commutes with integer translations $K_n := x \mapsto x + n$, and since $K_n^\#(f(x)\, dx) = f(x + n)\, dx$, we must have $f(x + n) = f(x)$, i.e., f is a periodic function.

The question is whether $\alpha = dg$ for some function $g\colon S^1 \to \mathbb{R}$. Now a function g on S^1 induces a function $\tilde{g}\colon \mathbb{R} \to \mathbb{R}$ by $\tilde{g} = \pi^\# g = g \circ \pi$, which is invariant under translations, and conversely a periodic function on $\mathbb{R}$ descends to a well-defined smooth function on the quotient S^1.

Hence we will have $\alpha = dg$ if and only if $\pi^\# \alpha = \pi^\#(dg) = d(g \circ \pi)$, or in other words if and only if $f(x)\, dx = \tilde{g}'(x)\, dx$ for some function $\tilde{g}$ on $\mathbb{R}$ which is equal to $g \circ \pi$ (i.e., $\tilde{g}$ is periodic on $\mathbb{R}$). Now if $f(x) = \tilde{g}'(x)$ for all x, then clearly we must

have $\tilde{g}(x) = \int_0^x f(s)\,ds$, and since f is periodic, we know that $\tilde{g}$ is periodic if and only if $\int_0^1 f(s)\,ds = 0$.

Finally, given any closed 1-form α on S^1, let $C = \int_{S^1} \alpha = \int_0^1 \pi^\# \alpha$. Define ϕ to be the 1-form on S^1 such that $\pi^\# \phi = dx$. Then

$$\pi^\#(\alpha - C\phi) = \big(f(x) - C\big)\,dx = \tilde{f}(x)\,dx$$

where $\int_0^1 \tilde{f}(x)\,dx = 0$ by definition of C, and thus we have $\pi^\#(\alpha - C\phi) = d\tilde{g}$ for some periodic function $\tilde{g}$ on $\mathbb{R}$ which descends to a smooth function $g\colon S^1 \to \mathbb{R}$ satisfying $g = \tilde{g} \circ \pi$. We therefore have $\pi^\#(\alpha - C\phi) = d(g \circ \pi) = \pi^\# dg$, which implies that $\alpha - C\phi = dg$.

In conclusion every closed 1-form α is equal to some constant multiple of the basic 1-form ϕ up to the differential of a function, and hence the quotient space is one-dimensional and spanned by $[\phi]$. □

18.2. **Homotopy invariance of cohomology.** In this Section we will construct some useful tools for understanding and computing cohomology spaces. The first is induced by the pull-back operation: given any map $\eta\colon M \to N$, there is an induced map $\eta^\#$ from $\Omega^k(N)$ to $\Omega^k(M)$ for each k, and since d commutes with $\eta^\#$, this descends to a linear map from $H^k(N)$ to $H^k(M)$. As a consequence it turns out that if η is a diffeomorphism, this linear map is an isomorphism. This should not be too surprising. What ends up being much more surprising is that smoothly homotopic manifolds have isomorphic de Rham cohomologies, and from this one can even prove that homeomorphic manifolds have isomorphic de Rham cohomologies.

First we discuss the map induced by pull-back.

Theorem 18.2.1. *Let M and N be smooth manifolds, possibly of different dimensions, and let $\eta\colon M \to N$ be a smooth map. Then for each $k \geq 0$ the linear map $\eta^\#\colon \Omega^k(N) \to \Omega^k(M)$ defined by* (16.4.1) *descends to a linear map $\eta^*\colon H^k(N) \to H^k(M)$ on the de Rham cohomology spaces defined by Definition 18.1.1.*

If $\eta\colon M \to N$ and $\xi\colon N \to P$ are both smooth maps, then the induced cohomology map $(\xi \circ \eta)^*\colon H^k(P) \to H^k(M)$ is given by $(\xi \circ \eta)^* = \eta^* \circ \xi^*$. In particular if $\eta\colon M \to N$ is a diffeomorphism, then $\eta^*\colon H^k(N) \to H^k(M)$ is an isomorphism of vector spaces.

Proof. We use Proposition 16.4.3, which says that $d(\eta^\#\omega) = \eta^\# d\omega$ for any k-form ω. Hence if $\omega \in \Omega^k(N)$ is closed, then so is $\eta^\#\omega$. Similarly if $\omega \in \Omega^k(N)$ is exact, then so is $\eta^\#\omega$. To show that η^* is a well-defined map on the quotient space $H^k(N)$, we want to show that given two closed k-forms ω_1 and ω_2 on N which are the same in $H^k(N)$, the closed k-forms $\eta^\#\omega_1$ and $\eta^\#\omega_2$ are the same in $H^k(M)$. That is, we want to show if $\omega_1 - \omega_2 = d\beta$ for some $(k-1)$-form β on N, then $\eta^\#\omega_1 - \eta^\#\omega_2 = d\alpha$ for some $(k-1)$-form α on M. Clearly this is true by picking $\alpha = \eta^\#\beta$. Hence we have a well-defined map on cohomology.

The formula $(\xi \circ \eta)^* = \eta^* \circ \xi^*$ on cohomology follows from the same formula for pull-backs of forms in general. This is true because for any k-form ω on P, and any point $p \in M$ and any vectors $v_1, \ldots, v_k \in T_pM$, we have by Definition 16.4.1 that

$$\begin{aligned}\big((\xi \circ \eta)^\#\omega\big)(v_1, \ldots, v_k) &= \omega\big((\xi \circ \eta)_* v_1, \ldots, (\xi \circ \eta)_* v_k\big) = \omega\big(\xi_*(\eta_* v_1), \ldots, \xi_*(\eta_* v_k)\big) \\ &= (\xi^\#\omega)(\eta_* v_1, \ldots, \eta_* v_k) = \big(\eta^\#(\xi^\#\omega)\big)(v_1, \ldots, v_k).\end{aligned}$$

Since the left and right sides are equal on all collections of k vectors at any point of M, they must be equal as k-forms on M.

The fact that diffeomorphisms induce isomorphisms comes from taking $P = M$ and $\xi = \eta^{-1}$, which implies that $(\eta^{-1})^* \circ \eta^* = \mathrm{id}^* = \mathrm{id}$. Hence if η is a diffeomorphism, then η^* is invertible. □

One goal of the de Rham cohomology theory (which de Rham accomplished in his thesis) is to show that if two smooth manifolds M and N are homeomorphic as topological spaces, then they have the same de Rham cohomology spaces. This isn't obvious since even spaces like $\mathbb{R}^4$ have a variety of nonequivalent differentiable structures: that is, there are homeomorphisms from $\mathbb{R}^4$ (with one smooth structure) to $\mathbb{R}^4$ (with a different smooth structure) which are not diffeomorphisms. So it is not reasonable to hope for homeomorphism invariance of de Rham cohomology, although fortunately it works anyway.

The concern is that even if homeomorphisms happen to be smooth, they need not be diffeomorphisms since they may collapse tangent spaces; for example $x \mapsto x^3$ is a smooth homeomorphism of $\mathbb{R}$ but not a diffeomorphism. It turns out that the collapsing doesn't matter for the pull-back on forms, precisely because the pull-back $\eta^\#$ is well-defined regardless of whether η is a diffeomorphism or just an arbitrary smooth map. Recall that we expected $d \circ \eta^\# = \eta^\# \circ d$ as in Proposition 16.4.3 when η is a diffeomorphism simply because d is defined in a coordinate-invariant way, and a diffeomorphism of manifolds is indistinguishable in coordinates from an ordinary coordinate-transition map. However because $\eta^\#$ makes sense for *arbitrary* smooth maps, we get $d \circ \eta^\# = \eta^\# \circ d$ also for arbitrary smooth maps, for free. This will happen again when we study deformations of manifolds: we expect that if we have a family of diffeomorphisms of a manifold (such as the flow of a vector field), the cohomology will be preserved. But it turns out that for free it's preserved even if the family consists only of smooth maps and not necessarily diffeomorphisms. This is homotopy invariance.

To motivate what we're about to do a little more, recall the proof of Theorem 18.1.6, where we proved $H^1(\mathbb{R}^n) = \{0\}$. The technique was essentially to use the Fundamental Theorem of Calculus $\int_\gamma df = f(\gamma(1)) - f(\gamma(0))$ when $\gamma\colon [0,1] \to M$, which is Theorem 17.1.6. To find $f(p)$ given ω which we hoped was df, we simply integrated along a particular piecewise-smooth path γ for which $\gamma(0) = \mathbf{0}$ and $\gamma(1) = p$. Certainly it shouldn't have mattered which path γ we used to get from $\mathbf{0}$ to p; we could have used any path. I used a piecewise path parallel to the coordinate axes in order to make the proof that $df = \omega$ a little easier. But the key here is that I can come up with some systematic way of building paths from a given point to any other point; integrating forms along these paths will hopefully invert the d operator. We'll have to figure out what this means in a moment.

Now let's digress (even more) to discuss how forms change under deformations of a manifold. The simplest example is the flow of a vector field, so suppose M is a smooth manifold with a smooth vector field X on it. Then X has a local flow Φ_t by Definition 14.3.3 which satisfies $\Phi_0(p) = p$ and $\frac{\partial \Phi_t}{\partial t}(p) = X_{\Phi_t(p)}$. For any k-form ω we get a pulled-back form $\Phi_t^\# \omega$, which we can differentiate with respect to t in each cotangent space one point at a time. This is the analogue of the Lie derivative of a vector field from Proposition 14.5.4, and is naturally called the Lie derivative of the k-form. In the same way as we obtained the simple formula $\mathcal{L}_X Y = [X, Y]$

for any vector field Y, we will obtain a simple formula for the Lie derivative of the k-form.

First we discuss the interior product ι_X, which takes any k-form to a $(k-1)$-form by filling one slot with X. The idea behind using it is that if we want to integrate a k-form along a path, we can plug in the tangent vector along the path as one of the arguments of the k-form and reduce it to a $(k-1)$-form (which we hope will help us invert the d operator, which goes the other way).

Definition 18.2.2. Let M be a manifold and X a vector field. The operator $\iota_X \colon \Omega^k(M) \to \Omega^{k-1}(M)$ is defined for any k-form ω by the formula

$$\iota_X \omega(Y_1, \ldots, Y_{k-1}) = \omega(X, Y_1, \ldots, Y_{k-1}).$$

For a 1-form ω, we have $\iota_X(\omega) = \omega(X)$, a function.

By tensoriality of ω, the value $(\iota_X \omega)_p$ depends only on X_p and ω_p, so we can also think of it as an operator in each tangent space.

We now prove Cartan's magic formula, showing how to compute the Lie derivative of a k-form in direction X in terms of d and ι_X. We will only do it in the case $k = 1$ right now for simplicity, since the result for k-forms will follow from the much more general result of Theorem 18.2.5.

Proposition 18.2.3. *Suppose M is a smooth manifold, ω is a smooth 1-form on M, and X is a smooth vector field on M. Let Φ_t be the local flow of X, and assume for simplicity that Φ_t is defined globally (for all t, everywhere on M). Then we have*

$$\frac{\partial}{\partial t} \Phi_t^\# \omega = \Phi_t^\# (\mathcal{L}_X \omega) \qquad \text{where} \quad \mathcal{L}_X \omega = \iota_X d\omega + d(\iota_X \omega) \tag{18.2.1}$$

and ι_X is the operator in Definition 18.2.2.

Proof. The first step is to reduce to computing at time $t = 0$, using the additive property of the flow map Φ_t in Proposition 14.4.1: $\Phi_{s+t_o} = \Phi_s \circ \Phi_{t_o}$. The consequence is that if we want to differentiate at time t_o, we can write $t = s + t_o$ and instead differentiate at $s = 0$. We have

$$\frac{\partial}{\partial t}\Big|_{t=t_o} \Phi_t^\# \omega = \frac{\partial}{\partial s}\Big|_{s=0} \Phi_{s+t_o}^\# \omega = \frac{\partial}{\partial s}\Big|_{s=0} \Phi_{t_o}^\# \Phi_s^\# \omega = \Phi_{t_o}^\# \frac{\partial}{\partial s}\Big|_{s=0} \Phi_s^\# \omega = \Phi_{t_o}^\# \mathcal{L}_X \omega.$$

Here we were able to pull the $\Phi_{t_o}^\#$ through the s-derivative since $\Phi_{t_o}^\#$ is really just a linear operator in each cotangent space $T_p^* M$.

Now to compute $\mathcal{L}_X \omega$, let Y be an arbitrary vector field on M; we will compute both sides of (18.2.1) on Y and show that we get the same answer. Recall the definition of $\mathcal{L}_X Y$ from Proposition 14.5.4, which can be written in the form

$$\mathcal{L}_X Y = \frac{\partial}{\partial t}\Big|_{t=0} (\Phi_{-t})_\# Y,$$

where the push-forward of a vector field Y is defined as in Definition 14.2.6 by $\big((\Phi_{-t})_\# Y\big)_p = (\Phi_{-t})_* (Y_{\Phi_t(p)})$. We therefore have

$$\begin{aligned} \big(\Phi_t^\# \omega\big)_p \big(((\Phi_{-t})_\# Y)_p\big) &= \Phi_t^\# \omega\big((\Phi_{-t})_* (Y_{\Phi_t(p)})\big) \\ &= \omega\big((\Phi_t)_* (\Phi_{-t})_* (Y_{\Phi_t(p)})\big) = \omega_{\Phi_t(p)} (Y_{\Phi_t(p)}). \end{aligned}$$

Thus if f denotes the function $\omega(Y)$, we have

$$(\Phi_t^\# \omega)\big((\Phi_{-t})_\# Y\big) = f \circ \Phi_t.$$

Differentiating both sides with respect to t and using the product rule (since after all, the operation of ω on the vector fields Y_i is just multiplication of the coefficients), we obtain

$$(\mathcal{L}_X\omega)(Y) + \omega(\mathcal{L}_X Y) = X\big(\omega(Y)\big).$$

Now using the fact from Proposition 14.5.4 that $\mathcal{L}_X Y = [X, Y]$, we get

$$\mathcal{L}_X\omega(Y) = X\big(\omega(Y)\big) - \omega([X, Y]). \tag{18.2.2}$$

On the other hand using Definition 16.1.2 for $d\omega$ we have

$$\begin{aligned}(\iota_X d\omega + d\iota_X\omega)(Y) &= d\omega(X, Y) + Y\big(\omega(X)\big)\\ &= X\big(\omega(Y)\big) - Y\big(\omega(X)\big) - \omega([X, Y]) + Y\big(\omega(X)\big)\\ &= X\big(\omega(Y)\big) - \omega([X, Y]).\end{aligned}$$

Matching against (18.2.2) gives Cartan's magic formula (18.2.1) for 1-forms. □

It turns out that we can generalize the formula (18.2.1) to any k-form and get the same thing: $\mathcal{L}_X\omega = \iota_X d\omega + d\iota_X\omega$. More significantly, however, we can prove an analogous formula for the derivative $\frac{\partial}{\partial t}\eta_t^\#\omega$ when η_t is *any* time-dependent family of maps, regardless of whether η_t is a diffeomorphism, and Cartan's magic formula will fall out if η_t happens to be the flow of a vector field. In fact if we don't demand that $(\eta_t^{-1})^\#\frac{\partial}{\partial t}\eta_t^\#\omega$ is independent of time, we can also discard the assumption that $\eta_{t+s} = \eta_t \circ \eta_s$. The fact that we can get such a formula which involves only d will be the main step toward showing that cohomology is homotopy-independent: we just let η_t be an arbitrary smooth family of maps whose endpoints are two homotopic spaces, and we will be able to conclude that the cohomology spaces are isomorphic.

Suppose $\eta\colon [0, 1] \times M \to N$ is a smooth map; we denote $\eta_t(p) = \eta(t, p)$ when we want to consider a particular map η_t. Given a k-form ω on N, we have a family $\alpha_t = \eta_t^\#\omega$ of k-forms on M, which of course depends smoothly on t. For each point $p \in M$, the function $t \mapsto \alpha_t(p)$ is a curve in $\Omega^k(T_pM)$, a single vector space, and so it makes sense to subtract them and compute a derivative. We therefore have

$$\alpha_1(p) - \alpha_0(p) = (\eta_1^\#\omega)_p - (\eta_0^\#\omega)_p = \int_0^1 \frac{\partial}{\partial t}(\eta_t^\#\omega)_p\, dt. \tag{18.2.3}$$

If we can find a formula for the derivative of a pullback in the general case which is similar to Cartan's magic formula (18.2.1) in the case where η is a diffeomorphism, we can hope that we can relate the exterior derivative on the set $\eta_1[M]$ to that on the set $\eta_0[M]$, and if we are clever we can make $\eta_0[M]$ a much simpler space than $\eta_1[M]$ is. For example, there is a homotopy from the plane minus a point to the unit circle, and reducing the dimension of a manifold is clearly a great way to simplify cohomology.

First we define a generalization of the interior product ι_X from Definition 18.2.2 and establish some of its properties.

Proposition 18.2.4. *Suppose $\eta\colon [0, 1] \times M \to N$ is a smooth map which gives a family $\eta_t\colon M \to N$ of smooth maps depending smoothly on t. For each $p \in M$, define $X_t(p) = \frac{\partial}{\partial t}\eta_t(t, p) \in T_{\eta_t(p)}N$. Then $X_t\colon M \to TN$, and we can define for any k-form ω on N the interior product $\iota_{X_t}\omega$, a time-dependent $(k-1)$-form on*

M, by the formula

$$\iota_{X_t}\omega(v_1,\dots,v_{k-1}) = \omega_{\eta_t(p)}\big(X_t(p),(\eta_t)_*v_1,\dots,(\eta_t)_*v_{k-1}\big) \quad \text{for any } v_1,\dots,v_{k-1}\in T_pM. \tag{18.2.4}$$

Then the operator ι_{X_t} satisfies the product rule

$$\iota_{X_t}(\alpha\wedge\beta) = \iota_{X_t}\alpha\wedge\eta_t^{\#}\beta + (-1)^k\eta_t^{\#}\alpha\wedge\iota_{X_t}\beta \tag{18.2.5}$$

for any k-form α and ℓ-form β on M.

Proof. We just have to check both sides of (18.2.5) on $(k+\ell-1)$ vectors in any tangent space T_pM. It is enough to check this when β is a 1-form, since we can build up the general case by induction using associativity of the wedge product. We have

$$\begin{aligned}
\iota_{X_t}(\alpha\wedge\beta)(v_1,\dots,v_k) &= (\alpha\wedge\beta)(X_t(p),(\eta_t)_*v_1,\dots,(\eta_t)_*v_k)\\
&= \sum_{j=1}^{k}(-1)^{k-j}\alpha\big(X_t(p),(\eta_t)_*v_1,\dots,\widehat{(\eta_t)_*v_j},\dots,(\eta_t)_*v_k\big)\cdot\beta\big((\eta_t)_*v_j\big)\\
&\qquad + (-1)^k\alpha\big((\eta_t)_*v_1,\dots,(\eta_t)_*v_k\big)\cdot\beta\big(X_t(p)\big)\\
&= \sum_{j=1}^{k}(-1)^{k-j}\iota_{X_t}\alpha(v_1,\dots,\hat{v_j},\dots,v_k)\cdot\eta_t^{\#}\beta(v_j)\\
&\qquad + (-1)^k\eta_t^{\#}\alpha(v_1,\dots,v_k)\cdot\iota_{X_t}\beta\\
&= (\iota_{X_t}\alpha\wedge\eta_t^{\#}\beta)(v_1,\dots,v_k) + (-1)^k(\eta_t^{\#}\alpha\wedge\iota_{X_t}\beta)(v_1,\dots,v_k).
\end{aligned}$$

Since this is true for any collection of vectors in T_pM, the k-forms are equal. Since any ℓ-form can be expressed as a linear combination of wedge products of 1-forms, we can use this result one step at a time to get the general formula for wedge products of k-forms and ℓ-forms. □

Theorem 18.2.5. *Suppose $\eta\colon[0,1]\times M\to N$ is a smooth map (in the sense of a manifold with boundary as in the end of Chapter 17; that is, η extends to a smooth map on some neighborhood of $[0,1]\times M$ in $\mathbb{R}\times M$. Denote $\eta_t = p\mapsto\eta(t,p)$. Then for any k-form ω we have*

$$\frac{\partial}{\partial t}\eta_t^{\#}\omega = \iota_{X_t}d\omega + d(\iota_{X_t}\omega). \tag{18.2.6}$$

Proof. The trick is to do it for 1-forms first and establish a product rule which will take care of it for higher-degree forms.

First let's prove (18.2.6) for 1-forms. Now any 1-form can be written (at least locally) as a sum of terms of the form $f\,dg$ where f and g are smooth functions on some open set of N. By linearity and locality, we will get the 1-form formula as

long as we can prove it if $\omega = f\,dg$. In this case the left side of (18.2.6) is

$$\begin{aligned}\frac{\partial}{\partial t}\eta_t^{\#}(f\,dg) &= \frac{\partial}{\partial t}\big(\eta_t^{\#} f \cdot \eta_t^{\#} dg\big)\\ &= \frac{\partial}{\partial t}\big((f\circ\eta_t)\cdot d(g\circ\eta_t)\big)\\ &= \left(\frac{\partial}{\partial t}(f\circ\eta_t)\right)\cdot d(g\circ\eta_t) + (f\circ\eta_t)\cdot d\left(\frac{\partial}{\partial t}(g\circ\eta_t)\right)\\ &= X_t(f)\cdot\eta_t^{\#}dg + (\eta_t^{\#}f)\cdot d\big(X_t(g)\big)\end{aligned}$$

Here we freely use $\eta_t^{\#}\circ d = d\circ\eta_t^{\#}$. The right side of (18.2.6) is

$$\begin{aligned}\iota_{X_t}d(f\,dg) + d\big(\iota_{X_t}(f\,dg)\big) &= \iota_{X_t}(df\wedge dg) + d\big(\eta_t^{\#}f\cdot X_t(g)\big)\\ &= X_t(f)\cdot\eta_t^{\#}dg - X_t(g)\cdot\eta_t^{\#}df\\ &\qquad + d(\eta_t^{\#}f)\cdot X_t(g) + \eta_t^{\#}f\cdot d\big(X_t(g)\big)\\ &= X_t(f)\cdot\eta_t^{\#}dg + \eta_t^{\#}f\cdot d\big(X_t(g)\big)\end{aligned}$$

Hence formula (18.2.6) works if $\omega = f\,dg$ for some smooth functions f and g, and thus it works for a general 1-form.

Now we know formula (18.2.6) works for 1-forms, and we want to prove it works for k-forms. The easiest thing to do it to use induction. That is, we are going to assume that the formula (18.2.6) holds for k-forms α and for ℓ-forms β, and then prove that it must also hold for the $(k+\ell)$-form $\alpha\wedge\beta$. Since any form of high degree can always be expressed as a linear combination of wedge products of forms of lower degree, and since both sides of (18.2.6) are linear, this will prove the formula for all degrees.

The first step is to notice that we have a product rule:

$$\begin{aligned}\frac{\partial}{\partial t}\eta_t^{\#}(\alpha\wedge\beta) &= \frac{\partial}{\partial t}(\eta_t^{\#}\alpha)\wedge(\eta_t^{\#}\beta)\\ &= \left(\frac{\partial}{\partial t}\eta_t^{\#}\alpha\right)\wedge(\eta_t^{\#}\beta) + (\eta_t^{\#}\alpha)\wedge\left(\frac{\partial}{\partial t}\eta_t^{\#}\beta\right).\end{aligned}\tag{18.2.7}$$

This works since in coordinates, the wedge product is just a sum of products of coefficients (with some signs arising from permutations) and we have a product rule on each of the coefficient products. On each of the lower-degree terms we already know (18.2.6), so we can write

$$\frac{\partial}{\partial t}\eta_t^{\#}(\alpha\wedge\beta) = (\iota_{X_t}d\alpha + d\iota_{X_t}\alpha)\wedge(\eta_t^{\#}\beta) + (\eta_t^{\#}\alpha)\wedge(\iota_{X_t}d\beta + d\iota_{X_t}\beta).\tag{18.2.8}$$

So really the question is whether it's true that the right side of (18.2.8) is what we want, i.e., whether

$$\begin{aligned}\iota_{X_t}d(\alpha\wedge\beta) + d\iota_{X_t}(\alpha\wedge\beta) &= (\iota_{X_t}d\alpha + d\iota_{X_t}\alpha)\wedge(\eta_t^{\#}\beta)\\ &\qquad + (\eta_t^{\#}\alpha)\wedge(\iota_{X_t}d\beta + d\iota_{X_t}\beta).\end{aligned}\tag{18.2.9}$$

So all we have to do is to combine the product rule for the exterior derivative d from Proposition 16.3.6 with the product rule for the interior product ι_{X_t} from Proposition 18.2.4, hoping that the various signs cancel themselves out.

