Differential Geometry I and II, 2021–2022, HU Berlin

Chris Wendl

Contents

These notes will be expanded gradually over the course of the semester. If you notice any typos or mathematical errors, please send e-mail about them to `wendl@math.hu-berlin.de` and they will be corrected.

While the notes are written in English, I make an effort to include the German translations (*geschrieben in dieser Schriftart*) of important terms wherever they are introduced. I will occasionally omit these translations in cases where the English and German words are identical, or if the word has already appeared before with its translation in a different context (e.g. the word "smooth" needs to be defined many times in different contexts, and its German translation is always the same), and also in cases where I can't reliably figure out what the German word is. The latter will happen more often as the course goes on, because the deeper one gets into advanced mathematics, the harder it becomes to find authoritative German sources for clarifying the terminology (and I am not linguistically qualified to invent terms in German myself).

First semester (Differentialgeometrie I)

1. Introduction

Before diving in with definitions, theorems and proofs, I want to give an informal taste of what differential geometry is all about. The word "informal" means, in this case, that you should try not to worry too much about the precise definitions or rigorous arguments behind what we are discussing, but focus instead on the big picture. Before the first lecture is finished, I will revert to being a proper mathematician and give some actual definitions.

1.1. A foretaste of Riemannian geometry. Let's assume for the moment that we all understand what a "smooth surface" is, e.g. you can picture it as a subset[1] of $\mathbb{R}^3$