We suppose α is a k-form and β is an ℓ-form. We have to keep track of the fact that ι_{X_t} and d change the degree of the form when computing, and use the fact that the sign depends on the degree of the first factor. We have

$$\begin{aligned}
\iota_{X_t} d(\alpha\wedge\beta) + d\iota_{X_t}(\alpha\wedge\beta) &= \iota_{X_t}\big(d\alpha\wedge\beta + (-1)^k\alpha\wedge d\beta\big)\\
&\qquad + d\big((\iota_{X_t}\alpha)\wedge(\eta_t^{\#}\beta) + (-1)^k(\eta_t^{\#}\alpha)\wedge(\iota_{X_t}\beta)\big)\\
&= (\iota_{X_t}d\alpha)\wedge(\eta_t^{\#}\beta) + (-1)^{k+1}(\eta_t^{\#}d\alpha)\wedge(\iota_{X_t}\beta)\\
&\qquad + (-1)^k(\iota_{X_t}\alpha)\wedge(\eta_t^{\#}d\beta) + (-1)^k(-1)^k(\eta_t^{\#}\alpha)\wedge(\iota_{X_t}d\beta)\\
&\qquad + (d\iota_{X_t}\alpha)\wedge(\eta_t^{\#}\beta) + (-1)^{k-1}(\iota_{X_t}\alpha)\wedge(d\eta_t^{\#}\beta)\\
&\qquad + (-1)^k(d\eta_t^{\#}\alpha)\wedge(\iota_{X_t}\beta) + (-1)^k(-1)^k(\eta_t^{\#}\alpha)\wedge(d\iota_{X_t}\beta)\\
&= (\iota_{X_t}d\alpha)\wedge(\eta_t^{\#}\beta) + (\eta_t^{\#}\alpha)\wedge(\iota_{X_t}d\beta)\\
&\qquad + (d\iota_{X_t}\alpha)\wedge(\eta_t^{\#}\beta) + (\eta_t^{\#}\alpha)\wedge(d\iota_{X_t}\beta)
\end{aligned}$$

as we hoped, using of course the fact that $\eta_t^{\#}$ commutes with d. We thus know (18.2.6) is true for forms of any order. □

Notice that this proof gives results already in the case where η_t is the flow of a vector field X: we proved Cartan's magic formula for 1-forms but not for k-forms, and this proof gets us the magic formula when we use $\eta_t = \Phi_t$ and incorporate the group homomorphism property $\Phi_t \circ \Phi_s = \Phi_{s+t}$. More importantly for the present purpose, having a nice formula for the derivative along a curve gives us a nice formula for the difference at the endpoints, just using the Fundamental Theorem of Calculus.

Theorem 18.2.6. *Suppose that $\eta_t\colon M \to N$ is a family of smooth maps that depends smoothly on t as in Theorem 18.2.5. Suppose that ω is a closed k-form on N; then $\eta_1^{\#}\omega = \eta_0^{\#}\omega + d\phi$ for some $(k-1)$-form ϕ on M.*

As a consequence, if $\eta_t^*\colon H^k(N) \to H^k(M)$ denote the map on cohomology induced by the pull-back $\eta_t^{\#}\colon \Omega^k(N) \to \Omega^k(M)$, then $\eta_0^* = \eta_1^*$.

Proof. To prove the first part, just integrate the derivative, which makes sense since the derivative is being taken point-by-point in particular vector spaces $\Omega^k(T_pM)$ for each p.

We have

$$\begin{aligned}
\eta_1^{\#}\omega - \eta_0^{\#}\omega = \int_0^1 \frac{\partial}{\partial t}\eta_t^{\#}\omega &= \int_0^1 (\iota_{X_t}d\omega)\,dt + \int_0^1 (d\iota_{X_t}\omega)\,dt\\
&= \int_0^1 (\iota_{X_t}d\omega)\,dt + d\left(\int_0^1 (\iota_{X_t}\omega)\,dt\right),
\end{aligned}$$

since in the last term d is taken with respect to the spatial variables and therefore commutes with the time integral. If ω is closed, then $d\omega = 0$ and we can write $\eta_1^{\#}\omega - \eta_0^{\#}\omega = d\phi$ where $\phi = \int_0^1 (\iota_{X_t}\omega)\,dt$.

Now to prove the second part, we know that each $\eta_t^{\#}$ descends to a map η_t^* from $H^k(N)$ to $H^k(M)$ by Theorem 18.2.1. Given a closed k-form $\omega \in Z^k(N)$, let $\alpha_t = \eta_t^{\#}\omega$. Let $[\omega]$ be the equivalence class of ω in the de Rham cohomology $H^k(N)$, and let $[\alpha_t]$ denote the equivalence class of α_t in $H^k(M)$. Since η_t^* is well-defined on equivalence classes, we have $[\alpha_t] = \eta_t^*[\omega]$ (i.e., the equivalence class of α_t

depends only on the equivalence class of ω). We want to show that $[\alpha_0] = [\alpha_1]$ in $H^k(M)$, which means we want to show that $\alpha_1 - \alpha_0 = d\phi$ for some $\phi \in \Omega^{k-1}(M)$. But this is precisely what we just did. □

18.3. **Applications of homotopy invariance.** To illustrate how useful Theorem 18.2.6 is, we will use it to compute the de Rham cohomology of a variety of spaces. The main technique is when $M = N$ and our family η_t satisfies $\eta_1 = \mathrm{id}$, while the image of η_0 is a hopefully simpler space (e.g., of smaller dimension). In such a situation, $\eta_0[M]$ is called a *smooth deformation retract* of M.

First we generalize Theorem 18.1.6, where we showed that $H^1(\mathbb{R}^n)$ is trivial for $n > 0$. We can easily prove the general case since $\mathbb{R}^n$ is contractible to the origin (i.e., there is a smooth deformation retract to a single point).

Proposition 18.3.1. *For any integers $n > 0$ and $k > 0$ we have $H^k(\mathbb{R}^n) = \{0\}$.*

Proof. Let $M = N = \mathbb{R}^n$, and define $\eta_t \colon \mathbb{R}^n \to \mathbb{R}^n$ by $\eta_t(\mathbf{x}) = t\mathbf{x}$. Obviously η_t is smooth in both t and $\mathbf{x}$. When $t = 1$ we get the identity map, and when $t = 0$ we get the trivial map $\eta_0(\mathbf{x}) = \mathbf{0}$.

The induced map $\eta_t^* \colon H^k(\mathbb{R}^n) \to H^k(\mathbb{R}^n)$ therefore does not depend on t. When $t = 1$ this is the identity map. When $t = 0$ we have to understand $H^k(\{\mathbf{0}\})$. Note that $\Omega^k(\{0\})$ is trivial for $k > 1$; since the tangent vector space is the trivial one containing only the zero vector, the cotangent vector space must also contain only the zero vector, and hence so do all the higher k-forms spaces. This means the map η_0^* must take all of $H^k(\mathbb{R}^n)$ to zero, which means $H^k(\mathbb{R}^n)$ must itself be the trivial vector space. □

The other space which has nontrivial cohomology is $M = \mathbb{R}^2 \backslash \{\mathbf{0}\}$. We saw in Example 17.1.9 that there is at least one 1-form

$$\omega = -\frac{y}{x^2 + y^2}\, dx + \frac{x}{x^2 + y^2}\, dy \tag{18.3.1}$$

such that $d\omega = 0$ but ω is not df for any $f \colon M \to \mathbb{R}$. We revisited this again in Example 18.1.2. This allowed us to conclude that $H^1(M)$ has dimension at least one, but we were not able to prove that the dimension is exactly one, or that $H^2(M)$ has dimension zero. Now the fact that the plane minus the origin can be deformed onto a circle will finish this computation.

Proposition 18.3.2. *If M is the plane $\mathbb{R}^2$ minus the origin $\{\mathbf{0}\}$, then M has the same cohomology as the circle, as in Theorem 18.1.7.*

Proof. Let $M = N = \mathbb{R}^2 \backslash \{\mathbf{0}\}$, and let $\eta_t \colon M \to M$ be the map $\eta_t(\mathbf{x}) = \|\mathbf{x}\|^{t-1}\mathbf{x}$. Clearly η_t depends smoothly on both $\mathbf{x}$ and t as long as $\mathbf{x} \neq \mathbf{0}$. And clearly η_1 is the identity while η_0 maps all of M onto the circle. By Theorem 18.2.6, the identity map on cohomology of M must be the same as the map to cohomology of S^1. Hence for $k \geq 2$ we have $H^k(M) = \{0\}$. On the other hand, for $k = 1$ we just need to check that the cohomology map is not trivial into S^1, and this comes from the example ω from (18.3.1); the pull-back $\eta_0^\#$ maps this ω to the nontrivial 1-form ϕ from Theorem 18.1.7, and thus the image of $H^1(M)$ is both at least and at most one-dimensional. So $H^1(M)$ itself must be one-dimensional. □

We can also give a proof that there is no smooth vector field on S^2 which is nowhere-zero; recall that we gave a proof in Example 12.2.5, but obviously this

relied on a very special property of the 2-sphere which we cannot generalize. The following proof makes it a bit clearer what's going on.

Proposition 18.3.3. *There is no vector field X on S^2 which is nowhere-zero.*

Proof. Assume to get a contradiction that X is a nowhere-zero vector field on S^2. Let $\iota\colon S^2 \to \mathbb{R}^3$ be the inclusion, and define $f(p) = \|\iota_*(X_p)\|$. Since $X_p \neq 0$, this is a smooth function from S^2 to $\mathbb{R}^+$, and thus $Y = (1/f)X$ is also a smooth nowhere-zero vector field on S^2.

Define $\tilde{\eta}_t(p) = (\cos t)\iota(p) + (\sin t)\iota_*(Y_p)$ for $p \in S^2$ and $t \in \mathbb{R}$, where the addition and multiplication takes place in $\mathbb{R}^3$. Since $\iota(p)$ and $\iota_*(Y_p)$ are orthonormal in $\mathbb{R}^3$, it's easy to see that $\tilde{\eta}_t(p)$ is actually a unit vector in $\mathbb{R}^3$, and hence we have $\tilde{\eta}_t(p) = \iota(\eta_t(p))$ for some smooth function $\eta_t\colon S^2 \to S^2$. Clearly η_0 is the identity map and $\eta_1(p)$ is the antipodal map on S^2.

Now the antipodal map $A\colon S^2 \to S^2$ is the restriction of the map $J\colon \mathbb{R}^3 \to \mathbb{R}^3$ given by $J(\mathbf{x}) = -\mathbf{x}$, and we have shown that A is homotopic to the identity. This means that the induced map A^* on cohomology $H^2(S^2)$ must be the same as the identity map on $H^2(S^2)$, using Theorem 18.2.6. By the general result Theorem 18.1.4 we know that the top-dimensional cohomology $H^2(S^2)$ is nontrivial.

In fact we can compute at least one nontrivial element. Let $\beta = x\,dy \wedge dz + y\,dz \wedge dx + z\,dx \wedge dy$ in $\Omega^2(\mathbb{R}^3)$, and let $\omega = \iota^\# \beta$ be the corresponding 2-form in $\Omega^2(S^2)$. It is easy to compute that $\int_{S^2} \omega > 0$, and thus although $d\omega = 0$ we cannot have $\omega = d\alpha$ for any 1-form α.

Now since A is the restriction of the negation map J, we have $J \circ \iota = \iota \circ A$, which means that $\iota^\# \circ J^\# = A^\# \circ \iota^\#$. Applying both sides to ω we get $\iota^\# J^\# \omega = A^\# \circ \beta$. Since J negates every component (x, y, z), it is easy to see that $J^\# \beta = -\beta$, and thus $A^\# \omega = -\omega$, and we obtain that $A^*[\omega] = -[\omega]$. But the identity map has $\mathrm{id}^*[\omega] = [\omega]$, and since $[\omega] \neq 0$, we get a contradiction because A is homotopic to the identity. □

More generally one can prove that for any compact orientable n-dimensional manifold M, then $H^n(M) \cong \mathbb{R}$. The idea is basically that any particular choice of volume form (orientation) as in Theorem 17.4.1 will generate the cohomology $H^n(M)$, and that every n-form which integrates to zero must actually be the exterior derivative of some $(n-1)$-form. There is an easy way to build a form if we take for granted a Riemannian metric, using a trick of Moser, but without that it's kind of a pain. Hence we will skip the proof for now. But a consequence is that since the space $H^n(M)$ is always one-dimensional, for any map $\eta\colon M \to M$ we have an induced vector space homomorphism $\eta^*\colon H^n(M) \to H^n(M)$ which must be $\eta^*[\mu] = k[\mu]$ for some integer k. This integer k is then called the *degree of* $\eta\colon M \to M$. There are a variety of other ways to define degree, many of them purely topological, and this generates more relationships between de Rham cohomology and topology.

The computation of de Rham cohomology is generally made fairly routine by using things like the Mayer-Vietoris sequence, which works for differential forms in much the same way it works in algebraic topology. We will not actually discuss the Mayer-Vietoris sequence, instead deferring it to algebraic topology where it properly belongs. But this method allows us to explicitly compute the cohomology of various spheres, along with all the compact surfaces, and many other spaces just by using the homotopy invariance Theorem 18.2.6.

The last thing we will discuss here is the homeomorphism invariance. The essential point is to start with continuous maps and approximate them closely by smooth maps, using the Weierstrass approximation theorem (a result often proved, at least in one space dimension, in undergraduate real analysis). For a full proof see Lee's "Introduction to Smooth Manifolds." The technical details are too involved to make it worth presenting here, I think, so I will just discuss it in the way de Rham originally thought of it.

Recall that for k-chains defined as in Definition 17.2.5 as formal integer linear combinations of k-cubes, we have a boundary operator ∂. We claimed earlier that $\partial\partial c = 0$ for any k-chain c, but we only proved this in Example 17.2.7 for 2-chains. Now let's prove it in general. It's certainly reasonable to expect that it holds since $d^2 = 0$ for forms and for any $(k-2)$-form ω we have

$$\int_{\partial\partial c} \omega = \int_{\partial c} d\omega = \int_c d^2\omega = 0.$$

Hence $\partial(\partial c)$ is a $(k-2)$-chain such that for any $(k-2)$-form ω the integral is zero. We would like to conclude that $\partial\partial c = 0$. But this is kind of a pain; it's easier to just use the algebraic definition of the boundary from Definition 17.2.5.

Proposition 18.3.4. *Let c be any k-cube in a manifold M. Then $\partial(\partial c) = 0$ in the sense of Definition 17.2.5.*

Proof. When we take the boundary of a k-cube c, we basically just plug in either 0 or 1 to one of the entries of c. When we take the boundary of the boundary, we end up with two entries which are independently either 1 or 0, and we just want to make sure every such entry appears exactly once with a plus sign and once with a minus sign.

Consider a k-cube c, and let $\{i, j\}$ be indices such that $1 \leq i < j \leq n$. Consider all terms of $\partial\partial c$ where the i^{th} term is 1 and the j^{th} term is 0. Either we plugged in 1 to the i^{th} term first and then plugged in 0 to the j^{th} term, or the other way around.

Now if we plugged in 1 to the i^{th} term first, then we must have been dealing with the face ι_i^+, which appears in ∂c in the form $\partial c(u^1, \ldots, u^{k-1}) = \cdots + (-1)^{i+1} c(u^1, \ldots, u^{i-1}, 1, u^i, \ldots, u^{k-1})$. When we then take the boundary to plug in 0 to the j^{th} place of the *original* k-cube, we actually notice it's the $(j-1)^{\text{st}}$ place of the boundary cube (since we lost one of the variables by plugging in for the i^{th} place). We therefore see the term

$$\partial(\partial c)(t^1, \ldots, t^{k-2}) = \cdots + (-1)^{i+1}(-1)^{j-1} c(t^1, \ldots, t^{i-1}, 1, t^i, \ldots, t^{j-1}, 0, t^j, \ldots, t^{k-2}).$$

On the other hand if we had plugged in j first then we would have gotten

$$\partial c(u^1, \ldots, u^{k-1}) = \cdots + (-1)^j c(u^1, \ldots, u^i, \ldots, u^{j-1}, 0, u^j, \ldots, u^{k-1}).$$

Taking the boundary again and plugging in 1 for the i^{th} place (since $i < j$ this doesn't change anything) we get the term

$$\partial(\partial c)(t^1, \ldots, t^{k-2}) = \cdots + (-1)^j (-1)^{i+1} c(t^1, \ldots, t^{i-1}, 1, t^i, \ldots, t^{j-1}, 0, t^j, \ldots, t^{k-2}).$$

Now the sign $(-1)^{i+1}(-1)^{j-1}$ obviously cancels out the sign $(-1)^j(-1)^{i+1}$, and this term disappears in the overall sum of $\partial(\partial c)$.

Once we understand how it works in this case, the other three cases are exactly the same (since changing a 1 to a 0 in one place changes *both* signs). So all the terms must be canceling out. □

Now for each k we can, just as in Definition 18.1.1, define closed chains and exact chains. To make the theory look more like the cohomology theory, we will extend the definition of k-chains to be *real linear combinations* of k-cubes, not just *integer* linear combinations of them as we did in Definition 17.2.5. Note that, inspired by the duality, we use subscripts on the spaces Z, B, and H instead of superscripts as we did in Definition 18.1.1.

Definition 18.3.5. Let $Q_k(M)$ denote the set of all real linear combinations of k-cubes on a manifold M, and denote the elements of $Q_k(M)$ as k-chains. Define $Z_k(M)$ to be the set of all k-chains c such that $\partial c = 0$, and define $B_k(M)$ to be the set of all boundaries of $(k+1)$-chains. Since $\partial \circ \partial = 0$, we know that $B_k(M)$ is a vector subspace of $Z_k(M)$. Define $H_k(M) = Z_k(M)/B_k(M)$ to be the *k-dimensional smooth homology of M*.

As we have mentioned repeatedly, the boundary operator ∂ makes sense even if c is only a *continuous* k-cube on a topological space M. Hence there is a notion of *continuous homology* which is in fact is the usual starting point in algebraic topology.

Now for any smooth k-form ω, we have a map $\Lambda_\omega \colon Q_k(M) \to \mathbb{R}$ given by $\Lambda_\omega(c) = \int_c \omega$. By definition of integration over chains, this is linear in c. The idea now is that Stokes' Theorem should tell us that this actually gives a map from the cohomology space $H^k(M)$ to the dual of $H_k(M)$.

Theorem 18.3.6. *Suppose that ω is a closed k-form, and let Λ_ω be the operator on closed k-chains given by $\Lambda_\omega(c) = \int_c \omega$. Then $\omega \mapsto \Lambda_\omega$ respects our equivalence relation and therefore descends to a map from $H^k(M)$ to $H_k(M)^*$, the dual space of the smooth homology space $H_k(M)$.*

Proof. We just have to show that if $\omega_1 - \omega_2 = d\phi$ for some $(k-1)$-form ϕ, and if $c_1 - c_2 = \partial b$ for some $(k+1)$-chain b, then $\int_{c_1} \omega_1 = \int_{c_2} \omega_2$. To do this, notice that

$$\begin{aligned}\int_{c_1} \omega_1 &= \int_{c_2+\partial b} (\omega_2 + d\phi) \\ &= \int_{c_2} \omega_2 + \int_{\partial b} \omega_2 + \int_{c_2} d\phi + \int_{\partial b} d\phi \\ &= \int_{c_2} \omega_2 + \int_b d\omega_2 + \int_{\partial c_2} \phi + \int_b d^2\phi.\end{aligned}$$

Now since $d\omega_1 = d\omega_2 = 0$ and $\partial c_1 = \partial c_2 = 0$ and $d^2\phi = 0$, all terms vanish except the first one, and we get $\int_{c_1} \omega_1 = \int_{c_2} \omega_2$. □

In Theorem 17.1.8 we proved that for closed 1-forms ω and 1-chains γ, if $\int_\gamma \omega = 0$ whenever $\partial\gamma = 0$, then $\omega = df$ for some function f. What this essentially accomplished is to show that the linear map $\Lambda \colon H^1(M) \to H_1(M)^*$ is an isomorphism: if $\Lambda(\omega)(\gamma) = 0$ for all closed 1-chains γ, then ω is the differential of a function, which means that $[\omega] = 0$ in $H^1(M)$. In other words, the induced map from cohomology to dual-homology is injective and therefore an isomorphism (if we knew that all these homology and cohomology spaces were finite-dimensional).

This can be generalized: cohomology classes integrate over homology classes to give an isomorphism, and thus it is natural to think of $H^k(M)$ as the dual space of $H_k(M)$. If we can then prove that the smooth homology is isomorphic to the

continuous homology (which is true since we can approximate continuous functions by smooth functions), it stands to reason that the smooth de Rham cohomology should have a continuous analogue which is the dual space of the continuous homology space.

There is one other thing to be concerned about. The spaces Z^k and B^k and Z_k and B_k are all infinite-dimensional vector spaces of smooth maps, and there is no obvious reason why the quotient spaces $H^k(M)$ and $H_k(M)$ should actually be finite-dimensional. The basic idea is that smooth manifolds can be triangulated, and thus they are homeomorphic to certain graph-like structures (i.e., built up out of vertices, edges, faces, etc.). The homology of such structures can be computed explicitly and is always finite-dimensional, and thus all the other isomorphic spaces are also finite-dimensional.

An alternate proof uses Hodge theory (which makes sense if one has a Riemannian metric on the manifold) to pick out a distinguished closed k-form in each cohomology class which minimizes the norm. The fact that there is only a finite-dimensional space of minimizers and that there is a unique one in each class comes from the fact that the corresponding differential operators end up being *elliptic operators*, which implies compactness of the solution space, and compactness in any linear Banach space implies finite-dimensionality. This sort of thing extends to any elliptic operator on a smooth manifold, and the idea that the dimension of its kernel and cokernel are finite suggests that we might hope that these dimensions can be calculated in terms of purely topological information. This often turns out to be correct, and the general theory is called *index theory*, a rather popular topic of current research.

For now, however, we will abandon all these trains of thought, since it's a huge topic to which we cannot hope to do any more justice.

19. Riemannian metrics

"We meet again, at last. The circle is now complete. When I left you, I was but the learner; now I am the master."

19.1. **Definition and examples.** Notice that *everything* we've done in the previous Chapters does not depend on inner products. Specifically, we've never used the fact that the basic vectors in Cartesian coordinates, $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$, are orthonormal. This is the power of differential geometry in general; even if those vector fields were *not* orthonormal, everything we've done so far would still work.

However, most manifolds have more structure than what we've assumed. They may have a Riemannian metric (a positive definite symmetric inner product on each tangent space), or a Lorentzian metric (a symmetric inner product which is nondegenerate but not necessarily positive), or a symplectic form (a nondegenerate antisymmetric tensor of type $(2,0)$), which represents something important about it, and we want to study the things that make sense if we incorporate this structure.

Riemannian geometry is what most people think of as "geometry," since it represents curved spaces for which we can measure lengths and areas and such. Lorentzian geometry is used in general relativity; many things carry over from Riemannian to Lorentzian geometry, while some do not.

For now we will focus on Riemannian metrics.

Definition 19.1.1. If M is a manifold, the space of symmetric tensors of type $(2,0)$ on M is a vector bundle, the elements of which are inner products. It may be denoted by $\operatorname{Sym}(M)$. A *Riemannian metric* is a section g of this bundle which is positive definite at each point. A *Riemannian manifold* (M,g) is a manifold with some preferred Riemannian metric.

If g is a Riemannian metric and $u \in T_pM$ is a tangent vector at $p \in M$, the *length of* u is $\|u\| = \sqrt{g(p)(u,u)}$. If $\gamma\colon [a,b] \to M$ is a smooth curve, the *length of* γ is

$$L(\gamma) = \int_a^b \left\| \frac{d\gamma}{dt} \right\| dt.$$

We frequently denote a Riemannian metric in coordinates by

$$g = ds^2 = \sum_{i=1}^n \sum_{j=1}^n g_{ij}(x^1, \cdots, x^n)\, dx^i \otimes dx^j.$$

The notation here is inspired by the use of s for an arc length parameter. The symmetry condition is that $g_{ij} = g_{ji}$. Nondegeneracy implies that $\det g > 0$ at every point.

We have already encountered Riemannian metrics, for example in Section 13.3 where we showed that every manifold with a partition of unity has a Riemannian metric, and we mentioned that a Riemannian metric can be used to construct an isomorphism between TM and T^*M. Let's do this now.

Definition 19.1.2. If g is a nondegenerate inner product on a single vector space V, we define the *musical isomorphisms between V and V^** as follows:

- the index-lowering map is $\flat\colon V \to V^*$ given by $v^\flat(w) = g(v,w)$ for any $w \in V$;
- the index-raising map is $\sharp\colon V^* \to V$ given by $g(\alpha^\sharp, w) = \alpha(w)$ for any $w \in V$.

Nondegeneracy of g (the fact that $g(u,v) = 0$ for all v implies $u = 0$) is precisely the condition that $\flat$ is an isomorphism, which is why $\sharp$ exists.

The names arise from the action on a basis. Let $\{e_i\}$ be a basis of V, and let $\{\alpha^i\}$ be the dual basis of V^* satisfying $\alpha^i(e_j) = \delta^i_j$. Set $g_{ij} = g(e_i, g_j)$. If $v = \sum_i v^i e_i$ and $v^\flat = \sum_j \tilde{v}_j \alpha^j$, we must have

$$v^\flat(e_j) = g(v, e_j) = \sum_i v^i g(e_i, e_j) = \sum_i v^i g_{ij}.$$

We conclude that

$$v^\flat = \sum_j \tilde{v}_j \alpha^j = \sum_j \sum_i v^i g_{ij} \alpha^j.$$

Hence in going from v represented by $(v^1, \ldots, v^n)$ to $v^\flat$ represented by $(\tilde{v}_1, \ldots, \tilde{v}_n)$, we have lowered the indices. Similarly inverting this operation raises indices.

Note that since an inner product generates an isomorphism from V to V^*, we obtain an inner product on V^* by the formula $\tilde{g}(\beta, \gamma) = g(\beta^\sharp, \gamma^\sharp)$. Recall that $e_i^\flat = \sum_j g_{ij}\alpha^i$, and define g^{ij} to be the inverse matrix of g_{ij}, so that $\sum_k g_{ik} g^{kj} = \delta_i^j$. Then we must have $(\alpha^i)^\flat = \sum_j g^{ij} e_j$, which implies that

$$\tilde{g}(\alpha^i, \alpha^j) = g\big(\sum_k g^{ik} e_k, g^{j\ell} e_\ell\big) = \sum_{k,\ell} g^{ik} g^{j\ell} g(e_k, e_\ell) == \sum_{k,\ell} g^{ik} g^{j\ell} g_{k\ell} = \sum_k g^{ik} \delta_k^j = g^{ij}.$$

Thus the components of the inner product on V^* in a dual basis are exactly the components of the inverse matrix of the inner product on V in the basis.

It is now trivial to prove that the tangent bundle TM is isomorphic to the cotangent bundle T^*M, given any Riemannian metric on M. It is important to note however that this isomorphism depends on the metric and is not in any sense "natural."

Theorem 19.1.3. *Suppose M is a smooth manifold. Define a Riemannian metric g on M as in Section 13.3 using a partition of unity. Then the operator $v \mapsto v^\flat$ given by Definition 19.1.2 extends to a smooth map $\flat\colon TM \to T^*M$ given by $\flat(v) = v^\flat$, and this map is a bundle-isomorphism in the sense of Definition 12.1.6.*

Proof. It is easy to see that the map $\flat$ is linear in each tangent space, and it is an isomorphism of each tangent space by definition since the inner product is nondegenerate. The only thing we have to worry about is that it is smooth. But this follows in coordinates, since it's a local property. The map $\flat$ is given in a coordinate chart by

$$\frac{\partial}{\partial x^i}\Big|_p \mapsto \sum_j g_{ij}(p) dx^j\big|_p,$$

and since the metric components are smooth, so is this map. □

That's as much as one needs to do to relate metrics to bundles. For now we really just want to study simple examples which lead to interesting geometry.

Example 19.1.4. If $M = \mathbb{R}^n$, the *Euclidean metric* is the one obtained in Cartesian coordinates by

$$ds^2 = \sum_{k=1}^{n} dx^k \otimes dx^k,$$

in other words $g_{ij} = \delta_{ij}$.

In $\mathbb{R}^2$ we have $ds^2 = dx \otimes dx + dy \otimes dy$, which is frequently abbreviated by $ds^2 = dx^2 + dy^2$. We can change coordinates using the formulas $dx = \frac{\partial x}{\partial u}\, du + \frac{\partial x}{\partial v}\, dv$ and $dy = \frac{\partial y}{\partial u}\, du + \frac{\partial y}{\partial v}\, dv$. For example in polar coordinates we have

$$\begin{aligned} ds^2 &= (\cos\theta\, dr - r\sin\theta\, d\theta)^2 + (\sin\theta\, dr + r\cos\theta\, d\theta)^2 \\ &= dr^2 + r^2\, d\theta^2. \end{aligned}$$

In three dimensions we have $ds^2 = dx^2 + dy^2 + dz^2$ and using the spherical coordinates $(x, y, z) = (r\sin\theta\cos\phi, r\sin\theta\sin\phi, r\cos\theta)$ we obtain

$$ds^2 = dr^2 + r^2\, d\theta^2 + r^2\sin^2\theta\, d\phi^2.$$

☺

It is not a coincidence that the diagonal terms cancel out in these formulas; in fact the coordinates themselves were chosen to make this happen. Such coordinates are called "orthogonal," since the diagonal terms being zero is equivalent to the coordinate vectors $\frac{\partial}{\partial x^i}$ being orthogonal to each other. This is common in two dimensions, since if we take an analytic function on $\mathbb{C}$ given by $z = f(w)$, or $x + iy = g(u, v) + ih(u, v)$, then the Cauchy-Riemann equations imply $\frac{\partial x}{\partial u} = \frac{\partial y}{\partial v}$ and $\frac{\partial y}{\partial u} = -\frac{\partial x}{\partial v}$, so that the metric in (u, v) coordinates will be

$$\begin{aligned} ds^2 &= (x_u\, du + x_v\, dv)^2 + (y_u\, du + y_v\, dv)^2 \\ &= (x_u\, du - y_u\, dv)^2 + (y_u\, du + x_v\, dv)^2 \\ &= (x_u^2 + y_u^2)\,(du^2 + dv^2). \end{aligned}$$

Since analytic functions are easy to find, so are orthogonal coordinates. As an example, parabolic coordinates (6.2.4) and elliptic coordinates (6.2.5) both arise from analytic functions: parabolic coordinates $(x, y) = \left(\sigma\tau, \frac{1}{2}(\tau^2 - \sigma^2)\right)$ come from $x + iy = f(\sigma + i\tau)$ where $f(w) = -iw^2/2$, while elliptic coordinates $(x, y) = (\cosh\mu\cos\nu, \sinh\mu\sin\nu)$ come from $x + iy = f(\mu + i\nu)$ where $f(w) = \cosh w$. The metric in parabolic coordinates thus ends up being $ds^2 = (\tau^2 + \sigma^2)\,(d\tau^2 + d\sigma^2)$, while the metric in elliptic coordinates is $ds^2 = (\sinh^2\mu + \sin^2\nu)\,(d\mu^2 + d\nu^2)$. In three dimensions it is much harder to find orthogonal coordinate systems, and the formulas end up being more complicated.

The most common way to get other Riemannian metrics is to have a submanifold of Euclidean space and have the submanifold inherit the Riemannian metric.

Definition 19.1.5. If N is a Riemannian manifold with Riemannian metric h, then a *Riemannian submanifold* M is a manifold with an immersion $\iota\colon M \to N$, and such that the Riemannian metric g on M is given by

$$g(u, v) = h(\iota_* u, \iota_* v) \tag{19.1.1}$$

for any vectors $u, v \in T_p M$.

The inner product g is a genuine Riemannian metric since it is positive-definite: if $u \neq 0$ then $\iota_* u \neq 0$, so that $h(\iota_* u, \iota_* u) > 0$, so that $g(u, u) > 0$.

We can abbreviate the formula (19.1.1) by $g = \iota^* h$.

In most applications we have $N = \mathbb{R}^n$ and h is the Euclidean metric.

Example 19.1.6. The sphere S^2 is a submanifold of Euclidean space $\mathbb{R}^3$, and the immersion ι is given in coordinates by

$$(x, y, z) = \iota(\theta, \phi) = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta).$$

We have

$$\begin{aligned}
\iota_*\left(\frac{\partial}{\partial\theta}\right) &= \cos\theta\cos\phi\,\frac{\partial}{\partial x} + \cos\theta\sin\phi\,\frac{\partial}{\partial y} - \sin\theta\,\frac{\partial}{\partial z} \\
\iota_*\left(\frac{\partial}{\partial\phi}\right) &= -\sin\theta\sin\phi\,\frac{\partial}{\partial x} + \sin\theta\cos\phi\,\frac{\partial}{\partial y}
\end{aligned}$$

so that

$$\begin{aligned}
g_{11} &= \left\langle \frac{\partial}{\partial\theta}, \frac{\partial}{\partial\theta} \right\rangle = 1 \\
g_{12} &= \left\langle \frac{\partial}{\partial\theta}, \frac{\partial}{\partial\phi} \right\rangle = 0 \\
g_{22} &= \left\langle \frac{\partial}{\partial\phi}, \frac{\partial}{\partial\phi} \right\rangle = \sin^2\theta,
\end{aligned}$$

so the metric on S^2 is

$$ds^2 = d\theta^2 + \sin^2\theta\, d\phi^2.$$

Notice that we could also have computed this metric just by using the transformation formulas

$$\begin{aligned}
dx &= \cos\theta\cos\phi\, d\theta - \sin\theta\sin\phi\, d\phi \\
dy &= \cos\theta\sin\phi\, d\theta + \sin\theta\cos\phi\, d\phi \\
dz &= -\sin\theta\, d\theta,
\end{aligned}$$

and plugging into $ds^2 = dx^2 + dy^2 + dz^2$.

This is frequently a much faster shortcut for computing the metric for an embedded submanifold (or more generally for any immersed submanifold). The reason it works is because in the shortcut we compute

$$\begin{aligned}
g = \iota^* h &= \iota^*\left(\sum_{i,j} h_{ij}(x)\, dx^i \otimes dx^j\right) \\
&= \sum_{i,j} (\iota^* h_{ij})\, (\iota^* dx^i) \otimes (\iota^* dx^j) \\
&= \sum_{i,j} h_{ij}(x \circ \iota)\, d(x^i \circ \iota) \otimes d(x^j \circ \iota),
\end{aligned}$$

so it's just using the product rule for pull-backs (a special case of which is (16.4.2)).

In stereographic coordinates (u, v) the immersion is given by

$$(x, y, z) = \iota(u, v) = \left(\frac{2u}{u^2+v^2+1}, \frac{2v}{u^2+v^2+1}, \frac{u^2+v^2-1}{u^2+v^2+1}\right),$$

and we can compute that the metric is

$$ds^2 = \frac{4(du^2 + dv^2)}{(u^2 + v^2 + 1)^2}.$$

We can generalize both coordinate systems easily. For example, spherical coordinates on S^3 would be given by

$$(w, x, y, z) = (\sin\psi \sin\theta \cos\phi, \sin\psi \sin\theta \sin\phi, \sin\psi \cos\theta, \cos\psi),$$

i.e., we just attach a sine of a new coordinate to all the old coordinates and throw in a cosine to the new last coordinate. The metric on S^3 is

$$ds^2 = d\psi^2 + \sin^2\psi \, d\theta^2 + \sin^2\psi \sin^2\theta \, d\phi^2, \tag{19.1.2}$$

and it should be clear how to get any other dimension to work out the same way.

Stereographic coordinates are even easier to generalize. On S^3 we just set

$$(w, x, y, z) = \left(\frac{2t}{t^2 + u^2 + v^2 + 1}, \frac{2u}{t^2 + u^2 + v^2 + 1}, \frac{2v}{t^2 + u^2 + v^2 + 1}, \frac{t^2 + u^2 + v^2 - 1}{t^2 + u^2 + v^2 + 1} \right),$$

and the metric is

$$ds^2 = \frac{4(dt^2 + du^2 + dv^2)}{(t^2 + u^2 + v^2 + 1)^2}.$$

Again it is easy to see how to generalize this. ☺

The simplest general class of examples comes from surfaces in $\mathbb{R}^3$, which were historically the beginning of the theory. If we have a parametrization $(x, y, z) = (f(u, v), g(u, v), h(u, v))$ of some two-dimensional manifold M, then the metric on M is

$$ds^2 = (f_u^2 + g_u^2 + h_u^2)\, du^2 + (f_u f_v + g_u g_v + h_u h_v)\, (du \otimes dv + dv \otimes du) + (f_v^2 + g_v^2 + h_v^2)\, dv^2.$$

So we can just write it as

$$ds^2 = E(u, v)\, du^2 + F(u, v)\, (du \otimes dv + dv \otimes du) + G(u, v)\, dv^2$$

for some functions E, F, G. (Notice that if you take the common shortcut

$$ds^2 = E\, du^2 + 2F\, du\, dv + G\, dv^2,$$

you have to remember that $g(\frac{\partial}{\partial u}, \frac{\partial}{\partial v}) = F$ rather than $2F$; all the cross terms will get an extra factor of two depending on your notation.)

Example 19.1.7. A common example is a surface of revolution. Take a parametric curve in the plane, and suppose it lies entirely to the right of the y-axis. We have $x = \alpha(u), y = \beta(u)$ where $\alpha(u) > 0$ for all u. Now suppose we want to revolve this curve around the y-axis. We introduce a z-axis perpendicular to the plane and consider a parametrized circle of x and z points around each fixed y point, of radius $\alpha(u)$. Thinking of it this way, the parametrization is

$$(x, y, z) = (\alpha(u) \cos v, \beta(u), \alpha(u) \sin v).$$

Then the metric is

$$ds^2 = (\alpha'(u)^2 + \beta'(u)^2)\, du^2 + \alpha(u)^2\, dv^2.$$

If we are fortunate enough that the original curve in the plane is parametrized by arc length, then we have $\alpha'(u)^2 + \beta'(u)^2 = 1$, and the metric becomes

$$ds^2 = du^2 + \alpha(u)^2\, dv^2.$$

The spherical coordinates metric on the sphere is obviously a special case of this.

Another example is a paraboloid, obtained by revolving a parabola $x = u, y = u^2$; we get

$$ds^2 = (1 + 4u^2)\,du^2 + u^2\,dv^2.$$

An ellipsoid comes from revolving $x = a\cos u, y = b\sin u$, so we get

$$ds^2 = (a^2\sin^2 u + b^2\cos^2 u)\,du^2 + a^2\cos^2 u\,dv^2.$$

Unsurprisingly when $a = b = 1$ we get the standard metric on the sphere.

A cylinder comes from revolving a vertical line such as $x = 1$, $y = u$. In this case we get $ds^2 = du^2 + dv^2$. Note that this is the same formula as the flat plane's metric. This will be important in a bit.

A cone comes from revolving $x = u$, $y = u$, which leads to the metric $ds^2 = 2du^2 + u^2\,dv^2$. Rescaling the coordinates to $u = p/\sqrt{(2)}$ and $v = q\sqrt{2}$ we get the metric $ds^2 = dp^2 + p^2\,dq^2$, which is the same formula as the flat plane in polar coordinates. ☺

19.2. **Invariance and curvature.** To do Riemannian geometry, we imagine people living in the manifold. They can measure lengths of curves, and therefore they can measure lengths of vectors (which are just infinitesimal curves). So in any coordinate system they use, they can tell you what the metric components are. If you give them a surface, they'll divide it up into little approximately flat quadrilaterals, and compute the areas using the fact that they know the lengths of all the sides. So they can compute areas. Similarly they can build up volumes and such. Since they can also measure angles between vectors using

$$g(u, v) = |u||v|\cos\theta,$$

they can do most of what we think of as geometry. They'd build "straight lines" out of geodesics, which are curves between two points minimizing lengths. In this way they'd build "triangles." By finding the shortest curve joining one point to another, they'd compute distances. They could build "circles" or "spheres" by finding all points equidistant to the center. The formulas they'd get for the relation between the circumference or surface area of their circles or spheres in terms of their radii would probably be different from what we have in Euclidean space, and that's how they could figure out whether they live in a curved manifold or not. The only Euclidean constructions they might not be able to imitate are sliding things around rigidly and rotating things, although even then some things would still work depending on whether their manifold had symmetry.

The important thing is that the *only* things they can do geometrically, living inside the manifold, are things that involve the metric. Locally (in a single coordinate chart), only the metric components matter. Hence for example locally there is no way for someone living on a cylinder or cone to tell the difference between their surface and the Euclidean plane, since as we saw, in the right coordinates the Riemannian metrics are all the same. Globally one could tell the difference since one coordinate chart works for the plane and multiple are needed for the cone or cylinder, and of course it's an important problem to distinguish manifolds globally that are the same locally, but the first thing to do is understand the local structure in coordinates.

This is harder than it seems at first. Figuring it out for surfaces gets us most of the main ideas we need, so we will specialize to that case. We know that the metric

is completely determined by three functions appearing in the formula

$$ds^2 = E(u,v)\,du^2 + 2F(u,v)\,du\,dv + G(u,v)\,dv^2.$$

However the same space may have many different representations. The Euclidean plane is described by $E(u,v) = 1, F(u,v) = 0, G(u,v) = 1$ (Cartesian coordinates), but it's also described by $E(u,v) = u^2+v^2, F(u,v) = 0, G(u,v) = u^2+v^2$ (parabolic coordinates), and it's also described by $E(u,v) = 1, F(u,v) = 0, G(u,v) = u^2$ (polar coordinates). So the geometry is encoded in these three functions in a rather subtle way. The sphere is described by the three functions $E(u,v) = 1, F(u,v) = 0, G(u,v) = \sin^2 u$; how would I know just by looking at these formulas (without knowing where they came from) that all the Euclidean formulas represent the same space and the spherical formula represents a different space?

The original motivation for this problem is the question we discussed in the introduction to these notes: why is there no good map of the earth? There are good maps of a cylinder (you just cut it along some irrelevant line and unfold the cylinder until it flattens so you see a rectangle) and of the cone (again, cut it and unfold it until it flattens, and you'll see a Pac-Man type of shape in the plane). Something about the sphere prevents us from doing this, and we should be able to see it in the metric components.

The issue here is that three functions determine the metric components, but there's a huge freedom to change coordinates. Roughly speaking I can change coordinates using two new functions, and I expect to set it up so that I can change two of the functions to whatever I want. So there should really only be *one* function that determines the geometry. How would I find it?

It was actually found in a somewhat indirect way historically, and the reason why will be clear once you see the formula.

Here's the basic question. Let's imagine we had some metric given to us:

$$ds^2 = E(u,v)\,du^2 + 2F(u,v)\,du\,dv + G(u,v)\,dv^2.$$

Is this actually a Euclidean metric in some other coordinates? That is, are there functions $x = f(u,v), y = g(u,v)$ such that

$$ds^2 = dx^2 + dy^2?$$

Once we decide what we're looking for, we just apply the coordinate change formula. We want to see

$$dx^2 + dy^2 = E(u,v)\,du^2 + 2F(u,v)\,du\,dv + G(u,v)\,dv^2,$$

which means

$$\begin{aligned}(f_u^2 + g_u^2)\,du^2 + 2(f_u f_v + g_u g_v)\,du\,dv + (f_v^2 + g_v^2)\,dv^2 \\ = E(u,v)\,du^2 + 2F(u,v)\,du\,dv + G(u,v)\,dv^2.\end{aligned}$$

So we get three partial differential equations for the unknown functions f, g:

$$\begin{aligned} f_u^2 + g_u^2 &= E(u,v) \\ f_u f_v + g_u g_v &= F(u,v) \\ f_v^2 + g_v^2 &= G(u,v). \end{aligned} \tag{19.2.1}$$

Three equations for two functions: again there is one constraint that the data will have to satisfy, and this should be where the geometry is.

Our equations involve differentiating with respect to both u and v, so a compatibility condition comes from the fact that *if* there is a solution f, g, then we must have $f_{uv} = f_{vu}$ and $g_{uv} = g_{vu}$. This is exactly the same sort of thing that came up in Example 15.2.9 when we tried to solve the two equations

$$\frac{\partial h}{\partial x} = \alpha(x,y) \qquad \frac{\partial h}{\partial y} = \beta(x,y)$$

for one function h (that is, trying to solve $dh = \omega$ for a 1-form ω). There's no reason for a solution to exist without some special condition, and the condition you find is that α and β must be related by $\frac{\partial \alpha}{\partial y} = \frac{\partial \beta}{\partial x}$, since that's what comes from $\frac{\partial^2 h}{\partial x \partial y} = \frac{\partial^2 h}{\partial y \partial x}$. In other words, $d\omega = 0$. Conversely if this condition is satisfied, then we can solve for h by integrating one equation, and the other one will end up being satisfied, using the Poincaré Lemma, Proposition 15.2.10.

Unfortunately the equations (19.2.1) are nonlinear. Now there is a general technique to figure out the conditions under which you can solve a nonlinear system of partial differential equations, especially when they're overdetermined (like this one is). It's called the Cartan-Kahler method; at the University of Colorado, Jeanne Clelland is the local expert. Since we're only doing this one special case, we won't discuss the general technique, but it's worth knowing that there is one.

So the first thing we can do is differentiate them, because what we'll end up with will be linear in the highest-order terms. So we differentiate each of the three equations with respect to whatever we can, to get six equations.

$$2f_u f_{uu} + 2g_u g_{uu} = E_u \qquad 2f_u f_{uv} + 2g_u g_{uv} = E_v \tag{19.2.2}$$

$$2f_v f_{uv} + 2g_v g_{uv} = G_u \qquad 2f_v f_{vv} + 2g_v g_{vv} = G_v. \tag{19.2.3}$$

$$f_u f_{uv} + f_v f_{uu} + g_u g_{uv} + g_v g_{uu} = F_u \tag{19.2.4}$$

$$f_u f_{vv} + f_v f_{uv} + g_u g_{vv} + g_v g_{uv} = F_v \tag{19.2.5}$$

(19.2.6)

The last two equations simplify, using the others, to

$$f_v f_{uu} + g_v g_{uu} = F_u - \tfrac{1}{2} E_v, \tag{19.2.7}$$

and

$$f_u f_{vv} + g_u g_{vv} = F_v - \tfrac{1}{2} G_u. \tag{19.2.8}$$

We've now got six *linear* equations for six quantities $f_{uu}, f_{uv}, f_{vv}, g_{uu}, g_{uv}, g_{vv}$. In fact they occur in pairs: if we just look at f_{uu} and g_{uu}, then we have

$$\begin{aligned} f_u f_{uu} + g_u g_{uu} &= \tfrac{1}{2} E_u \\ f_v f_{uu} + g_v g_{uu} &= F_u - \tfrac{1}{2} E_v, \end{aligned} \tag{19.2.9}$$

and we can think of this as a matrix equation with coefficient matrix $A = \begin{pmatrix} f_u & g_u \\ f_v & g_v \end{pmatrix}$. Of course, this is the transformation matrix from (u,v) to (x,y), so in particular it should be nonsingular. In fact we can rewrite (19.2.1) as the single matrix equation $AA^T = \Lambda$ where Λ is the matrix of metric coefficients, $\Lambda = \begin{pmatrix} E & F \\ F & G \end{pmatrix}$.

So thinking of (19.2.9) as a matrix equation, we have

$$A \begin{pmatrix} f_{uu} \\ g_{uu} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} E_u \\ 2F_u - E_v \end{pmatrix}.$$

We want to multiply by A^{-1}, but doing this directly would involve finding the determinant $f_u g_v - f_v g_u$, which is not hard but which makes the equations more complicated. Instead we work indirectly: since $AA^T = \Lambda$, we know $A^{-1} = A^T \Lambda^{-1}$; the nice thing about this is that our equations end up *linear* in the unknowns f and g (and rather nonlinear in the knowns, but sacrifices must be made). Then we get

$$\begin{pmatrix} f_{uu} \\ g_{uu} \end{pmatrix} = \frac{1}{2(EG-F^2)} \begin{pmatrix} GE_u f_u + G(2F_u - E_v) f_v - FE_u g_u + (2F_u - E_v) g_v \\ -FE_u f_u - F(2F_u - E_v) f_v + EE_u g_u + E(2F_u - E_v) g_v \end{pmatrix}.$$

Similarly with the other terms.

Summarizing it all is a bit easier in index notation: write $u^1 = u, u^2 = v, x^1 = x = f(u,v), x^2 = y = g(u,v)$. Then the equations we end up with are of the form

$$\frac{\partial^2 x^k}{\partial u^i \partial u^j} = \sum_{m=1}^{2} \Gamma_{ij}^m \frac{\partial x^k}{\partial u^m}, \tag{19.2.10}$$

where the functions Γ_{ij}^m are called the *Christoffel symbols* and are given by

$$\begin{aligned} \Gamma_{11}^1 &= \frac{GE_u - 2FF_u + FE_v}{2(EG-F^2)} & \Gamma_{11}^2 &= \frac{-FE_u + 2EF_u - EE_v}{2(EG-F^2)} \\ \Gamma_{12}^1 &= \frac{GE_v - FG_u}{2(EG-F^2)} & \Gamma_{12}^2 &= \frac{EG_u - FE_v}{2(EG-F^2)} \\ \Gamma_{22}^1 &= \frac{2GF_v - GG_u - FG_v}{2(EG-F^2)} & \Gamma_{22}^2 &= \frac{EG_v + FG_u - 2FF_v}{2(EG-F^2)}. \end{aligned} \tag{19.2.11}$$

We set $\Gamma_{ij}^m = \Gamma_{ji}^m$.

Now finally we have linear equations (19.2.10) for the unknowns $x^1 = f$ and $x^2 = g$, and the question is whether we can solve them. Actually note that derivatives of f and g satisfy the *same* equation, so it's just a question of whether there is a solution of the equation

$$\frac{\partial^2 f}{\partial u^i \partial u^j} = \sum_{m=1}^{2} \Gamma_{ij}^m \frac{\partial f}{\partial u^m}.$$

We want a compatibility condition that will come from equating mixed partials; right now we're not getting one (just $\Gamma_{ij}^m = \Gamma_{ji}^m$), so we have to differentiate one more time. Let's say with respect to u^k.

$$\begin{aligned} \frac{\partial^3 f}{\partial u^i \partial u^j \partial u^k} &= \sum_{m=1}^{2} \frac{\partial \Gamma_{ij}^m}{\partial u^k} \frac{\partial f}{\partial u^m} + \sum_{m=1}^{2} \Gamma_{ij}^m \frac{\partial^2 f}{\partial u^m \partial u^k} \\ &= \sum_{p=1}^{2} \frac{\partial \Gamma_{ij}^p}{\partial u^k} \frac{\partial f}{\partial u^p} + \sum_{m=1}^{2} \sum_{p=1}^{2} \Gamma_{ij}^m \Gamma_{mk}^p \frac{\partial f}{\partial u^p}. \end{aligned}$$

The equality of mixed partials means that if we rearrange (i,j,k) in any way whatsoever, we'll get the same thing on the left side, and so we should get the same thing on the right side. Now we already know (and it's easy to see) that the order between i and j doesn't matter. So the only interesting term comes from interchanging k

with j, or equivalently interchanging k with i. So the condition is that

(19.2.12)
$$\sum_{p=1}^{2} \frac{\partial \Gamma_{ij}^{p}}{\partial u^k} \frac{\partial f}{\partial u^p} + \sum_{m=1}^{2}\sum_{p=1}^{2} \Gamma_{ij}^{m}\Gamma_{mk}^{p} \frac{\partial f}{\partial u^p} = \sum_{p=1}^{2} \frac{\partial \Gamma_{ik}^{p}}{\partial u^j} \frac{\partial f}{\partial u^p} + \sum_{m=1}^{2}\sum_{p=1}^{2} \Gamma_{ik}^{m}\Gamma_{mj}^{p} \frac{\partial f}{\partial u^p}$$

for all combinations i, j, k. And actually we can write the whole equation as a vector field (for each fixed i, j, k) which acts by differentiating the function f; we set

$$R_{jki}^{p} = \frac{\partial \Gamma_{ij}^{p}}{\partial u^k} - \frac{\partial \Gamma_{ik}^{p}}{\partial u^j} + \sum_{m=1}^{2} \Gamma_{ij}^{m}\Gamma_{mk}^{p} - \sum_{m=1}^{2} \Gamma_{ik}^{m}\Gamma_{mj}^{p}, \tag{19.2.13}$$

and equation (19.2.12) says that

$$\sum_{p=1}^{2} R_{jki}^{p} \frac{\partial f}{\partial u^p} = 0.$$

Recall this equation must be true for both f *and* g since (19.2.10) holds for both f and g. So for each fixed i, j, k we have a vector field which, applied to both coordinate functions, gives zero. That means the vector field itself must be zero. So we obtain finally that

$$R_{jki}^{p} = 0 \qquad \text{for all } i, j, k, p. \tag{19.2.14}$$

This looks like a lot of conditions to satisfy (16 in principle), but it turns out there's really only one condition (as we expect heuristically), because R ends up having a lot of symmetries. The tensor R is called the *Riemann curvature tensor*; many authors also define what I wrote to be the *negative* of the curvature tensor, and the differential geometry community has never agreed which sign is correct. It's probably not at all obvious that it actually is a tensor (i.e., that you'd have the same formula under a coordinate change), but in principle you could check that, and historically that's what would have been done. The nice thing about this approach is that the exact same technique works just as well in higher dimensions (though the definitions of the Christoffel symbols are different there).

As an example of the type of symmetries that exist in two dimensions, we can (ask Maple to) compute that

$$FR_{121}^{2} + ER_{121}^{1} = 0, \qquad GR_{121}^{1} - FR_{122}^{1} = 0, \qquad FR_{122}^{1} + GR_{122}^{2} = 0.$$

So really if we know for example R_{121}^{2} then we know all the others. So only curvature term that matters is the sectional curvature, normalized by $K = R_{121}^{2}/E$, which becomes, after plugging all the Christoffel symbols (19.2.11) into the curvature formula (19.2.13), the big mess

$$\begin{aligned} K = -\frac{1}{4(EG-F^2)^2}\Big(&2EGE_{vv} - 2F^2E_{vv} + 2EGG_{uu} - 2F^2G_{uu} \\ &- 4EGF_{uv} + 4F^2F_{uv} - EE_vG_v - EG_u^2 - FE_uG_v - GE_uG_u + FG_uE_v \\ &+ 2GE_uF_v - 4FF_uF_v + 2FE_vF_v + 2FF_uG_u - GE_v^2 + 2EF_uG_v\Big) \end{aligned} \tag{19.2.15}$$

Obviously nobody actually wants to work with this formula, but the beauty of it is that it's coordinate-invariant and depends only on the metric components. This formula was Gauss' "Remarkable Theorem." The point of everything in this section

so far has been to prove that if some metric components are given and it turns out that $K \equiv 0$, then there is a new coordinate system in which the Riemannian metric is Euclidean. Conversely if $K \neq 0$ at any point, then no coordinate chart will make the metric Euclidean.

Example 19.2.1. The general formula (19.2.15) is probably a fair bit more complicated than you might expect for something so seemingly natural, geometrically. Practically we almost always have some special case that makes things easier.

The most common situation is when the coordinates are orthogonal, in which case $F \equiv 0$. The curvature formula then reduces to

$$K = \frac{-1}{4E^2G^2}\Big(2EGE_{vv} + 2EGG_{uu} - EE_vG_v - EG_u^2 - GE_uG_u - GE_v^2\Big). \tag{19.2.16}$$

Specializing even further, if it happens that the cross-term F is zero and also that the metric components don't depend explicitly on v (such as for a surface of revolution), then the curvature is

$$K = \frac{-1}{4E^2G^2}\Big(2EGG_{uu} - EG_u^2 - GE_uG_u\Big). \tag{19.2.17}$$

On the other hand if we have a polar representation for the surface, then we can write $E = 1$ and $F = 0$ with $G(u, v)$ some arbitrary positive function. Write $G(u, v) = \alpha(u, v)^2$; then the curvature becomes the much simpler formula

$$K = -\frac{1}{\alpha^4}\Big(2GG_{uu} - G_u^2\Big) = -\frac{\frac{\partial^2\alpha}{\partial u^2}(u, v)}{\alpha(u, v)}. \tag{19.2.18}$$

For example if we have a surface of revolution arising from a curve $(\alpha(u), \beta(u))$ parametrized by arc length, then $E = 1$ and $G = \alpha(u)^2$, so that the curvature simplifies to

$$K = -\alpha''(u)/\alpha(u). \tag{19.2.19}$$

Finally if we have an *isothermal coordinate chart*, in which the metric is such that $F = 0$ and $E(u, v) = G(u, v)$ (with possibly both depending on u and v), then we get

$$K = \frac{-1}{2E^3}\Big(EE_{vv} + EE_{uu} - E_u^2 - E_v^2\Big). \tag{19.2.20}$$

Now let's do some things explicitly. First for Cartesian coordinates we have $E = 1, F = 0, G = 1$, so that obviously $K = 0$ by (19.2.18). For the Euclidean metric in polar coordinates we have $E = 1, F = 0, G = u^2$ and (19.2.19) gives $K = 0$ as expected. For the Euclidean metric in parabolic coordinates we have $E = G = u^2 + v^2$, and (19.2.20) gives $K = 0$. For the Euclidean metric in elliptical coordinates we have $E = G = \sinh^2 u + \sin^2 v$, and again (19.2.20) gives $K = 0$.

For the standard sphere S^2 we have $E = 1$, $F = 0$, and $G = \sin^2 u$, and (19.2.19) gives $K = 1$ (independently of u). So the sphere has *constant positive curvature*. The fact that the curvature is constant means that the local geometry is essentially the same everywhere, which implies (if we didn't know it already) that there are a lot of isometries of the sphere.

Taking advantage of (19.2.19) we see that if $E = 1$, $F = 0$, and $G = \sinh^2 u$, then we get $K = -1$, independently of u. This gives another space with a lot of symmetry; the geometry is called *hyperbolic geometry*. This was historically the first non-Euclidean geometry, and this was not at all the method used to find it. (There

are a lot of different ways of thinking about hyperbolic geometry, corresponding to many different models.) What's a bit strange about this is that hyperbolic geometry does *not* come from a surface of revolution of an arc-length parametrized curve. If it did, we'd have to have $\alpha'(u)^2 + \beta'(u)^2 = 1$ and $\alpha(u) = \sinh u$, but that would imply $\beta'(u)^2 = 1 - \cosh^2 u = -\sinh^2 u$, which is impossible. The geometry is basically the Euclidean plane with a different metric, but because the change from sine to hyperbolic sine switches the curvature from $+1$ to -1, the hyperbolic plane is sometimes called the "pseudosphere." ☺

This was not quite the historical development. Actually the approach was to study the curvature of a surface S (which was of course a subset of $\mathbb{R}^3$) by using the Frenet-Serret formulas for curvature of a curve. You take any point p you're interested in, take any unit vector $\mathbf{s}$ parallel to the surface, and the normal unit vector $\mathbf{n}$ to the surface, and form the plane π spanned by $\mathbf{s}$ and $\mathbf{n}$. The intersection of the plane π with the surface S forms a curve C through p, which has a certain curvature at the point p (found as the magnitude of the second derivative vector when the curve is parametrized by arc length). Then you consider all possible unit vectors $\mathbf{s}$ in the tangent plane and see how the curvature κ depends on $\mathbf{s}$. Now it turns out that $\kappa(\mathbf{s}) = \langle \mathbf{s}, \Pi\mathbf{s} \rangle$ for some symmetric matrix Π, and so the maximum and minimum of κ are the eigenvalues κ_1 and κ_2 which are called the *principal curvatures*. These always occur in perpendicular directions (since eigenvectors of a symmetric matrix are orthogonal), so the curvature and thus the geometry is completely determined by κ_1 and κ_2 at every point.

The principal curvatures will tell you how a surface bends in space (the extrinsic geometry), but they won't characterize the geometry that can be measured from inside the surface. Two spaces might have the same local geometry (i.e., the same metric coefficients after coordinates are changed) but different principal curvatures. For example on the plane, $\kappa_1 = \kappa_2 = 0$, while on the cylinder of radius R, we have $\kappa_1 = \frac{1}{R}$ and $\kappa_2 = 0$. For the sphere of radius R we have $\kappa_1 = \kappa_2 = \frac{1}{R}$. Yet the plane and cylinder are the same in terms of local geometry, while the sphere is genuinely different.

Gauss discovered that while the principal curvatures were not intrinsic geometric invariants, the product of them $K = \kappa_1\kappa_2$ *is* a geometric invariant. He showed this by computing it directly: he parametrized an arbitrary surface in terms of some functions f, g, h, found the metric coefficients E, F, G in terms of those functions, and also computed $K = \kappa_1\kappa_2$ in terms of those functions. Then by simplifying the formula he found that K depended only on E, F, G. This made Gauss very happy, and he named the result his "Theorema Egregium" (Remarkable Theorem). You can imagine how he might have felt if you imagine what you'd have to do to compute K directly from my description in terms of the Fernet-Serret formulas, *then* try to find the cancellation which gets everything in terms of the metric coefficients (when dealing with second-derivatives of everything, of course).

19.3. **The covariant derivative.** We now want to do something that will both allow us to finally differentiate vector fields (and pretty much anything else). In addition it will give us a much nicer invariant formula (in terms of vector fields, not coordinates) for the Riemann curvature tensor.

We still have not defined derivatives of vector fields except for the Lie derivative, and the problem with the Lie derivative $\mathcal{L}_U V$ is that it's not tensorial in either U

or V. We'd expect an actual derivative in the standard sense, which we'll denote by $\nabla_U V$, to be tensorial in U and not tensorial in V. It should be a first-order differential operator in V, so we should have the product rule

$$\nabla_U(fV) = U(f)V + f\nabla_U V$$

for any smooth function f.

To get a clue about what it should be, let's compute the covariant derivative we expect in Cartesian coordinates. Write $U = \sum_i u^i(x^1, \cdots, x^n) \frac{\partial}{\partial x^i}$ and $V = \sum_j v^j(x^1, \cdots, x^n) \frac{\partial}{\partial x^j}$. Then we get

$$\nabla_U V = \sum_{i,j} u^i(x^1, \cdots, x^n) \frac{\partial v^j}{\partial x^i}(x^1, \cdots, x^n) \frac{\partial}{\partial x^j}.$$

Let's see what happens to this in another coordinate system $(y^1, \cdots, y^n)$.

If $U = \sum_k \tilde{u}^k \frac{\partial}{\partial y^k}$ and $V = \sum_l \tilde{v}^l \frac{\partial}{\partial y^l}$, then we have $u^i = \sum_k \frac{\partial x^i}{\partial y^k} \tilde{u}^k$ and $v^j = \sum_l \frac{\partial x^j}{\partial y^l} \tilde{v}^l$. So

$$\begin{aligned}
\nabla_U V &= \sum_{i,j} u^i \frac{\partial v^j}{\partial x^i} \frac{\partial}{\partial x^j} \\
&= \sum_{i,j,k,m} \tilde{u}^i \frac{\partial}{\partial y^i} \left(\frac{\partial x^j}{\partial y^k} \tilde{v}^k \right) \frac{\partial y^m}{\partial x^j} \frac{\partial}{\partial y^m} \\
&= \sum_{i,j,k,m} \tilde{u}^i \tilde{v}^k \frac{\partial^2 x^j}{\partial y^i \partial y^k} \frac{\partial y^m}{\partial x^j} \frac{\partial}{\partial y^m} + \sum_{i,j,k,m} \tilde{u}^i \frac{\partial \tilde{v}^k}{\partial y^i} \frac{\partial x^j}{\partial y^k} \frac{\partial y^m}{\partial x^j} \frac{\partial}{\partial y^m} \\
&= \sum_{i,k,m} \Gamma^m_{ik} \tilde{u}^i \tilde{v}^k \frac{\partial}{\partial y^m} + \sum_{i,j,k} \tilde{u}^i \frac{\partial \tilde{v}^k}{\partial y^i} \frac{\partial}{\partial y^k},
\end{aligned}$$

where

$$\Gamma^m_{ik} = \sum_j \frac{\partial y^m}{\partial x^j} \frac{\partial^2 x^j}{\partial y^i \partial y^k}. \tag{19.3.1}$$

The terms Γ^m_{ik} are the Christoffel symbols that already came up in Section 19.2; to see this, recall by formula (19.2.10) that

$$\frac{\partial^2 x^k}{\partial y^i \partial y^j} = \sum_{m=1}^n \Gamma^m_{ij} \frac{\partial x^k}{\partial y^m}$$

when the x-coordinate system is Euclidean. Thus the Christoffel symbols are obtained by multiplying both sides by the inverse matrix $\frac{\partial y}{\partial x}$.

They are *not* the components of a tensor of type $(2, 1)$, despite the notation. The Christoffel symbols are zero for Cartesian coordinates and nonzero for most other coordinate systems, even on Euclidean space. If a tensor's components are zero in one coordinate system, then they must be zero in *all* coordinate systems. They are not coordinate-invariant objects; they depend on the Euclidean metric. If we are willing to incorporate Christoffel symbols, then we can define the covariant derivative this way, but this only works on Euclidean spaces.

But we want a definition that will work for any Riemannian manifold. Notice that we needed to incorporate the Riemannian metric in Euclidean space to get this to work: we ended up with a formula that's simple in Cartesian coordinates

and more complicated in other coordinates. So we're going to find some invariant-looking properties of the covariant derivative we want.

The first thing is that in Euclidean space we have, for any vector fields U and V that

$$\nabla_U V - \nabla_V U = \sum_i \sum_j \left(u^i \frac{\partial v^j}{\partial x^i} - v^i \frac{\partial u^j}{\partial x^i} \right) \frac{\partial}{\partial x^j}.$$

We recognize the right-hand side as the formula for the Lie bracket $[U, V]$ in coordinates. In fact the right-hand side doesn't depend on choice of coordinates, by Proposition 14.5.4. So it seems perfectly reasonable to demand

$$\nabla_U V - \nabla_V U = [U, V]. \tag{19.3.2}$$

This isn't enough though. Notice that if you took *any* functions ζ_{ij}^k which satisfied $\zeta_{ij}^k = \zeta_{ji}^k$, and declared

$$\nabla_U V = \sum_{i,j} u^i \frac{\partial v^j}{\partial x^i} \frac{\partial}{\partial x^j} + \sum_{i,j,k} \zeta_{ij}^k u^i v^j \frac{\partial}{\partial x^k},$$

then we would have $\nabla_U V - \nabla_V U = [U, V]$. So we need some more properties to completely determine $\nabla_U V$.

Consider any three vector fields U, V, W. Write $\langle V, W \rangle = g(V, W)$ to simplify the notation a bit. Then $\langle V, W \rangle$ is a function on the manifold, and $U(\langle V, W \rangle)$ makes sense and is another function. The operation $(U, V, W) \mapsto U(\langle V, W \rangle)$ is linear (over constants) and is obviously tensorial in U. It is also a first-order differential operator in V and W, since it satisfies the product rule

$$U(\langle fV, W \rangle) = U\big(f \langle V, W \rangle\big) = U(f)\langle V, W \rangle + fU\big(\langle V, W \rangle\big).$$

So it would make sense to ask that

$$U\big(\langle V, W \rangle\big) = \langle \nabla_U V, W \rangle + \langle V, \nabla_U W \rangle; \tag{19.3.3}$$

Both sides are tensorial in U and satisfy a product rule in V and W separately. Furthermore this formula obviously works out nicely for the Euclidean metric, when $\langle V, W \rangle = \sum_k V^k W^k$. Since we're trying to define $\nabla_U V$ in a coordinate-invariant way, and the left side of (19.3.3) is coordinate-independent, the fact that this formula works is independent of the choice of Cartesian coordinates. (We could verify it in another coordinate system on Euclidean space using the Christoffel symbols from (19.3.1), and if you want some good practice with partial derivative manipulations, you should try it.)

Now we haven't asked for very much, just the absolute basics, but it turns out the two requirements (19.3.2) and (19.3.3) completely determine the covariant derivative in *any* Riemannian manifold. The proof is due to Koszul and is kind of fun.

Proposition 19.3.1. *Suppose M is a Riemannian manifold with Riemannian metric $g = \langle \cdot, \cdot \rangle$. Then there is a unique covariant derivative map on vector fields, $(U, V) \mapsto \nabla_U V$, satisfying the properties*

$$[U, V] = \nabla_U V - \nabla_V U,$$
$$U\big(\langle V, W \rangle\big) = \langle \nabla_U V, W \rangle + \langle V, \nabla_U W \rangle,$$

for any vector fields U, V, W, along with the properties

$$\begin{aligned} \nabla_{fU} V &= f \nabla_U V, \\ \nabla_U (fV) &= U(f) V + f \nabla_U V, \end{aligned} \tag{19.3.4}$$

for any smooth function f and any vector fields U and V.

Proof. The basic idea is to just start computing $\langle \nabla_U V, W \rangle$ using the assumed properties. It doesn't look like it'll work out at first, but it does. Then we conclude that if we know $\langle \nabla_U V, W \rangle$ for every vector field W, then we know exactly what $\nabla_U V$ must be. So let's go.

$$\begin{aligned} \langle \nabla_U V, W \rangle &= U\big(\langle V, W \rangle\big) - \langle V, \nabla_U W \rangle \\ &= U\big(\langle V, W \rangle\big) - \langle V, [U, W] \rangle - \langle V, \nabla_W U \rangle \\ &= U\big(\langle V, W \rangle\big) - \langle V, [U, W] \rangle - W\big(\langle V, U \rangle\big) + \langle \nabla_W V, U \rangle \\ &= U\big(\langle V, W \rangle\big) - \langle V, [U, W] \rangle - W\big(\langle V, U \rangle\big) + \langle [W, V], U \rangle + \langle \nabla_V W, U \rangle \\ &= U\big(\langle V, W \rangle\big) - \langle V, [U, W] \rangle - W\big(\langle V, U \rangle\big) + \langle [W, V], U \rangle + V\big(\langle W, U \rangle\big) \\ &\qquad - \langle W, \nabla_V U \rangle \\ &= U\big(\langle V, W \rangle\big) - \langle V, [U, W] \rangle - W\big(\langle V, U \rangle\big) + \langle [W, V], U \rangle + V\big(\langle W, U \rangle\big) \\ &\qquad - \langle W, [V, U] \rangle - \langle W, \nabla_U V \rangle, \end{aligned}$$

and we end up back where we started with $\langle \nabla_U V, W \rangle$. Which is usually a bad thing, except in this case it means we can pull both terms over to the left side and get the Koszul formula

$$\begin{aligned} \langle \nabla_U V, W \rangle = \frac{1}{2}\Big(& U\big(\langle V, W \rangle\big) - \langle V, [U, W] \rangle - W\big(\langle V, U \rangle\big) \\ & + \langle [W, V], U \rangle + V\big(\langle W, U \rangle\big) - \langle W, [V, U] \rangle \Big). \end{aligned} \tag{19.3.5}$$

For a nice intricate exercise in understanding tensoriality, you can actually prove using this formula that if $\langle \nabla_U V, W \rangle$ is defined by the right-hand side of (19.3.5), then despite appearances it's actually tensorial in U, tensorial in W, and satisfies a product rule in V. So it does what we expect automatically, even without explicitly incorporating those assumptions. Thus we end up with a definition of the covariant derivative in terms of arbitrary vector fields. □

The formula (19.3.5) is rather cute, but we frequently want to do our computations in coordinates. So let's see what happens in a coordinate chart. Let $U = \sum_i u^i \frac{\partial}{\partial x^i}$, $V = \sum_j v^j \frac{\partial}{\partial x^j}$. We get (using tensoriality in U and the product rule in V) that

$$\begin{aligned} \nabla_U V &= \sum_i u^i \nabla_{\frac{\partial}{\partial x^i}} \Big(\sum_j v^j \frac{\partial}{\partial x^j} \Big) \\ &= \sum_i \sum_j u^i \frac{\partial v^j}{\partial x^i} \frac{\partial}{\partial x^j} + \sum_i \sum_j u^i v^j \nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j}. \end{aligned}$$

Now whatever $\nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j}$ is, it's a vector field for each fixed i and j and can therefore also be expressed in the coordinate field basis. At long last, let's give the proper definition of the Christoffel symbols for a general Riemannian manifold.

Definition 19.3.2. Suppose M is a Riemannian manifold, and ∇ is the unique Riemannian covariant derivative determined by Propositon 19.3.1.

Then in any coordinate chart $(x^1, \cdots, x^n)$, the *Christoffel symbols* Γ_{ij}^k are defined by the formula

$$\nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} = \sum_k \Gamma_{ij}^k \frac{\partial}{\partial x^k}. \tag{19.3.6}$$

Since $\left[\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right] = 0$ in any coordinate chart, we have $\Gamma_{ij}^k = \Gamma_{ji}^k$ for any i, j, k.

If U and V are vector fields expressed in coordinates as $U = \sum_i u^i \frac{\partial}{\partial x^i}$ and $V = \sum_j v^j \frac{\partial}{\partial x^j}$, then the covariant derivative in these coordinates is

$$\nabla_U V = \sum_{i,j} u^i \frac{\partial v^j}{\partial x^i} \frac{\partial}{\partial x^j} + \sum_{i,j,k} \Gamma_{ij}^k u^i v^j \frac{\partial}{\partial x^k}. \tag{19.3.7}$$

Proposition 19.3.3. *In a coordinate chart $(x^1, \cdots, x^n)$ on a Riemannian manifold M, with metric g given in components by $g_{ij} = g\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right)$ and having inverse metric $g^{-1} = g^{ij}$, the Christoffel symbols are given by*

$$\Gamma_{ij}^k = \sum_\ell \frac{1}{2} g^{k\ell} \big(\partial_i g_{\ell j} + \partial_j g_{\ell i} - \partial_\ell g_{ij}\big). \tag{19.3.8}$$

Proof. We just have to apply the Koszul formula when $U = \frac{\partial}{\partial x^i}$, $V = \frac{\partial}{\partial x^j}$, and $W = \frac{\partial}{\partial x^k}$. All the Lie brackets are zero, and only the derivatives of the metric matter. We get

$$\Gamma_{ij}^m g_{mk} = \frac{1}{2}\left(\frac{\partial}{\partial x^i} g_{jk} + \frac{\partial}{\partial x^j} g_{ik} - \frac{\partial}{\partial x^k} g_{ij}\right).$$

Now apply the inverse matrix $g^{k\ell}$ to both sides to solve for Γ_{ij}^m, and rename indices to get (19.3.8). □

We should verify this result in the Euclidean case, since we've already been dealing with a different-looking formula (19.3.1). In x-coordinates the metric is δ_{ij}, so that in y-coordinates we have

$$g_{ij} = \left\langle \frac{\partial}{\partial y^i}, \frac{\partial}{\partial y^j} \right\rangle = \sum_m \frac{\partial x^m}{\partial y^i} \frac{\partial x^m}{\partial y^j}.$$

Thus

$$\begin{aligned} \partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij} &= \sum_m \partial_i \big(\tfrac{\partial x^m}{\partial y^j} \tfrac{\partial x^m}{\partial y^k}\big) + \partial_j \big(\tfrac{\partial x^m}{\partial y^i} \tfrac{\partial x^m}{\partial y^k}\big) - \partial_k \big(\tfrac{\partial x^m}{\partial y^i} \tfrac{\partial x^m}{\partial y^j}\big) \\ &= \sum_m 2 \tfrac{\partial^2 x^m}{\partial y^i \partial y^j} \tfrac{\partial x^m}{\partial y^k}. \end{aligned}$$

Plugging the left side into (19.3.8), we get (19.3.1), as expected.

We now know how to differentiate one vector field in the direction of another, and in general you can replace any "directional derivative" in a Euclidean formula with a covariant derivative to get a Riemannian version of that formula. (This is useful when doing physics on manifolds, for example studying fluids on the surface of a sphere.)

There is a different sort of derivative which is closely related and just as useful. It involves differentiating a vector which depends on a parameter. Now if the vector always stays in the same tangent space, there is no problem whatsoever, and we

can differentiate it without even thinking about a Riemannian metric. The only thing to worry about is when the vector's base point moves around as well. So let's suppose we have a curve $\gamma\colon (-\varepsilon,\varepsilon) \to M$ and a vector $V\colon (-\varepsilon,\varepsilon) \to TM$ whose base point is always $\gamma(t)$. We say that V is a *vector field along the curve* γ.

Definition 19.3.4. Suppose V is a vector field along a curve γ. Then there is a unique covariant derivative $\frac{DV}{dt}$ such that

$$\begin{aligned}\frac{D}{dt}\big(f(t)V(t)\big) &= \frac{df}{dt}V(t) + f(t)\frac{DV}{dt},\\ \frac{d}{dt}\langle U(t), V(t)\rangle &= \left\langle \frac{DU}{dt}, V(t)\right\rangle + \left\langle U(t), \frac{DV}{dt}\right\rangle,\end{aligned}$$

and if W is a vector field on M such that $V(t) = W(\gamma(t))$, then

$$\frac{DV}{dt} = \nabla_{\frac{d\gamma}{dt}} W.$$

In coordinates $(x^1,\cdots,x^n)$, where the curve is $\gamma(t) = \big(\gamma^1(t),\cdots,\gamma^n(t)\big)$ and the vector field is $V(t) = \sum_{j=1}^n v^j(t)\,\frac{\partial}{\partial x^j}\big|_{\gamma(t)}$, we have

$$\frac{DV}{dt} = \sum_{i=1}^n \frac{dv^i}{dt}\,\frac{\partial}{\partial x^i}\Big|_{\gamma(t)} + \sum_{i,j,k=1}^n v^j(t)\frac{d\gamma^k}{dt}\Gamma^i_{jk}(\gamma(t))\,\frac{\partial}{\partial x^i}\Big|_{\gamma(t)}. \tag{19.3.9}$$

Uniqueness of this derivative is straightforward from the coordinate formula, and the coordinate formula can be used to establish existence just by checking that the same formula works in any coordinate system.

Using the covariant derivative along a curve, we can compute the second derivative of a curve just by setting $V(t) = \frac{d\gamma}{dt}$. We get

$$\frac{D}{dt}\frac{d\gamma}{dt} = \sum_{i=1}^n \left(\frac{d^2\gamma^i}{dt^2} + \sum_{j,k}\Gamma^i_{jk}(\gamma(t))\,\frac{d\gamma^j}{dt}\,\frac{d\gamma^k}{dt}\right)\,\frac{\partial}{\partial x^i}\Big|_{\gamma(t)}.$$

Similarly we can compute derivatives of all orders of a curve. It is important just to note that it depends on the Riemannian metric.

Example 19.3.5. As an application, we consider the covariant derivative on the Euclidean space $\mathbb{R}^2$ in polar coordinates. Suppose a curve is described by $\gamma(t) = \big(r(t),\theta(t)\big)$. The metric components are $g_{11} = 1$, $g_{12} = 0$, and $g_{22} = r^2$. From this we compute that the Christoffel symbols are

$$\begin{aligned}\Gamma^1_{22} &= -\frac{1}{2}\frac{\partial}{\partial r}g_{22} = -r\\ \Gamma^2_{12} &= \frac{1}{2r^2}\frac{\partial}{\partial r}g_{22} = \frac{1}{r},\end{aligned}$$

with all other symbols equaling zero. Hence the second derivative in polar coordinates is

$$\frac{D}{dt}\frac{d\gamma}{dt} = \left(\frac{d^2r}{dt^2} - r\left(\frac{d\theta}{dt}\right)^2\right)\,\frac{\partial}{\partial r} + \left(\frac{d^2\theta}{dt^2} + \frac{2}{r}\,\frac{dr}{dt}\,\frac{d\theta}{dt}\right)\,\frac{\partial}{\partial\theta}.$$

(Remember that the Christoffel symbols are symmetric, so we get a factor of two because of Γ^2_{12} and Γ^2_{21}.)

These extra terms in the second derivative are important in physics, since they essentially represent "forces." Newton's equation on a Riemannian manifold is

$$\frac{D}{dt}\frac{d\gamma}{dt} = F\big(\gamma(t)\big)$$

where F is the vector field representing the force. Frequently the force acts only along the radial vector (i.e., a "central force"), so it's proportional to $\frac{\partial}{\partial r}$. So it's natural to write Newton's equations in polar coordinates. When you do this, you end up with equations that look like

$$\begin{aligned}\frac{d^2r}{dt^2} &= r\left(\frac{d\theta}{dt}\right)^2 + F(r)\\ \frac{d^2\theta}{dt^2} &= -\frac{2}{r}\frac{d\theta}{dt}\frac{dr}{dt}.\end{aligned}$$

It looks like we have mysterious extra forces on the right side, like "centrifugal force" and "centripetal force" and such. These aren't actual forces, they're just the Christoffel symbols. Similar sorts of "forces" will show up in any coordinate system where the metric components are not constant (which is typical in any non-Cartesian system).

To make this look slightly more familiar, note that we can solve the second equation for $\frac{d\theta}{dt}$ since it implies that $\frac{d}{dt}(r^2\frac{d\theta}{dt}) = 0$. So $\frac{d\theta}{dt} = \frac{\omega}{r^2}$ for some constant ω (the angular momentum), and then the equation for r becomes

$$\frac{d^2r}{dt^2} = \frac{\omega^2}{r^3} + F(r).$$

☺

The analogue of a straight line in Riemannian geometry is a *geodesic*, which is a curve $\gamma(t)$ such that $\frac{D}{dt}\frac{d\gamma}{dt} = 0$. This is the curve a particle in the manifold will follow when there are no external forces. In Riemannian geometry these are fundamental, and one builds many other concepts in terms of them. Note that the geodesic equation in coordinates is just a second-order differential equation, and thus it is uniquely determined once we specify the initial conditions $\gamma(0)$ and $\gamma'(0)$.

Now finally let's see how to define the Riemann curvature tensor for a general Riemannian manifold. (In Section 19.2 we defined it only for a manifold where the metric was Euclidean after some coordinate change.)

Definition 19.3.6. Suppose M is a Riemannian manifold, with covariant derivative defined as in Proposition 19.3.1. Then the Riemann curvature tensor is a tensor field of type $(3,1)$, i.e., it takes three vector fields and gives another vector field, or you can think of it as taking three vector fields and a covector field and giving you a function. It is defined as

$$R(U,V)W = \nabla_V\nabla_U W - \nabla_U\nabla_V W + \nabla_{[U,V]}W. \tag{19.3.10}$$

To call this a definition, we should of course verify that R is a tensor field, which means that it's tensorial in U, in V, and in W.

Proposition 19.3.7. *The Riemann tensor is a tensor.*

Proof. First check tensoriality in U. We can easily see that $[fU, V] = f[U, V] - V(f)U$. So we have

$$\begin{aligned} R(fU,V)W &= \nabla_V \nabla_{fU} W - \nabla_{fU}\nabla_V W + \nabla_{[fU,V]} W \\ &= \nabla_V (f\nabla_U W) - f\nabla_U \nabla_V W + \nabla_{f[U,V]-V(f)U} W \\ &= V(f)\nabla_U W + f\nabla_V \nabla_U W - f\nabla_U \nabla_V W + f\nabla_{[U,V]} W - V(f)\nabla_U W \\ &= fR(U,V)W. \end{aligned}$$

Tensoriality in V is obvious since $R(U, V)W = -R(V, U)W$.

Now finally let's check tensoriality in W. We have

$$\begin{aligned} R(U,V)(fW) &= \nabla_V \nabla_U (fW) - \nabla_U \nabla_V (fW) + \nabla_{[U,V]}(fW) \\ &= \nabla_V \big(U(f)W + f\nabla_U W\big) - \nabla_U \big(V(f)W + f\nabla_V W\big) \\ &\qquad + [U,V](f)W + f\nabla_{[U,V]} W \\ &= V(U(f))W + U(f)\nabla_V W + V(f)\nabla_U W + f\nabla_V \nabla_U W \\ &\qquad - U(V(f))W - V(f)\nabla_U W - U(f)\nabla_V W - f\nabla_U \nabla_V W \\ &\qquad + U(V(f))W - V(U(f))W + f\nabla_{[U,V]} W \\ &= fR(U,V)W. \end{aligned}$$

□

You can easily compute what the coefficients R^p_{ijk} of this tensor are in terms of the Christoffel symbols, and you come up with the formula (19.2.13) (of course with 2 replaced by n).

So we already have two different interpretations of the Riemann curvature tensor. One is a compatibility condition for a partial differential equation that sort of comes out of nowhere, while another is an explicit check of the failure of the covariant derivative to be commutative. When studying the details of geodesics one can get another interpretation of the curvature tensor as a correction term of the deviation between nearby geodesics (which in Euclidean space would obviously be linear in the time parameter, but in general may have higher-order dependence on time). The various things you can do with curvature are the subject of an entire course in Riemannian geometry, but this is just a taste.

The curvature tensor has the obvious symmetry $R(U, V)W = -R(V, U)W$, but we can make it look even more symmetric by defining a tensor of order $(4, 0)$ by lowering indices.

Definition 19.3.8. The *index-lowered Riemann curvature tensor* is a tensor field of type $(4, 0)$ defined by

$$R(U,V,W,X) = \langle \nabla_V \nabla_U W - \nabla_U \nabla_V W + \nabla_{[U,V]} W, X \rangle. \tag{19.3.11}$$

Proposition 19.3.9. *The index-lowered curvature tensor has the following symmetries:*

$$R(U,V,W,X) + R(V,U,W,X) = 0 \tag{19.3.12}$$

$$R(U,V,W,X) + R(U,V,X,W) = 0 \tag{19.3.13}$$

$$R(U,V,W,X) - R(W,X,U,V) = 0 \tag{19.3.14}$$

$$R(U,V,W,X) + R(V,W,U,X) + R(W,U,V,X) = 0. \tag{19.3.15}$$

Proof. Equation (19.3.12) is obvious. For equation (19.3.13), the easiest thing to do is to prove $R(U,V,W,W)=0$, since multilinearity does the rest. Note that

$$\begin{aligned}
R(U,V,W,W) &= \langle \nabla_V \nabla_U W - \nabla_U \nabla_V W + \nabla_{[U,V]} W, W\rangle \\
&= V(\langle \nabla_U W, W\rangle) - \langle \nabla_U W, \nabla_V W\rangle - U(\langle \nabla_V W, W\rangle) \\
&\qquad + \langle \nabla_V W, \nabla_U W\rangle + \frac{1}{2}[U,V](\langle W, W\rangle) \\
&= \frac{1}{2}V\big(U(\langle W,W\rangle)\big) - \frac{1}{2}U\big(V(\langle W,W\rangle)\big) + \frac{1}{2}[U,V](\langle W,W\rangle) \\
&= 0
\end{aligned}$$

by definition of the Lie bracket.

Equation (19.3.15) follows from the Bianchi identity

$$R(U,V)W + R(V,W)U + R(W,U)V$$

and the Jacobi identity

$$[[U,V],W] + [[V,W],U] + [[W,U],V] = 0.$$

The Jacobi identity is easy, and the Bianchi identity is proved via

$$\begin{aligned}
R(U,V)W + R(V,W)U + R(W,U)V &= \nabla_V \nabla_U W - \nabla_U \nabla_V W + \nabla_{[U,V]} W \\
&\qquad + \nabla_W \nabla_V U - \nabla_V \nabla_W U + \nabla_{[V,W]} U \\
&\qquad + \nabla_U \nabla_W V - \nabla_W \nabla_U V + \nabla_{[W,U]} V \\
&= -\nabla_V [W,U] - \nabla_W [U,V] - \nabla_U [V,W] \\
&\qquad + \nabla_{[U,V]} W + \nabla_{[V,W]} U + \nabla_{[W,U]} V \\
&= [[U,V],W] + [[V,W],U] + [[W,U],V] \\
&= 0.
\end{aligned}$$

Finally (19.3.14) is surprisingly difficult to prove; it comes from using (19.3.15) four times:

$$\begin{aligned}
0 &= 0+0+0+0 \\
&= R(U,V,W,X) + R(V,W,U,X) + R(W,U,V,X) \\
&\qquad + R(V,W,X,U) + R(W,X,V,U) + R(X,V,W,U) \\
&\qquad + R(W,X,U,V) + R(X,U,W,V) + R(U,W,X,V) \\
&\qquad + R(X,U,V,W) + R(U,V,X,W) + R(V,X,U,W) \\
&= \big[R(U,V,W,X) + R(U,V,X,W)\big] + \big[R(V,W,U,X) + R(V,W,X,U)\big] \\
&\qquad + \big[R(X,U,W,V) + R(X,U,V,W)\big] + \big[R(W,X,V,U) + R(W,X,U,V)\big] \\
&\qquad + 2R(W,U,V,X) - 2R(V,X,W,U).
\end{aligned}$$

All the terms in square brackets are zero by (19.3.13), so what's left is zero as well. □

These symmetries look somewhat mysterious, but they're quite natural if you think about them in the right way. What I'll describe is the way Riemann originally presented the curvature tensor publicly.

Suppose we want to study the Riemannian metric near a point. A natural thing to do is to construct a Taylor expansion of its components. However it's not completely obvious how to do this, since the Taylor expansion depends on

the coordinates we choose. We know we have a lot of freedom to choose convenient coordinates, so what's the best choice? There are two choices, depending on whether we want the lower-order terms of the metric to look more like Cartesian coordinates or like polar coordinates. The Cartesian approach was Riemann's approach; the polar approach was used by Gauss for surfaces and later generalized. First note that given any point p, we can *always* change coordinates so that $g_{ij}(p) = \delta_{ij}$. (Of course we can't necessarily make this true at other points simultaneously; we're just asking to do it one point at a time.) The method for doing so is just to find some constant linear combination of the coordinate basis vectors which is orthonormal at p, then use these as new coordinates (since they all commute). So if we're doing a Taylor expansion of g, we can assume the first term is always the Euclidean metric. Then the question is what the first derivative of g is. It turns out that by doing more coordinate changes, we can always assume that $\frac{\partial g_{ij}}{\partial x^k}\Big|_p = 0$.

Riemann's trick was to take coordinates defined by the geodesics of the manifold. We saw the geodesic equation was second order, so solutions are uniquely determined by a point and vector. If we look at all the geodesics through a particular point, then they are determined just by their initial tangent vectors at that point. So we map $v \in T_pM$ to $\gamma(1)$ where γ is the geodesic with $\gamma(0) = p$ and $\gamma'(0) = v$. This map from T_pM to M is called the *Riemannian exponential map*. It's usually extremely hard to compute explicitly, but it's got nice theoretical properties. On Euclidean space the exponential map looks like $\exp_p(v) = p + v$, so it's basically the identity map. On a general manifold it sends geodesics in a manifold to straight lines in a tangent space, and hence it sort of does the "best possible job" of flattening the manifold.

Since geodesics through p correspond to straight lines in T_pM, the geodesic equation at p must have all Christoffel symbols equal to zero at p. So the metric takes the very nice form that $g_{ij}(p) = \delta_{ij}$ and $\Gamma^m_{ij}(p) = 0$. Now the Christoffel symbols were defined by (19.3.8), and if you play around a bit with this formula you can solve for $\frac{\partial}{\partial x^k} g_{ij}$ in terms of them. So if all Christoffel symbols are zero, then $\frac{\partial g_{ij}}{\partial x^k}$ is zero. Of course this is only true at the point p: we've set it up so that geodesics through p look like straight lines through the origin in T_pM, but there's no reason to expect that geodesics not passing through p will look like straight lines in T_pM.

So the linear approximation of any metric near a point is just the Euclidean metric, in the right coordinates. The quadratic terms cannot be transformed away by a coordinate transformation. In fact since the Riemann curvature tensor is defined by operations involving two derivatives of the metric, we might expect to see something like the Riemann curvature showing up in the quadratic terms. We can compute the curvature in these coordinates, using the fact that $g_{ij} = \delta_{ij}$ and $\Gamma^k_{ij} = 0$ at the point p, to get

$$R_{ijk\ell} = \frac{1}{2}\left(\frac{\partial^2 g_{i\ell}}{\partial x^j \partial x^k} + \frac{\partial^2 g_{jk}}{\partial x^i \partial x^\ell} - \frac{\partial^2 g_{ik}}{\partial x^j \partial x^\ell} - \frac{\partial^2 g_{j\ell}}{\partial x^i \partial x^k}\right).$$

Observe the symmetries in this formula. Again some index trickery will tell us how to solve for $\frac{\partial^2 g_{i\ell}}{\partial x^j \partial x^k}$ in terms of the Riemann curvature. We end up with the Taylor

expansion

$$g = \sum_{i,j} g_{ij} dx^i \otimes dx^j = \sum_{i,j} \delta_{ij} dx^i \otimes dx^j + \frac{1}{3} \sum_{i,j,k,\ell} R_{ijk\ell}(x^k\, dx^i - x^i dx^k) \otimes (x^\ell\, dx^j - x^j\, dx^\ell) + O(x^3).$$

This was how Riemann actually discovered the curvature tensor, although it would clearly be very hard to compute the curvature in general using this Taylor expansion. Thus the roundabout way, which at least makes the computations straightforward if not always easy. Note the appearance of terms like $x^k\, dx^i - x^i\, dx^k$, which you should essentially think of as something like $d\theta^{ik}$ for an angular coordinate θ^{ik}. (Recall that in polar coordinates on the plane, $r^2 d\theta = y\, dx - x\, dy$.)

Notice also, and this is perhaps the most important thing in Riemannian geometry, that if you want to compute *any* measurement in a small neighborhood of a point, you can always use the Euclidean formulas to lowest order; the first correction term will always involve the Riemann curvature. Hence you can define the curvature by for example looking at the sum of the angles in a small triangle whose edges are made of geodesics; this will be π plus a small correction term involving the curvature. Subtract π, divide by the area, and take a limit to get the curvature back out. You can do the same thing measuring the difference in the area of a geodesic circle from to πr^2, or the difference in circumference from $2\pi r$. This sort of thing is why the Riemann curvature tensor ends up being so important in geometry.

We can generalize a fair amount of this. In terms of covariant derivatives, the Riemannian case is fairly special; in general the only properties one needs from a covariant derivative $\nabla_U V$ are the basic tensoriality in U and product rule in V, as in (19.3.4). On a manifold where there is no distinguished Riemannian metric, we certainly shouldn't expect (19.3.3) to hold, and in fact we don't even need (19.3.2) to hold. A general covariant derivative is frequently referred to as a *connection*, the name coming from the fact that it's essentially allowed us to connect different tangent spaces. (Recall the fact that different vectors were in different tangent spaces was the biggest problem with finding some way to differentiate vector fields; you can't even subtract vectors unless they're in the same tangent space.)

20. Orthonormal frames

"Remember, your focus determines your reality."

If we have a Riemannian metric given in coordinates, then we can understand everything in terms of the coordinate basis vector fields $X_j = \frac{\partial}{\partial x^j}$ and their inner products $g_{ij} = \langle X_i, X_j \rangle$. Of course we have $[X_i, X_j] = 0$ since mixed partials of smooth functions commute. We can compute the covariant derivative and curvature tensor and such things just knowing these, since the inner products and Lie brackets of arbitrary vector fields are all we need for those definitions.

For some purposes, however, coordinate computations are less convenient. An alternative is to deal with orthonormal vector fields. These are vector fields E_i (possibly defined only on some open set in the manifold) which satisfy $\langle E_i, E_j \rangle = \delta_{ij}$ everywhere they're defined. Such vector fields are called *an orthonormal frame* or classically (since Cartan) a *moving frame*.

20.1. **Basic properties.** Once we do this, of course, we no longer expect to have $[E_i, E_j] = 0$. Instead $[E_i, E_j] = \sum_k c_{ijk} E_k$ for some functions c_{ijk}. We only know two things about the functions c_{ijk}: first, since $[X, Y] = -[Y, X]$, we have $c_{ijk} = -c_{jik}$. Second, by the Jacobi identity

$$[[X, Y], Z] + [[Y, Z], X] + [[Z, X], Y] = 0,$$

we see that

$$\sum_\ell (c_{ij\ell} c_{\ell km} + c_{jk\ell} c_{\ell im} + c_{ki\ell} c_{\ell jm}) = 0$$

for all i, j, k, m.

It's important to note that we can no longer take advantage of the rule that upper and lower indices combined is an indication that something is invariant. When dealing with orthonormal vector fields, we are using the metric in a crucial way, and it's easy to end up with covectors identified with vectors. (Prior to this we've been making a distinction, which is what's responsible for the importance of the index position.)

Example 20.1.1. On the Euclidean plane in polar coordinates, the coordinate vector fields $X_r = \frac{\partial}{\partial r}$ and $X_\theta = \frac{\partial}{\partial \theta}$ are orthogonal but not orthonormal, since $\langle X_r, X_\theta \rangle \equiv 0$, and $|X_r| \equiv 1$, but $|X_\theta| = r$. However we can rescale these to get an orthonormal basis

$$E_r = X_r \qquad E_\theta = \frac{1}{r} X_\theta.$$

Then we have only one Lie bracket to compute:

$$[E_r, E_\theta] = \frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial}{\partial \theta}\right) - \frac{1}{r}\frac{\partial}{\partial \theta}\frac{\partial}{\partial r} = -\frac{1}{r^2}\frac{\partial}{\partial \theta} = -\frac{1}{r} E_\theta.$$

So $c_{121} = 0$ and $c_{122} = -\frac{1}{r}$. Of course these vector fields are defined everywhere except the origin; in Cartesian coordinates they are given by

$$E_r = \frac{x}{r}\frac{\partial}{\partial x} + \frac{y}{r}\frac{\partial}{\partial y} \qquad \text{and} \qquad E_\theta = \frac{y}{r}\frac{\partial}{\partial x} - \frac{x}{r}\frac{\partial}{\partial y}$$

where $r = \sqrt{x^2 + y^2}$.

On the 2-sphere S^2 we can similarly define $E_\theta = \frac{\partial}{\partial\theta}$ and $E_\phi = \frac{1}{\sin\theta}\frac{\partial}{\partial\phi}$ and get $c_{121} = 0$ and $c_{122} = -\tan\theta$.

On S^3 we have a nice orthonormal basis given by the three vector fields on $\mathbb{R}^4$ restricted to S^3, given by

$$\begin{aligned} E_1 &= -x\frac{\partial}{\partial w} + w\frac{\partial}{\partial x} - z\frac{\partial}{\partial y} + y\frac{\partial}{\partial z}, \\ E_2 &= -y\frac{\partial}{\partial w} + z\frac{\partial}{\partial x} + w\frac{\partial}{\partial y} - x\frac{\partial}{\partial z}, \\ E_3 &= -z\frac{\partial}{\partial w} - y\frac{\partial}{\partial x} + x\frac{\partial}{\partial y} + w\frac{\partial}{\partial z}. \end{aligned}$$

We easily compute that $[E_1, E_2] = -2E_3$, $[E_2, E_3] = -2E_1$, and $[E_3, E_1] = -2E_2$. ☺

The last example on S^3 works out the way it does (with constants for the c_{ijk}) because of the Lie group structure of S^3. We will discuss this in a bit.

The typical way to obtain an orthonormal basis is the Gram-Schmidt process. We suppose we have some vector fields $U_1, \cdots, U_n$ which are not orthonormal. Set $F_1 = U_1$, set $F_2 = U_2 - \frac{\langle U_2, F_1\rangle}{\langle F_1, F_1\rangle}F_1$, and in general set

$$F_{k+1} = U_{k+1} - \sum_{j=1}^{k} \frac{\langle U_{k+1}, F_j\rangle}{\langle F_j, F_j\rangle} F_j$$

for $1 \leq k < n$. The field F_2 is well-defined and smooth whenever $F_1 \neq 0$, while F_3 is well-defined whenever F_2 is defined and nonzero, etc. The set of points where any F_k is zero is closed, so the vector fields F_k are all defined on some open set. Then $\langle F_k, F_j\rangle = 0$ whenever $j < k$, so the vector fields F_j are orthogonal on the open set where they're defined and nonzero. Finally set $E_k = \frac{F_k}{|F_k|}$ for each k to get orthonormal fields.

Example 20.1.2. Suppose we have the metric given by $ds^2 = y^2\,dx^2 + 2xy\,dx\,dy + (1+x^2)\,dy^2$. Find an orthonormal basis. To do this, start with any known basis, say $U_1 = \frac{\partial}{\partial x}$ and $U_2 = \frac{\partial}{\partial y}$, and apply Gram-Schmidt. Set $F_1 = \frac{\partial}{\partial x}$ and compute $\langle F_1, F_1\rangle = y^2$, while $\langle F_1, U_2\rangle = \langle\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\rangle = xy$. So

$$F_2 = \frac{\partial}{\partial y} - \frac{xy}{y^2}\frac{\partial}{\partial x}.$$

We have $|F_1| = y$ and $|F_2|^2 = 1 + x^2 - \frac{2x}{y}(xy) + \frac{x^2}{y^2}y^2 = 1$, so that $E_1 = \frac{1}{y}\frac{\partial}{\partial x}$ and $E_2 = -\frac{x}{y}\frac{\partial}{\partial x} + \frac{\partial}{\partial y}$. ☺

Of course, if we had started with $F_1 = \frac{\partial}{\partial y}$, then we would have gotten a totally different orthonormal basis. It's important to understand how to change from one to another, since we frequently want to define objects using an orthonormal basis, and we need to know it doesn't actually matter which one we picked.

Proposition 20.1.3. *Suppose $\{E_1, \cdots, E_n\}$ and $\{F_1, \cdots, F_n\}$ are two orthonormal frames defined on the same open set. Then there is a unique linear operator P_i^j such that*

$$E_i = \sum_j P_i^j F_j,$$

for all i, and P is orthogonal in the sense that

$$\sum_{k=1}^{n} P_i^k P_j^k = \delta_{ij}$$

for all i and j.

Proof. That P_i^j exists is a consequence of the fact that both sets of vectors form a basis for each vector space. Once we have it, we just compute:

$$\langle E_i, E_j\rangle = \delta_{ij} = \sum_{k,\ell} P_i^k P_j^\ell \langle F_k, F_\ell\rangle = \sum_{k,\ell} P_i^k P_j^\ell \delta_{k\ell} = \sum_k P_i^k P_j^k.$$

□

We can compute the geometric objects like covariant derivatives and curvature in an orthonormal basis. Write $U = \sum_{i=1}^n u^i E_i$ and $V = \sum_{j=1}^n v^j E_j$, and we get

$$\nabla_U V = \sum_{i,j} \nabla_{u^i E_i}(v^j E_j) = \sum_{i,j} u^i E_i(v^j) E_j + u^i v^j \nabla_{E_i} E_j.$$

The term $\nabla_{E_i} E_j$ can be computed using the Koszul formula (19.3.5), and we get

$$\begin{aligned}\nabla_{E_i} E_j &= \sum_{k=1}^n \langle \nabla_{E_i} E_j, E_k\rangle E_k \\ &= \frac{1}{2}\sum_{k=1}^n (c_{kij} + c_{kji} + c_{ijk}) E_k.\end{aligned}$$

Writing $a_{ijk} = \frac{1}{2}(c_{kij} + c_{kji} + c_{ijk})$, we can compute the curvature tensor (19.3.11) as

$$\begin{aligned}R(E_i, E_j, E_k, E_\ell) &= \sum_m \langle \nabla_{E_j}(a_{ikm} E_m), E_\ell\rangle - \langle \nabla_{E_i}(a_{jkm} E_m), E_\ell\rangle + c_{ijm}\langle \nabla_{E_m} E_k, E_\ell\rangle \\ &= E_j(a_{ik\ell}) - E_i(a_{jk\ell}) + \sum_m a_{ikm} a_{jm\ell} - \sum_m a_{jkm} a_{im\ell} + \sum_m c_{ijm} a_{mk\ell}.\end{aligned}$$

20.2. **Lie groups.** The most common way to get an orthonormal basis is when the manifold is a Lie group, with a left-invariant metric. Recall that a Lie group G is a smooth manifold with a group structure such that the group operations are smooth functions. If e is the identity element, then $\mathfrak{g} = T_e G$ is called the *Lie algebra*. We can think of the vectors $v \in \mathfrak{g}$ as vector fields, by defining the *left-invariant vector field* $V_g = (L_g)_*(v)$ for each $g \in G$, where $(L_g)_* \colon \mathfrak{g} \to T_g G$ is the differential of the left translation. Then since L_h is a diffeomorphism for any h, it makes sense to consider the push-forward vector field $(L_h)_\# V$. For any element $g \in G$, we have

$$\begin{aligned}((L_h)_\# V)(g) &= (L_h)_*(V(L_h^{-1}(g))) = (L_h)_*(V(h^{-1}g)) \\ &= (L_h)_*(L_{h^{-1}g})_*(v) = (L_{hh^{-1}g})_*(v) = V(g),\end{aligned}$$

using the definition of push-forward, the definition of V, and the chain rule. So $(L_h)_\# V = V$ for any $h \in G$.

Now recall that whenever η is a diffeomorphism and X and Y are vector fields, we have

$$\eta_\# [X, Y] = [\eta_\# X, \eta_\# Y]$$

by Proposition 14.2.9. This general formula implies that if U and V are left-invariant vector fields, then since $(L_h)_\# U = U$ and $(L_h)_\# V = V$, we must have

$$(L_h)_\#[U,V] = [(L_h)_\# U, (L_h)_\# V] = [U,V].$$

So if U and V are left-invariant, then their Lie bracket $[U,V]$ is also left-invariant. Hence there must be some vector $w \in \mathfrak{g}$ such that $[U,V](g) = (L_g)_*(w)$ for all $g \in G$. We abbreviate this as $[u,v] = w$, and in this way we get an operation on $\mathfrak{g}$ itself, also called the Lie bracket. It satisfies the same properties as the bracket of vector fields:

$$[u,v] = -[v,u], \qquad [[u,v],w] + [[v,w],u] + [[w,u],v] = 0.$$

If we have a basis $\{e_1, \cdots, e_n\}$ of the vector space $\mathfrak{g}$, then we can write

$$[e_i, e_j] = \sum_{k=1}^{n} c_{ijk} e_k$$

just as before. Notice however that the c_{ijk} are constants (as of course they should be, since we're only working at the identity in the group). But that means that even if we think of the left-translated vector fields E_i coming from e_i, then the Lie brackets satisfy $[E_i, E_j] = \sum_k c_{ijk} E_k$, where the functions c_{ijk} are actually just constants.

Hence by far the nicest Riemannian metrics on Lie groups are left-invariant metrics. We get one by just constructing a basis of left-invariant vector fields and declaring it to be orthonormal. This tells us the inner product in general, since we have $U = \sum_i u^i E_i$ and $V = \sum_j v^j E_j$, and thus

$$\langle U, V \rangle = \sum_{i,j} u^i v^j \langle E_i, E_j \rangle = \sum_i u^i v^i.$$

Of course, the fact that the metric looks like this in the orthonormal basis does not mean the metric is Euclidean, since $[E_i, E_j] \neq 0$. In fact clearly the only manifolds for which there is an orthonormal basis satisfying $[E_i, E_j] = 0$ are Euclidean manifolds, since vector fields that commute can be used as coordinates.

There's nothing to prevent us from doing all of the above using *right-translations*. It all works in the same way, although in general you don't get the same specific formulas. In fact to see it's equivalent, note that for any Lie group G, the inverse map $\Phi\colon g \mapsto g^{-1}$ is a diffeomorphism, and we have $\Phi(gh) = (gh)^{-1} = h^{-1}g^{-1} = \Phi(h)\Phi(g)$. So it's an antihomomorphism, i.e., it switches the order of group operations, and hence turns all left-translations into right-translations. So while the actual structure may not be exactly the same, it will be *isomorphic*, so we don't really lose anything by just considering left-translations.

Example 20.2.1. Consider the upper half-plane with group operation $(a,b) \cdot (c,d) = (a + bc, bd)$. The identity element is $(0,1)$. At $(0,1)$ we declare the vectors $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$ to be orthonormal. Now let's compute the left-translation. We have $L_{(a,b)}(x,y) = (a + bx, by)$ so that

$$(L_{(a,b)})_* \left(\tfrac{\partial}{\partial x}\big|_{(0,1)} \right) = b \, \tfrac{\partial}{\partial x}\big|_{(a,b)}$$
$$(L_{(a,b)})_* \left(\tfrac{\partial}{\partial y}\big|_{(0,1)} \right) = b \, \tfrac{\partial}{\partial y}\big|_{(a,b)}.$$

Therefore the left-invariant vector fields are

$$E_1(x,y) = (L_{(x,y)})_* \left(\tfrac{\partial}{\partial x}\big|_{(0,1)} \right) = y \, \tfrac{\partial}{\partial x}\big|_{(x,y)}$$
$$E_2(x,y) = (L_{(x,y)})_* \left(\tfrac{\partial}{\partial y}\big|_{(0,1)} \right) = y \, \tfrac{\partial}{\partial y}\big|_{(x,y)}.$$

Thus we compute that $[E_1, E_2] = -y \frac{\partial}{\partial x} = -E_1$.

The metric comes from the equations

$$\left\langle \frac{\partial}{\partial x}, \frac{\partial}{\partial x} \right\rangle = \frac{1}{y^2} \langle E_1, E_1 \rangle = \frac{1}{y^2}$$
$$\left\langle \frac{\partial}{\partial x}, \frac{\partial}{\partial y} \right\rangle = \frac{1}{y^2} \langle E_1, E_2 \rangle = 0$$
$$\left\langle \frac{\partial}{\partial y}, \frac{\partial}{\partial y} \right\rangle = \frac{1}{y^2} \langle E_2, E_2 \rangle = \frac{1}{y^2}.$$

So in coordinates it's $ds^2 = \frac{dx^2+dy^2}{y^2}$, which is the metric of hyperbolic space.

If we were to compute the right-invariant metric, we would have $R_{(a,b)}(u,v) = (u + va, vb)$. Then we would have

$$F_1 = \frac{\partial}{\partial u}\big|_{(u,v)} \qquad \text{and} \qquad F_2 = u \, \frac{\partial}{\partial u}\big|_{(u,v)} + v \, \frac{\partial}{\partial v}\big|_{(u,v)}.$$

We compute the Lie bracket and obtain $[F_1, F_2] = \frac{\partial}{\partial u}\big|_{(u,v)} = F_1$. Note we get the same type of formula as with left-translations, just with a minus sign. Also if we compute the metric in these coordinates, we obtain

$$ds^2 = du^2 - \frac{2u}{v} \, du \, dv + \frac{1+u^2}{v^2} \, dv^2.$$

Notice that $(x,y)^{-1} = (-x/y, 1/y)$. So the inversion map is $(u,v) = (-x/y, 1/y)$. So we see that

$$\begin{aligned} du^2 - \frac{2u}{v} \, du \, dv + \frac{1+u^2}{v^2} \, dv^2 &= \Big(-\frac{1}{y} \, dx + \frac{x}{y^2} \, dy \Big)^2 + 2x \Big(-\frac{1}{y} \, dx + \frac{x}{y^2} \, dy \Big)\Big(-\frac{1}{y^2} \, dy\Big) \\ &\qquad + \Big(1 + \frac{x^2}{y^2}\Big) \, dy^2 \\ &= \frac{1}{y^2} \, dx^2 - \frac{2x}{y^3} \, dx \, dy + \frac{x^2}{y^4} \, dy^2 + \frac{2x}{y^3} \, dx \, dy \\ &\qquad - \frac{2x^2}{y^4} \, dy^2 + \frac{dy^2}{y^2} + \frac{x^2}{y^4} \, dy^2 \\ &= \frac{dx^2 + dy^2}{y^2}. \end{aligned}$$

So we get an isometry between the two rather different-looking metrics, as expected since left-translations and right-translations are equivalent. ☺

20.3. **Inner products of k-forms.** The easiest way to set up a metric on covector fields (and more generally on k-form fields) is to use orthonormal bases. For example, if $\{E_i\}$ is an orthonormal frame, then we have dual covector field bases α_i defined by

$$\alpha_i(E_j) = \delta_i^j.$$

The most natural thing to do is to define the metric on covector fields by declaring α_i to be an orthonormal basis whenever E_i is an orthonormal basis. For this to

work, of course, we have to check that we get the same definition no matter what orthonormal basis we use. That is, if $\{E_i\}$ is one orthonormal basis and $\{\alpha_i\}$ is its dual basis, and if $\{F_i\}$ is another orthonormal basis with dual basis $\{\beta_i\}$, then $\{\alpha_i\}$ is orthonormal if and only if $\{\beta_i\}$ is orthonormal.

Now using Proposition 20.1.3, we know that

$$F_i = \sum_j P_i^j E_i$$

where $\sum_k P_i^k P_j^k = \delta_{ij}$. From this we derive that

$$\alpha^j(F_i) = \alpha^j\left(\sum_k P_i^k E_k\right) = \sum_k P_i^k \delta_k^j = P_i^j.$$

So $\alpha^j = \sum_i P_i^j \beta^i$. If the β^i are orthonormal, then we have

$$\begin{aligned}\langle \alpha^k, \alpha^\ell\rangle &= \langle \sum_i P_i^k \beta^i, \sum_j P_j^\ell \beta^j\rangle \\ &= \sum_{i,j} P_i^k P_j^\ell \langle \beta^i, \beta^j\rangle \\ &= \sum_j P_j^k P_j^\ell.\end{aligned}$$

Now we know that $\sum_k P_i^k P_j^k = \delta_{ij}$, which says that $P^T P = I$, and that implies that $PP^T = I$, which implies that $\sum_j P_j^k P_j^\ell = \delta^{k\ell}$. Hence if the β^i are orthonormal, then the α^i are as well. So the metric on 1-forms does not depend on the choice of orthonormal basis.

In fact we can generalize this. Notice that the map that takes E_i to α^i is just the lifting map from Definition 19.1.2, in a certain basis: we have $E_i^\flat(E_j) = \langle E_i, E_j\rangle = \delta_j^i = \alpha^i(E_j)$, so that $E_i^\flat = \alpha^i$. However the map $v \mapsto v^\flat$ doesn't depend on any particular basis, just on the metric. That means the same principle works in coordinates. In other words, the map $\flat$ must *always* be an isometry, no matter how we're computing it. So in a coordinate chart we have $\frac{\partial}{\partial x^i}^\flat = \sum_j g_{ij}\, dx^j$, and so

$$g_{ij} = \langle \tfrac{\partial}{\partial x^i}, \tfrac{\partial}{\partial x^j}\rangle = \sum_{k,\ell} g_{ik} g_{j\ell} \langle dx^k, dx^\ell\rangle.$$

Now multiply by the inverse matrices g^{ip} and by g^{jq}, and sum over i and q. We get

$$\begin{aligned}\sum_{i,j} g_{ij} g^{ip} g^{jq} &= \sum_{k,\ell,i,j} g_{ik} g^{ip} g_{j\ell} g^{jq} \langle dx^k, dx^\ell\rangle \\ \sum_j \delta_j^p g^{jq} &= \sum_{k,\ell} \delta_k^p \delta_\ell^q \langle dx^k, dx^\ell\rangle \\ g^{pq} &= \langle dx^p, dx^q\rangle.\end{aligned}$$

So the metric on the 1-forms is

$$\sum_{p,q} g^{pq} \frac{\partial}{\partial x^p} \otimes \frac{\partial}{\partial x^q}.$$

On other k-forms the metric is somewhat more complicated, and because there's no fantastic notation for k-forms when $k > 1$, it just doesn't work out very well in coordinates. However it's very easy in terms of orthonormal vector fields.

Definition 20.3.1. Suppose $\{E_i\}$ is an orthonormal basis of vector fields, and $\{\alpha^i\}$ is the dual basis which is also orthonormal. Then we can define the metric on k-forms by declaring the basis k-forms

$$\alpha^{i_1} \wedge \cdots \wedge \alpha^{i_k}$$

to be orthonormal as well; that is,

(20.3.1)

$$\langle \alpha^{i_1} \wedge \cdots \wedge \alpha^{i_k}, \alpha^{j_1} \wedge \cdots \wedge \alpha^{j_k} \rangle = \delta^{i_1 j_1} \cdots \delta^{i_k j_k} \text{ if } i_1 < \cdots < i_k \text{ and } j_1 < \cdots < j_k.$$

The metric thus obtained does not depend on the initial choice of orthonormal vector fields.

This is not really a definition until we prove the statement made there.

Proposition 20.3.2. *The metric defined by* (20.3.1) *does not depend on choice of orthonormal basis* α^i.

Proof. The general proof is a mess of indices; it's the sort of thing that's not too hard to do for oneself but is really difficult to do for someone else. Thus we'll skip the general proof and just prove it for 2-forms, where all the same issues already appear in a simpler fashion. (The general proof can actually be constructed from this case using induction on the size k of the form.)

So first we write $\alpha^i = \sum_j P^i_j \beta^j$, where P is an orthogonal matrix satisfying both $\sum_k P^k_i P^k_j = \delta_{ij}$ and $\sum_k P^i_k P^j_k = \delta^{ij}$. Suppose that

$$\langle \beta^{i_1} \wedge \beta^{i_2}, \beta^{j_1} \wedge \beta^{j_2} \rangle = \delta^{i_1 j_1} \delta^{i_2 j_2} - \delta^{i_1 j_2} \delta^{i_2 j_1}.$$

We want to prove the same formula for α.

We have

$$\begin{aligned}
\langle \alpha^{k_1} \wedge \alpha^{k_2}, \alpha^{\ell_1} \wedge \alpha^{\ell_2} \rangle &= \sum_{i_1, i_2, j_1, j_2} P^{k_1}_{i_1} P^{k_2}_{i_2} P^{\ell_1}_{j_1} P^{\ell_2}_{j_2} \langle \beta^{i_1} \wedge \beta^{i_2}, \beta^{j_2} \wedge \beta^{j_2} \rangle \\
&= \sum_{i_1, i_2, j_1, j_2} P^{k_1}_{i_1} P^{k_2}_{i_2} P^{\ell_1}_{j_1} P^{\ell_2}_{j_2} (\delta^{i_1 j_1} \delta^{i_2 j_2} - \delta^{i_1 j_2} \delta^{i_2 j_1}) \\
&= \sum_{i_1, i_2} P^{k_1}_{i_1} P^{k_2}_{i_2} P^{\ell_1}_{i_1} P^{\ell_2}_{i_2} - \sum_{i_1, i_2} P^{k_1}_{i_1} P^{k_2}_{i_2} P^{\ell_1}_{i_2} P^{\ell_2}_{i_1} \\
&= \delta^{k_1 \ell_1} \delta^{k_2 \ell_2} - \delta^{k_1 \ell_2} \delta^{k_2 \ell_1}.
\end{aligned}$$

□

As a result there are two distinguished choices of n-form field μ on any open set of a Riemannian manifold where an orthonormal basis is defined. We just take any orthonormal basis $\{\alpha^1, \cdots, \alpha^n\}$ and set $\mu = \pm \alpha^1 \wedge \cdots \wedge \alpha^n$. If the manifold is oriented, then only one of these is compatible with the orientation, and we can use it to globally define the Riemannian volume form.

Definition 20.3.3. If M is an oriented Riemannian manifold, then the *Riemannian volume form* μ is the unique n-form which has unit length and is compatible at every point with the orientation. It satisfies $\mu(E_1, \cdots, E_n) = 1$ whenever $\{E_1, \cdots, E_n\}$ is an oriented orthonormal basis.

Of course if M is not oriented, we may not be able to define a global volume form. However we can typically find an oriented subset of full measure and use the volume form on that subset to compute integrals on the entire manifold.

It's important to know what μ is in coordinates.

Proposition 20.3.4. *Suppose M is an oriented Riemannian manifold with a coordinate chart $(x^1, \cdots, x^n)$ on some open set. Write the metric as*

$$g = \sum_{i,j=1}^{n} g_{ij}(x^1, \cdots, x^n) dx^i \otimes dx^j.$$

Then the Riemannian volume form is

$$\mu = \sqrt{\det g}\, dx^1 \wedge \cdots \wedge dx^n. \tag{20.3.2}$$

Proof. Let $F_i = \frac{\partial}{\partial x^i}$, and construct an orthonormal basis E_i using the Gram-Schmidt process; suppose both bases are oriented. Then we have $F_i = \sum_j P_i^j E_j$ for some matrix P_i^j. Since $\langle F_i, F_j \rangle = g_{ij}$ and $\langle E_k, E_\ell \rangle = \delta_{k\ell}$, we have

$$g_{ij} = \langle F_i, F_j \rangle = \sum_{k,\ell} P_i^k P_j^\ell \langle E_k, E_\ell \rangle = \sum_k P_i^k P_j^k.$$

Hence as matrices we have $g = P^T P$. So we have $(\det P)^2 = \det g$.

Now on the other hand, the volume form satisfies

$$\mu(F_1, \cdots, F_n) = (\det P)\mu(E_1, \cdots, E_n) = \det P.$$

Hence $\mu(F_1, \cdots, F_n) = \sqrt{\det g}$. □

Using the Riemannian volume form, we can define the divergence of a vector field. Physically, the divergence is a quantity that tells you how much of whatever the vector field represents (fluid, current, force) is moving into or out of any given volume. Mathematically it is defined in terms of the volume form using Stokes' Theorem.

Definition 20.3.5. Suppose M is a Riemannian manifold with Riemannian volume form μ. Suppose V is any vector field. Then the *divergence of V* is a function $\operatorname{div} V$ defined by

$$d(\iota_V \mu) = (\operatorname{div} V)\, \mu, \tag{20.3.3}$$

where the inner product $\iota_V \mu$ is an $(n-1)$-form defined by

$$\iota_V \mu(W_1, \cdots, W_{n-1}) = \mu(V, W_1, \cdots, W_{n-1}).$$

Stokes' Theorem 17.3.1 for the $(n-1)$-form $\iota_u \mu$ thus takes the form of the divergence theorem: if M is the image of an n-chain c, then

$$\int_M \operatorname{div} V\, \mu = \int_{\partial M} \iota_V \mu. \tag{20.3.4}$$

The term on the left is $d(\iota_V \mu)$, so this is just Stokes' Theorem. For the term on the right, note that to do the integral we compute $(\partial c)^*(\iota_V \mu)$. So take any basis $w_1, \cdots, w_{n-1}$ of $\mathbb{R}^{n-1}$ and plug in:

$$(\partial c)^*(\iota_V \mu) = \mu(V, c_* w_1, \cdots, c_* w_{n-1}).$$

Now the vectors $c_* w_i$ are all in $T_p M$, so that only the component of V in the normal direction contributes; that is, if $\mathbf{n}$ is the unit normal vector to ∂M at p, then

$$\mu(V, c_* w_1, \cdots, c_* w_{n-1}) = \langle V, \mathbf{n} \rangle \mu(\mathbf{n}, c_* w_1, \cdots, c_* w_{n-1}).$$

Hence denoting the *surface area element* by the $(n-1)$-form $dS = \iota_{\mathbf{n}}\mu$, we get

$$\int_M \operatorname{div} V \, \mu = \int_{\partial M} \langle V, \mathbf{n} \rangle \, dS,$$

which is the standard form of the divergence theorem you'd see in vector calculus.

Finally we define the Hodge star operator, which is the second most useful isometry in Riemannian geometry, after the musical isomorphisms between vector fields and covector fields.

Definition 20.3.6. Let $\Omega^k(M)$ denote the space of k-form fields on M. Then the *Hodge star operator* is the map $\star\colon \Omega^k(M) \to \Omega^{n-k}(M)$ such that

$$\alpha \wedge \star\beta = \langle \alpha, \beta \rangle \, \mu \tag{20.3.5}$$

for every pair of k-forms α and β.

Proposition 20.3.7. *The Hodge star operator exists and is unique.*

Proof. This is one of those operations where the easiest thing to do is to say what it does to a basis; but if we do that, we need to show that it doesn't depend on which choice of basis we use. A direct proof of that would be pretty nightmarish (even more complicated than the proof of Proposition 20.3.2) since we hardly even understand inner products of k-forms in a general basis. (We've only been using orthonormal bases so far.) Instead we'll use a trick which gets used again and again: first prove that if we *had* such an operation satisfying all the desired properties, it would have to be unique. Then define it however you want and check that it satisfies the desired properties. Now you've established that there is one, and shown that any two must be the same.

So first let us prove uniqueness; we just have to show that if $\star_1$ and $\star_2$ are two operations with

$$(\langle \alpha, \beta \rangle) \, \mu = \alpha \wedge \star_1 \beta = \alpha \wedge \star_2 \beta$$

for all k-forms α and β, then $\star_1\beta = \star_2\beta$ for all k-forms β. By linearity, this comes down to showing that if $\alpha \wedge (\star_1\beta - \star_2\beta) = 0$ for every k-form α, then the $(n-k)$-form $\star_1\beta - \star_2\beta$ must be zero. This is a general fact about forms which has nothing to do with inner products.

Lemma 20.3.8. *If γ is an $(n-k)$-form field with $\alpha \wedge \gamma = 0$ for every k-form field α, then $\gamma = 0$ everywhere.*

Proof. By the proof of Proposition 4.3.8, we know that for any basis $\{\alpha^1, \cdots, \alpha^n\}$ of T^*M, we can express

$$\gamma = \sum_{i_1 < \cdots < i_{n-k}} \gamma_{i_1 \cdots i_{n-k}} \alpha^{i_1} \wedge \cdots \wedge \alpha^{i_{n-k}}$$

for some functions $\{\gamma_{i_1 \cdots i_{n-k}}\}$. Now $\gamma = 0$ if and only if all of these coefficients are zero. So assume that $\gamma_{i_1 \cdots i_{n-k}} \neq 0$ for some particular choice of $(n-k)$ indices $(i_1, \cdots, i_{n-k})$. Let $\{j_1, \cdots, j_k\}$ be the elements of $\{1, \cdots, n\} \backslash \{i_1, \cdots, i_{n-k}\}$, ordered so that $j_1 < \cdots < j_k$, and let $\alpha = \alpha^{j_1} \wedge \cdots \wedge \alpha^{j_k}$. Then

$$\alpha \wedge \gamma = (\alpha^{j_1} \wedge \cdots \wedge \alpha^{j_k}) \wedge (\gamma_{i_1 \cdots i_{n-k}} \, \alpha^{i_1} \wedge \cdots \wedge \alpha^{i_{n-k}}) = \pm \gamma_{i_1 \cdots i_{n-k}} \, \alpha^1 \wedge \cdots \wedge \alpha^n \neq 0,$$

a contradiction. □

So Lemma 20.3.8 establishes that if there are any operators $\star$ at all satisfying (20.3.5), then there is only one. To show there is one, we just define it on a particular basis. Let $\{\alpha^1, \cdots, \alpha^n\}$ be an oriented orthonormal basis of V^*, and set

$$\star\alpha^{i_1} \wedge \cdots \wedge \alpha^{i_k} = \operatorname{sgn}(j_1 \cdots j_{n-k} i_1 \cdots i_k)\alpha^{j_1} \wedge \cdots \wedge \alpha^{j_{n-k}}, \tag{20.3.6}$$

where the indices $\{j_1, \cdots, j_{n-k}\}$ are chosen so that $\{i_1, \cdots, i_k, j_1, \cdots, j_{n-k}\} = \{1, \cdots, n\}$, ordered $j_1 < \cdots < j_{n-k}$. Extend $\star$ by linearity. We see that $\star$ takes an orthonormal basis of $\Omega^k(M)$ to an orthonormal basis of $\Omega^{n-k}(M)$ by Proposition 20.3.2, and the equation 20.3.5 follows. □

We can easily extend $\star$ to an operator between 0-forms (functions) and n-forms (volume elements) by setting $\star 1 = \mu$ and $\star\mu = 1$. This is already implicit in (20.3.5) using the definition of the wedge product for 0-forms.

This operator is very easy to understand once you see it in action; in fact in some sense we've been dancing around it throughout our discussion of forms.

Example 20.3.9 (The Hodge star on M^2 and M^3)**.** Consider a two-dimensional manifold M, and consider an orthonormal basis of vector fields $\{E_1, E_2\}$ with dual basis $\{\alpha^1, \alpha^2\}$, so that $\mu = \alpha^1 \wedge \alpha^2$. Then the self-map $\star\colon \Omega^1(M) \to \Omega^1(M)$ must satisfy $\star\alpha^1 = \alpha^2$ and $\star\alpha^2 = -\alpha^1$. Thus we clearly have $\star\star = -1$; it corresponds loosely to "rotation by 90°." We get a more traditional version of this if we apply it to vector fields: define (in terms of the index-lowering map from Definition 19.1.2) the operator

$$\star V = (\star V^\flat)^\sharp,$$

and we get that

$$\star E_1 = E_2, \qquad \star E_2 = -E_1.$$

Now consider a three-dimensional manifold M and choose an orthonormal basis $\{E_1, E_2, E_3\}$ with dual basis $\{\alpha^1, \alpha^2, \alpha^3\}$ so that $\mu = \alpha^1 \wedge \alpha^2 \wedge \alpha^3$. Then

$$\star\alpha^1 = \alpha^2 \wedge \alpha^3, \qquad \star\alpha^2 = \alpha^3 \wedge \alpha^1, \qquad \star\alpha^3 = \alpha^1 \wedge \alpha^2$$
$$\star\alpha^1 \wedge \alpha^2 = \alpha^3 \qquad \star\alpha^3 \wedge \alpha^1 = \alpha^2 \qquad \star\alpha^2 \wedge \alpha^3 = \alpha^1.$$

As a result $\star\star = 1$ on 1-forms, hence also on 2-forms.

There is no operation of rotating vectors in a 3-dimensional space, since vectors map to 2-forms. However we can define another operation that only makes sense in three dimensions: the cross product. It's defined on a general manifold as follows: in terms of the musical isomorphisms from Definition 19.1.2, set

$$U \times V = [\star(U^\flat \wedge V^\flat)]^\sharp.$$

Notice that it satisfies antisymmetry obviously, and also that

$$\langle U \times V, U\rangle = \langle \star(U^\flat \wedge V^\flat), U^\flat\rangle = U^\flat \wedge \star\star\,(U^\flat \wedge V^\flat) = U^\flat \wedge U^\flat \wedge V^\flat = 0.$$

So as expected, the cross product is perpendicular to each of the vectors.

☺

These computations suggest the following proposition.

Proposition 20.3.10. *On an n-dimensional manifold M, the Hodge star operator has the following self-inverse property:*

$$\star\star\,\alpha = (-1)^{k(n-k)}\alpha \tag{20.3.7}$$

for any k-form α.

In addition $\star$ is an isometry between $\Omega^k(M)$ and $\Omega^{n-k}(M)$.

Proof. That $\star$ is an isometry is obvious from the fact that it takes an orthonormal basis of its domain to an orthonormal basis of its range.

For the doubling property (20.3.7), we just have to use a particular orthonormal basis. We know that if β is a basis element of $\Omega^k(M)$, then $\star\beta$ is a basis element of $\Omega^{n-k}(M)$, and therefore by formula (20.3.6) we know $\star\star\,\beta = \pm\beta$. So we just have to determine the sign. That comes from the isometry property and (4.3.4):

$$\langle \beta, \beta\rangle\,\mu = \langle \star\beta, \star\beta\rangle\,\mu = \star\beta \wedge \star\star\,\beta = (-1)^{k(n-k)}\star\star\,\beta \wedge \star\beta = (-1)^{k(n-k)}\langle \star\star\,\beta, \beta\rangle\,\mu.$$

□

20.4. **Vector calculus on functions and vector fields.** We have seen that all vector calculus operations seem to have a more natural analogue in terms of wedge products and the d operator (neither of which depend on a Riemannian metric). Now that we have a metric, the musical isomorphisms from Definition 19.1.2 and the Hodge star operator allow us to get the vector calculus operations in a more traditional appearance.

Definition 20.4.1. Suppose M is an n-dimensional oriented Riemannian manifold, with a Riemannian volume form μ. If $f\colon M \to \mathbb{R}$ is a smooth function, we define the *gradient of f*, $\operatorname{grad} f$, to be the vector field

$$(\operatorname{grad} f) = (df)^\sharp. \tag{20.4.1}$$

If V is a vector field on M, we define the *divergence of V* to be the function $\operatorname{div} V$ satisfying

$$\operatorname{div} V = \star d(\star V^\flat). \tag{20.4.2}$$

The composition of these two operators on a function f is another function, called the *Laplacian of f*, given by $\Delta f = \operatorname{div}\operatorname{grad} f$.

Note that we already defined the divergence in (20.3.3), so we should check that we get the same result.

Proposition 20.4.2.

$$\big(\star\, d(\star V^\flat)\big)\mu = d(\iota_V \mu).$$

Proof. By definition of $\star$ on n-forms, we have $(\star f)\mu = f$ whenever f is an n-form. So we just need to check that $d(\star V^\flat) = d(\iota_V\mu)$, and it's sufficient to check that $\star V^\flat = \iota_V\mu$. To see this, consider an orthonormal frame $\{E_1, \cdots, E_n\}$, and write $V = \sum_i v^i E_i$. Then $V^\flat = \sum_i v^i \alpha^i$, and

$$\begin{aligned}
\star V^\flat &= \sum_i (-1)^{i+1} v^i \alpha^1 \cdots \wedge \widehat{\alpha^i} \cdots \wedge \alpha^n \\
&= \sum_i \alpha^i \iota_{E_i} \alpha^1 \wedge \cdots \wedge \alpha^n \\
&= \iota_V \mu.
\end{aligned}$$

□

Let's see what these operations look like explicitly. First we'll do it in coordinates, then in an orthonormal frame. For the coordinate version, recall that the musical isomorphisms are given by

$$(dx^i)^\sharp = \sum_j g^{ij}\,\frac{\partial}{\partial x^j} \qquad \text{and} \qquad \Big(\frac{\partial}{\partial x^i}\Big)^\flat = \sum_j g_{ij}\,dx^j,$$

where g_{ij} is the metric on TM and g^{ij} are the components of the inverse of g.

Proposition 20.4.3. *In a coordinate chart $(x^1, \cdots, x^n)$ on a Riemannian manifold M with metric $ds^2 = \sum_{i,j} g_{ij} dx^i \otimes dx^j$, the gradient of a function is*

$$\operatorname{grad} f = \sum_{i,j} g^{ij}\,\frac{\partial f}{\partial x^i}\,\frac{\partial}{\partial x^j}. \tag{20.4.3}$$

The divergence is given by either of the formulas

$$\operatorname{div} V = \sum_i \frac{1}{\sqrt{\det g}}\,\frac{\partial}{\partial x^i}\big(\sqrt{\det g}\,v^i\big) \tag{20.4.4}$$

or

$$\operatorname{div} V = \sum_i \frac{\partial v^i}{\partial x^i} + \sum_{i,j} \Gamma^j_{ij} v^i. \tag{20.4.5}$$

Proof. The gradient is easy; we have

$$\operatorname{grad} f = (df)^\sharp = \Big(\sum_i \frac{\partial f}{\partial x^i}\,dx^i\Big)^\sharp = \sum_i \frac{\partial f}{\partial x^i}(dx^i)^\sharp = \sum_{i,j} \frac{\partial f}{\partial x^i} g^{ij}\,\frac{\partial}{\partial x^j}.$$

For the divergence, first recall from Proposition 20.3.4 that the Riemannian volume form is given by

$$\mu = \sqrt{\det g}\,dx^1 \wedge \cdots \wedge dx^n.$$

Thus

$$\iota_V \mu = \sum_i (-1)^{i+1} v^i \sqrt{\det g}\,dx^1 \wedge \cdots \wedge \widehat{dx^i} \wedge \cdots \wedge dx^n,$$

and so

$$\begin{aligned} d(\iota_V \mu) &= \sum_i (-1)^{i+1}\,\frac{\partial}{\partial x^i}\big(\sqrt{\det g} v^i\big) dx^i \wedge dx^1 \wedge \cdots \wedge \widehat{dx^i} \wedge \cdots \wedge dx^n \\ &= \sum_i \frac{\partial}{\partial x^i}\big(\sqrt{\det g} v^i\big) dx^1 \wedge \cdots \wedge dx^n. \end{aligned}$$

Formula (20.4.4) follows.

For (20.4.5) we just use the product rule on (20.4.4). We obtain

$$\operatorname{div} V = \sum_i \frac{\partial v^i}{\partial x^i} + \sum_i \frac{v^i}{2 \det g}\,\frac{\partial \det g}{\partial x^i}. \tag{20.4.6}$$

Now for any time-dependent matrix $A(t)$ whatsoever we have the general formula

$$\frac{d}{dt} \det A(t) = \operatorname{Tr}\left(A(t)^{-1}\frac{dA}{dt}\right) \det A(t).$$

Thus the coefficient of the second term in (20.4.6) is

$$\frac{1}{2\det g}\frac{\partial \det g}{\partial x^i} = \frac{1}{2}\sum_j g^{jk}\frac{\partial g_{jk}}{\partial x^i}.$$

Recalling from (19.3.8) that

$$\Gamma^k_{ij} = \frac{1}{2}\sum_\ell g^{k\ell}(\partial_i g_{\ell j} + \partial_j g_{\ell i} - \partial_\ell g_{ij}),$$

we see that

$$\sum_j \Gamma^j_{ij} = \frac{1}{2}\sum_{j,\ell} g^{j\ell}(\partial_i g_{\ell j} + \partial_j g_{\ell i} - \partial_\ell g_{ij}),$$

so we get what we want if we show that

$$\sum_{j,\ell} g^{j\ell}\partial_j g_{\ell i} = \sum_{j,\ell} g^{j\ell}\partial_\ell g_{ij}.$$

This is true since we can just interchange j and ℓ in the sum on the left to get the sum on the right, since $g^{j\ell} = g^{\ell j}$. □

The formula (20.4.4) is almost always easier to apply in practice than (20.4.5), since there's only one geometric computation to do (the volume form coefficient $\sqrt{\det g}$) rather than a bunch of Christoffel symbols. However (20.4.5) suggests that the divergence is a kind of "average" covariant derivative, which can be useful conceptually. This is made more explicit in the orthonormal frames version.

Proposition 20.4.4. *Suppose M is a Riemannian manifold and $\{E_1, \cdots, E_n\}$ is an orthonormal basis of vector fields. Then the gradient is given by*

$$\operatorname{grad} f = \sum_i E_i(f)\, E_i, \tag{20.4.7}$$

and the divergence is given by

$$\operatorname{div} V = \sum_i \langle \nabla_{E_i} V, E_i\rangle \tag{20.4.8}$$

or

$$\operatorname{div} V = \sum_i E_i(v^i) + \sum_{i,j} v^j \langle [E_i, E_j], E_i\rangle. \tag{20.4.9}$$

Proof. Since $df(E_i) = E_i(f)$ by definition, we have

$$df = \sum_i E_i(f)\alpha^i,$$

where $\alpha^i(E_j) = \delta^i_j$. By definition of the metric on 1-forms, if the E_i are orthonormal then the α^i are orthonormal as well, and we have $(\alpha^i)^\sharp = E_i$. Thus

$$\operatorname{grad} f = \sum_i E_i(f) E_i.$$

The proof of (20.4.8) is a bit subtler. Recall that by (19.3.7) we have for any vector field U that

$$\nabla_U V = \sum_{i,j} u^i \frac{\partial v^j}{\partial x^i}\frac{\partial}{\partial x^j} + \sum_{i,j,k}\Gamma^j_{ik} u^i v^k \frac{\partial}{\partial x^j}.$$

Now fix V and consider the operator $D_p\colon T_pM \to T_pM$ given by $U \mapsto \nabla_U V$. Since the covariant derivative is tensorial in U, the operator D_p really depends only on the point p. Any operator from a vector space to itself has a trace which is independent of any choice of basis, by Proposition 3.3.1. Now the components of D in the coordinate basis $\{\frac{\partial}{\partial x^i}\}$ are given by

$$D_i^j = \frac{\partial v^j}{\partial x^i} + \sum_k \Gamma_{ik}^j v^k.$$

The trace is obtained by just setting $i = j$ and summing over i, so we get by formula (20.4.5) that

$$\operatorname{div} V = \sum_i D_i^i = \frac{\partial v^i}{\partial x^i} + \sum_{i,k} \Gamma_{ik}^i v^k.$$

Now to prove (20.4.8) we just compute the trace of the operator D in the basis $\{E_i\}$ instead of the basis $\{\frac{\partial}{\partial x^i}\}$. We have

$$D(E_i) = \nabla_{E_i} V = \sum_j \langle \nabla_{E_i} V, E_j \rangle E_j,$$

so the components of D in this basis are $\tilde{D}_i^j = \langle \nabla_{E_i} V, E_j \rangle$. Hence the trace of this operator D is

$$\operatorname{div} V = \sum_i \tilde{D}_i^i = \sum_i \langle \nabla_{E_i} V, E_i \rangle,$$

so we obtain (20.4.8).

To get (20.4.9) we write $V = \sum_j v^j E_j$, and obtain

$$\begin{aligned}
\operatorname{div} V &= \sum_{i,j} \langle \nabla_{E_i}(v^j E_j), E_i \rangle \\
&= \sum_{i,j} \langle E_i(v^j) E_j + v^j \nabla_{E_i} E_j, E_i \rangle \\
&= \sum_i E_i(v^i) + \sum_{i,j} v^j \langle [E_i, E_j], E_i \rangle + \sum_{i,j} \langle \nabla_{E_j} E_i, E_i \rangle,
\end{aligned}$$

and the last term is zero since $\langle \nabla_{E_j} E_i, E_i \rangle = \frac{1}{2} E_j(\langle E_i, E_i \rangle) = \frac{1}{2} E_j(1) = 0$. □

It is worth noting that although I defined the divergence using a Riemannian volume form, the existence of which requires an orientation, the divergence makes sense without this. The reason is that in the formula

$$d(\iota_V \mu) = \operatorname{div} V \, \mu,$$

if we flip the sign of μ on both sides, the formula still works. The interpretation of the divergence as a trace of the covariant derivative makes it obvious that orientability is irrelevant.

Also notice that we end up with terms like $\langle [E_i, E_j], E_i \rangle = c_{iji}$ in formula (20.4.9), which is not surprising since that's where all the geometry is in an orthonormal basis.

Now the curl is more complicated since it depends on the dimension n of the manifold. We want to take the derivative of a vector field V which corresponds to a 1-form α. We have seen that d does the right sort of twisting, but of course $d\alpha$ is a 2-form. If $n = 1$ we get zero (there's no antisymmetric derivative in one dimension). If $n = 2$ we get a volume form, and using the Hodge star operator we

can get a more familiar 0-form out of this. If $n = 3$ we get a 2-form, and the Hodge star operator gives a 1-form, which we can then lift to get a vector field back. If $n = 4$ then the Hodge star just takes 2-forms to 2-forms, and there's no way to get back to functions or vector fields (and it's not just our lack of cleverness; there are six components of a 2-form in M^4, so you just can't match the dimension with anything natural). If $n > 4$ it's even worse. So curl is only going to be interesting in dimension 2 or 3, if our main concern is operators that can be reduced to actions on vector fields or functions and giving back vector fields and functions.

Definition 20.4.5. If M is a 2-dimensional Riemannian manifold and V is a vector field on M, we define the *curl of* V to be the function $\operatorname{curl} V$ given by

$$\operatorname{curl} V = \star dV^\flat, \tag{20.4.10}$$

in terms of the musical isomorphisms $\sharp$ and $\flat$ from Definition 19.1.2.

If M is a 3-dimensional Riemannian manifold and V is a vector field on M, then the *curl of* V is a vector field $\operatorname{curl} V$ given by

$$\operatorname{curl} V = (\star dV^\flat)^\sharp. \tag{20.4.11}$$

As before, we want to compute this explicitly, both in a coordinate system and in an orthonormal frame. The coordinate computation is an absolute mess in the most general case, and this comes from the fact that the Hodge star operator is only simple when it's acting on orthonormal k-forms. In the general case of coordinates it quickly becomes a mess, and 90% of the cases you need in a coordinate system will be in an orthogonal coordinate system anyway, where the metric is

$$ds^2 = h_1^2 dx^2 + h_2^2 dy^2 + h_3^2 dz^2.$$

In the other cases you either have a nice orthonormal basis of vector fields already given, or you should find one.

Proposition 20.4.6. *If M is a 2-dimensional Riemannian manifold with a coordinate system (x, y) in which the metric is*

$$ds^2 = h_1^2 dx^2 + h_2^2 dy^2,$$

then the curl of a vector field $V = u \frac{\partial}{\partial x} + v \frac{\partial}{\partial y}$ is

$$\operatorname{curl} V = \frac{1}{h_1 h_2} \left(\frac{\partial (h_2^2 v)}{\partial x} - \frac{\partial (h_1^2 u)}{\partial y} \right) \tag{20.4.12}$$

If M is a 3-dimensional Riemannian manifold with a coordinate system (x, y, z) in which the metric is

$$ds^2 = h_1^2 dx^2 + h_2^2 dy^2 + h_3^2 dz^2,$$

then for any vector field $V = u \frac{\partial}{\partial x} + v \frac{\partial}{\partial y} + w \frac{\partial}{\partial z}$, then the curl of V is

$$\begin{aligned} \operatorname{curl} V = {} & \frac{1}{h_1 h_2 h_3} \left(\frac{\partial (h_3^2 w)}{\partial y} - \frac{\partial (h_2^2 v)}{\partial z} \right) \frac{\partial}{\partial x} \\ & + \frac{1}{h_1 h_2 h_3} \left(\frac{\partial (h_1^2 u)}{\partial z} - \frac{\partial (h_3^2 w)}{\partial x} \right) \frac{\partial}{\partial y} + \frac{1}{h_1 h_2 h_3} \left(\frac{\partial (h_2^2 v)}{\partial x} - \frac{\partial (h_1^2 u)}{\partial y} \right) \frac{\partial}{\partial z}. \end{aligned} \tag{20.4.13}$$

Proof. In two dimensions it is easy to check from Definition 19.1.2 that we have

$$V^\flat = h_1^2 u\, dx + h_2^2 v\, dy,$$

so that

$$dV^\flat = \left(\frac{\partial}{\partial x}(h_2^2 v) - \frac{\partial}{\partial y}(h_1^2 u) \right) dx \wedge dy.$$

Since the determinant of the metric is $\det g = h_1^2 h_2^2$, the Riemannian volume form is $\mu = h_1 h_2\, dx \wedge dy$. And since $\star\mu = 1$, we know $\star(dx \wedge dy) = \frac{1}{h_1 h_2}$. So

$$\operatorname{curl} V = \star dV^\flat = \frac{1}{h_1 h_2} \left(\frac{\partial}{\partial x}(h_2^2 v) - \frac{\partial}{\partial y}(h_1^2 u) \right),$$

which is (20.4.12).

In three dimensions we have

$$V^\flat = h_1^2 u\, dx + h_2^2 v\, dy + h_3^2 w\, dz.$$

Therefore

$$\begin{aligned} dV^\flat = &\left(\frac{\partial(h_3^2 w)}{\partial y} - \frac{\partial(h_2^2 v)}{\partial z} \right) dy \wedge dz + \left(\frac{\partial(h_1^2 u)}{\partial z} - \frac{\partial(h_3^2 w)}{\partial x} \right) dz \wedge dx \\ &+ \left(\frac{\partial(h_2^2 v)}{\partial x} - \frac{\partial(h_1^2 u)}{\partial y} \right) dx \wedge dy. \end{aligned}$$

To compute the Hodge star operator in this case, we'd like to have an orthonormal basis for $T_p^* M$. Fortunately, this is provided by the 1-form fields $h_1\, dx$, $h_2\, dy$, and $h_3\, dz$. Thus we have $\star dx = \frac{h_2 h_3}{h_1}\, dy \wedge dz$, etc. So we have

$$\begin{aligned} \star dV^\flat = &\frac{h_1}{h_2 h_3} \left(\frac{\partial(h_3^2 w)}{\partial u^2} - \frac{\partial(h_2^2 v)}{\partial u^3} \right) dx + \frac{h_2}{h_1 h_3} \left(\frac{\partial(h_1^2 u)}{\partial z} - \frac{\partial(h_3^2 w)}{\partial x} \right) dy \\ &+ \frac{h_3}{h_1 h_2} \left(\frac{\partial(h_2^2 v)}{\partial x} - \frac{\partial(h_1^2 u)}{\partial y} \right) dz. \end{aligned}$$

Finally, applying the $\sharp$ operator, we get formula (20.4.13). □

As mentioned, since the Hodge star is easy in an orthonormal frame, we expect the curl formula to be easy as well. Actually the formulas in Proposition 20.4.6 are just a special case of those in the following proposition, when $E_1 = \frac{1}{h_1}\frac{\partial}{\partial x}$, $E_2 = \frac{1}{h_2}\frac{\partial}{\partial y}$, and $E_3 = \frac{1}{h_3}\frac{\partial}{\partial z}$.

Proposition 20.4.7. *If M is a two-dimensional Riemannian manifold with orthonormal frame $\{E_1, E_2\}$, then the curl of a vector field $V = v^1 E_1 + v^2 E_2$ is*

$$\operatorname{curl} V = E_2(v_1) - E_1(v^2) - \langle V, [E_1, E_2] \rangle. \tag{20.4.14}$$

If M is a three-dimensional Riemannian manifold with orthonormal frame $\{E_1, E_2, E_3\}$, then the curl of a vector field $V = v^1 E_1 + v^2 E_2 + v^3 E_3$ is

$$\begin{aligned} \operatorname{curl} V = &\big(E_2(v^3) - E_3(v^2)\big) E_1 + \big(E_3(v^1) - E_1(v^3)\big) E_2 \\ &+ \big(E_1(v^2) - E_2(v^1)\big) E_3 - \langle V, [E_2, E_3] \rangle E_1 - \langle V, [E_3, E_1] \rangle E_2 - \langle V, [E_1, E_2] \rangle E_3. \end{aligned} \tag{20.4.15}$$

Proof. To prove (20.4.14), note that $V^\flat = v^1 \alpha^1 + v^2 \alpha^2$, and thus by the product rule (16.3.6) we have

$$dV^\flat = dv^1 \wedge \alpha^1 + dv^2 \wedge \alpha^2 + v^1 d\alpha^1 + v^2 d\alpha^2.$$

We have

$$dv^1 \wedge \alpha^1 = (E_1(v^1)\alpha^1 + E_2(v^1)\alpha^2) \wedge \alpha^1 = -E_2(v^1)\alpha^1 \wedge \alpha^2.$$

Similarly $dv^2 \wedge \alpha^2 = E_1(v^2)\alpha^1 \wedge \alpha^2$. To compute $d\alpha^1$ we use (16.1.2); we have

$$\begin{aligned} d\alpha^1(E_1, E_2) &= E_1\big(\alpha^1(E_2)\big) - E_2\big(\alpha^1(E_1)\big) - \alpha^1([E_1, E_2]) \\ &= E_1(0) - E_2(1) - \langle E_1, [E_1, E_2]\rangle = -\langle E_1, [E_1, E_2]\rangle. \end{aligned}$$

Similarly $d\alpha^2(E_1, E_2) = -\langle E_2, [E_1, E_2]\rangle$. Putting it all together, we obtain (20.4.14).

Now to compute (20.4.15), we note that

$$\begin{aligned} dV^\flat &= d(v^1\alpha^1 + v^2\alpha^2 + v^3\alpha^3) \\ &= dv^1 \wedge \alpha^1 + dv^2 \wedge \alpha^2 + dv^3\alpha^3 \\ &\qquad + v^1 d\alpha^1 + v^2 d\alpha^2 + v^3 d\alpha^3 \\ &= E_2(v^1)\alpha^2 \wedge \alpha^1 + E_3(v^1)\alpha^3 \wedge \alpha^1 + E_1(v^2)\alpha^1 \wedge \alpha^2 \\ &\qquad + E_3(v^2)\alpha^3 \wedge \alpha^2 + E_1(v^3)\alpha^1 \wedge \alpha^3 + E_2(v^3)\alpha^2 \wedge \alpha^3 \\ &\qquad + v^1 d\alpha^1 + v^2 d\alpha^2 + v^3 d\alpha^3. \end{aligned}$$

Since $\star(\alpha^1 \wedge \alpha^2) = \alpha^3$, $\star(\alpha^2 \wedge \alpha^3) = \alpha^1$, and $\star(\alpha^3 \wedge \alpha^1) = \alpha^2$ by Example 20.3.9, we see the second-last and third-last lines give the first line of (20.4.15). So now we just have to compute

$$v^1 \star d\alpha^1 + v^2 \star d\alpha^2 + v^3 \star d\alpha^3.$$

Now for any i, j, k we have

$$d\alpha^k(E_i, E_j) = E_i\big(\alpha^k(E_j)\big) - E_j\big(\alpha^k(E_i)\big) - \alpha^k\big([E_i, E_j]\big) = -\langle E_k, [E_i, E_j]\rangle,$$

so that

$$d\alpha^k = -\langle E_k, [E_1, E_2]\rangle \alpha^1 \wedge \alpha^2 - \langle E_k, [E_2, E_3]\rangle \alpha^2 \wedge \alpha^3 - \langle E_k, [E_3, E_1]\rangle \alpha^3 \wedge \alpha^1,$$

and thus

$$\star d\alpha^k = -\langle E_k, [E_1, E_2]\rangle \alpha^3 - \langle E_k, [E_2, E_3]\rangle \alpha^1 - \langle E_k, [E_3, E_1]\rangle \alpha^2.$$

Plugging in, we get the second line of (20.4.15). □

Again notice that the geometry enters into the formulas via the Lie bracket coefficients $c_{ijk} = \langle [E_i, E_j], E_k\rangle$.

Example 20.4.8. Let's compute these things on S^3. In the spherical coordinate system (ψ, θ, ϕ) on S^3 we have by (19.1.2) that $h_1 = 1$, $h_2 = \sin\psi$, and $h_3 = \sin\psi \sin\theta$. So the curl of an arbitrary vector field

$$V = u\,\frac{\partial}{\partial\psi} + v\,\frac{\partial}{\partial\theta} + w\,\frac{\partial}{\partial\phi}$$

is, using (20.4.13),

$$\begin{aligned} \operatorname{curl} V = &\frac{1}{\sin^2\psi\sin\theta}\left(\frac{\partial(\sin^2\psi\sin^2\theta w)}{\partial\theta} - \frac{\partial(\sin^2\psi v)}{\partial\phi}\right)\frac{\partial}{\partial\psi} \\ &+ \frac{1}{\sin^2\psi\sin\theta}\left(\frac{\partial u}{\partial\phi} - \frac{\partial(\sin^2\psi\sin^2\theta w)}{\partial\psi}\right)\frac{\partial}{\partial\theta} \\ &+ \frac{1}{\sin^2\psi\sin\theta}\left(\frac{\partial(\sin^2\psi v)}{\partial\psi} - \frac{\partial u}{\partial\theta}\right)\frac{\partial}{\partial\phi}. \end{aligned}$$

So for example $\operatorname{curl} \frac{\partial}{\partial \psi} = 0$, $\operatorname{curl} \frac{\partial}{\partial \theta} = \frac{2\cos\psi}{\sin\psi\sin\theta}\,\frac{\partial}{\partial \phi}$, and $\operatorname{curl} \frac{\partial}{\partial \phi} = 2\cos\theta\,\frac{\partial}{\partial \psi} - \frac{2\cos\psi\sin\theta}{\sin\psi}\,\frac{\partial}{\partial \theta}$.

Now on S^3 we also have a nice orthonormal basis $\{E_1, E_2, E_3\}$ satisfying $[E_1, E_2] = -2E_3$, $[E_2, E_3] = -2E_1$, and $[E_3, E_1] = -2E_2$. So formula (20.4.15) gives for $V = v^1E_1 + v^2E_2 + v^3E_3$ that

$$\operatorname{curl} V = \big(E_2(v^3) - E_3(v^2)\big)E_1 + \big(E_3(v^1) - E_1(v^3)\big)E_2 + \big(E_1(v^2) - E_2(v^1)\big)E_3 + 2V.$$

As a consequence we see that $\operatorname{curl} E_k = 2E_k$ for any k. Thus these orthonormal vector fields are actually *curl eigenfields*, which are useful in plasma and fluid dynamics. ☺

Trying to prove a general formula like $\operatorname{curl}\operatorname{grad} f = 0$ from the coordinate formulas in Propositions 10.3.5 and 20.4.6, even in the special case of orthogonal coordinates, is silly. Instead it makes a lot more sense to use general, simple properties of forms. This is why k-forms end up being more fundamental in many ways than the basic objects of vector calculus (despite the fact that, of course, all of these things were first discovered using components of vector fields in coordinates, then generalized to forms later).

Proposition 20.4.9. *We have the following identities:*

$$\operatorname{curl}\operatorname{grad} f \equiv 0 \tag{20.4.16}$$

for any function f in two or three dimensions, and

$$\operatorname{div}\operatorname{curl} V \equiv 0 \tag{20.4.17}$$

for any vector field V in three dimensions.

Proof. These are easy consequences of $d^2 = 0$ and $\star\star = \pm 1$. □

The last thing to discuss is the Laplacian. On functions we have defined $\Delta f = \operatorname{div}\operatorname{grad} f$. In coordinates this ends up being

$$\Delta f = \frac{1}{\sqrt{\det g}}\,\frac{\partial}{\partial x^i}\left(\sqrt{\det g}\,g^{ij}\,\frac{\partial f}{\partial x^j}\right).$$

In Cartesian coordinates on Euclidean $\mathbb{R}^3$ we have the standard formula

$$\Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}.$$

In spherical coordinates on Euclidean $\mathbb{R}^3$ we have $ds^2 = dr^2 + r^2\,d\theta^2 + r^2\sin^2\theta\,d\phi^2$, and thus we read off $g^{11} = 1$, $g^{22} = \frac{1}{r^2}$, $g^{33} = \frac{1}{r^2\sin^2\theta}$, along with $\sqrt{\det g} = r^2\sin\theta$, so that

$$\begin{aligned}\Delta f &= \frac{1}{r^2\sin\theta}\,\frac{\partial}{\partial r}\left(r^2\sin\theta\,\frac{\partial f}{\partial r}\right) + \frac{1}{r^2\sin\theta}\,\frac{\partial}{\partial \theta}\left(\frac{r^2\sin\theta}{r^2}\,\frac{\partial f}{\partial \theta}\right) + \frac{1}{r^2\sin\theta}\,\frac{\partial}{\partial \phi}\left(\frac{r^2\sin\theta}{r^2\sin^2\theta}\,\frac{\partial f}{\partial \phi}\right)\\ &= \frac{1}{r^2}\,\frac{\partial}{\partial r}\left(r^2\,\frac{\partial f}{\partial r}\right) + \frac{1}{r^2\sin\theta}\,\frac{\partial}{\partial \theta}\left(\sin\theta\frac{\partial f}{\partial \theta}\right) + \frac{1}{r^2\sin^2\theta}\,\frac{\partial^2 f}{\partial \phi^2}.\end{aligned}$$

Even with the Euclidean metric, the Riemannian approach gives an easier-to-remember formula than just changing coordinates directly, and it explains all the cancellations that inevitably happen.

One of the most important things about the Laplacian is Green's identity for the Laplacian of a product. It's easy to verify the product rules for gradient and divergence:

$$\operatorname{grad}(\phi\psi) = \phi \operatorname{grad}\psi + \psi \operatorname{grad}\phi,$$
$$\operatorname{div}(\phi V) = \phi \operatorname{div} V + V(\phi),$$

and from these we conclude that

$$\Delta(\phi\psi) = \operatorname{div}(\phi \operatorname{grad}\psi + \psi \operatorname{grad}\phi) = \phi\Delta\psi + \psi\Delta\phi + 2\langle \operatorname{grad}\phi, \operatorname{grad}\psi\rangle.$$

Now notice that by the divergence theorem 20.3.4 that if a compact orientable Riemannian manifold M has no boundary, then the integral of any divergence is zero, and since the Laplacian is the divergence of a gradient, we have

$$\int_M \Delta(\phi\psi)\,\mu = 0.$$

Hence for example when $\phi = \psi$ we get

$$\int_M \phi\Delta\phi\,\mu = -\int_M |\operatorname{grad}\phi|^2\,\mu.$$

This says that the Laplacian is a negative-definite operator on any compact orientable Riemannian manifold; it also implies that the only solution of $\Delta\phi = 0$ is $\phi =$ constant.

Similarly we have

$$\begin{aligned}\int_M \phi\Delta\psi\,\mu &= \int_M \phi \operatorname{div}\operatorname{grad}\psi\,\mu = \int_M \operatorname{div}(\phi \operatorname{grad}\psi)\,\mu - \int_M \langle \operatorname{grad}\phi, \operatorname{grad}\psi\rangle\,\mu \\ &= -\int_M \langle \operatorname{grad}\phi, \operatorname{grad}\psi\rangle\,\mu = \int_M \psi\Delta\phi\,\mu.\end{aligned}$$

This formula establishes that Δ is a self-adjoint operator. Now functional analysis tells us that a negative-definite self-adjoint operator has eigenvalues and eigenfunctions, so that we have functions ϕ_k and numbers $\lambda_k > 0$ such that $\Delta\phi_k = -\lambda_k\phi_k$. The numbers λ_k are different for every Riemannian manifold and tell us something about the geometry. (Thus for example a famous problem is "Can you hear the shape of a drum?" which refers to the fact that modes of a vibrating surface are eigenfunctions of the Laplacian while frequencies are related to the eigenvalues, so the question is asking whether two manifolds with the same eigenvalues of the Laplacian must be isometric. It turns out to almost but not quite be true.)

The Laplacian of a vector field is more subtle, and in some sense more interesting. The interesting thing is that there are two reasonable definitions: for example on a 3-dimensional manifold we can write

$$\Delta_h V = \operatorname{grad}\operatorname{div} V - \operatorname{curl}\operatorname{curl} V, \tag{20.4.18}$$

which in the language of forms is the more easily-generalized formula

$$\Delta V = (d\star d\star V^\flat - \star d\star dV^\flat)^\sharp$$

in terms of the musical isomorphisms from Definition 19.1.2.

Another alternative is to consider the operator

$$(X, Y) \mapsto \nabla^2 V(X, Y)$$

defined by

$$(\nabla^2 V)(X, Y) = \nabla_X\nabla_Y V - \nabla_{\nabla_X Y} V,$$

with the correction term applied so that the operator is tensorial in both X and in Y. Notice that

$$(\nabla^2 V)(X,Y) - (\nabla^2 V)(Y,X) = R(Y,X)V,$$

by definition of the Riemann curvature tensor (19.3.10), so this operator is symmetric in X and Y if and only if the curvature is zero. By lifting indices and holding V fixed, we can think of $\nabla^2 V$ as a linear map Ξ from T_pM to T_pM, defined so that

$$\langle \Xi(X), Y\rangle = (\nabla^2 V)(X,Y).$$

This map has a trace which does not depend on choice of any vector fields, and from this idea we get a formula for the Laplacian in an orthonormal frame,

$$\Delta_b V = \sum_{i=1}^{n} \nabla_{E_i}\nabla_{E_i} V - \nabla_{\nabla_{E_i} E_i} V. \tag{20.4.19}$$

It sure would be nice if formulas (20.4.18) and (20.4.19) were the same. It turns out they almost are, except for a correction term involving the curvature. Formula (20.4.18) is called the Hodge Laplacian, while formula (20.4.19) is called the Bochner Laplacian. (Frequently in Riemannian geometry the signs are switched so that the Laplacian ends up being positive-definite instead of negative-definite.) The difference between them is

$$\Delta_b V - \Delta_c V = \pm \sum_i \langle R(E_i, V)E_i.$$

This is not a trivial issue: for example when writing the Navier-Stokes equations for a fluid on a manifold, one needs the vector Laplacian for the viscosity term, and there is some ambiguity over which one to use.

Speaking of fluids, let's do something fun with everything we've learned in this course. The Euler equation of ideal fluid mechanics for steady flow on a Riemannian manifold is

$$\nabla_V V = -\operatorname{grad} p, \qquad \operatorname{div} V = 0, \qquad \Delta p = -\operatorname{div}(\nabla_V V), \tag{20.4.20}$$

where V is the velocity field of the fluid and p is the pressure function, determined implicitly by the fact that the operator Δ is invertible. Now the fact that the pressure is determined indirectly makes these equations rather difficult to analyze, and so we can rewrite it to get rid of this term. Notice that if we lower indices using Definition 19.1.2, we get

$$(\nabla_V V)^\flat = -dp,$$

and so

$$d(\nabla_V V)^\flat = 0.$$

Let's try to simplify this. First notice that for any vector field Y we have

$$\begin{aligned} \langle \nabla_V V, Y\rangle &= V\big(\langle V, Y\rangle\big) - \langle V, \nabla_V Y\rangle \\ &= V\big(\langle V, Y\rangle\big) - \langle V, [V,Y]\rangle - \tfrac{1}{2} Y\big(\langle V, V\rangle\big). \end{aligned} \tag{20.4.21}$$

Now recall that

$$\begin{aligned} dV^\flat(V,Y) &= V\big(V^\flat(Y)\big) - Y\big(V^\flat(V)\big) - V^\flat([V,Y]) \\ &= V\big(\langle V, Y\rangle\big) - Y\big(\langle V, V\rangle\big) - \langle V, [V,Y]\rangle. \end{aligned} \tag{20.4.22}$$

Matching up terms between (20.4.21) and (20.4.22), we see that

$$\langle \nabla_V V, Y\rangle = dV^\flat(V,Y) + \tfrac{1}{2} Y\big(\langle V, V\rangle\big).$$

This is true for every vector field Y, and both sides are tensorial in Y, so we can eliminate it entirely and get

$$(\nabla_V V)^\flat = \iota_V dV^\flat + \tfrac{1}{2} d(|V|^2).$$

That means the Euler equation (20.4.20) becomes, in terms of forms,

$$\iota_V dV^\flat + d\big(\tfrac{1}{2}|V|^2 + p\big) = 0.$$

Taking d of both sides gives

$$d\iota_V dV^\flat = 0.$$

Now notice that by Cartan's magic formula, Proposition 18.2.3, we have

$$\mathcal{L}_V dV^\flat = d\iota_V dV^\flat + \iota_V d^2 V^\flat = d\iota_V dV^\flat$$

since $d^2 = 0$. So the Euler equation actually says that

$$\mathcal{L}_V dV^\flat = 0.$$

Let $\omega = dV^\flat$ be the *vorticity 2-form*. Then $\mathcal{L}_V \omega = 0$ says that vorticity is conserved along trajectories: that is, if Φ_t is the flow of V, then

$$\Phi_t^* \omega = \omega.$$

Since the conserved quantity is a 2-form, we get rather different results in two dimensions vs. three dimensions, and this is by far the biggest difference between two-dimensional fluids (which are relatively easy to understand) and three-dimensional fluids (which are the source of some very famous unsolved problems).

We can go further. Consider a surface bounded by a curve (that is, a 2-chain S whose boundary is a 1-chain C). Push the surface S by the flow Φ_t to get a time-dependent surface $S_t = \Phi_t \circ S$. Then

$$\int_{S_t} \omega = \int_{\Phi_t \circ S} \omega = \int_S \Phi_t^* \omega = \int_S \omega,$$

so that the *vorticity flux* is conserved along a surface moving with the fluid.

Furthermore we have

$$C_t = \Phi_t \circ C = \Phi_t \circ \partial S = \partial \Phi_t \circ S = \partial S_t.$$

Since $\omega = dV^\flat$, we have

$$\int_{S_t} \omega = \int_{S_t} dV^\flat = \int_{\partial S_t} V^\flat = \int_{C_t} V^\flat.$$

Since the vorticity flux is conserved, this quantity is also conserved. The integral $\int_C V^\flat$ is called the *circulation* of the fluid, and the fact that circulation is conserved along a closed curve that moves with the fluid is *Kelvin's circulation theorem*.

Of course conservation of vorticity and Kelvin's circulation theorem were historically derived without the benefit of Stokes' theorem or differential forms, but the nice thing about this approach is that this technique is applicable to all sorts of variations on the basic Euler equation to derive conservation laws in a straightforward and elegant way. This is generally what makes differential geometry techniques useful in practical fields, aside from the intrinsic beauty of the subject. You should now be able to apply our formulas and concepts to a wide variety of situations. Good luck!

"At an end your rule is. And not short enough it was."

Index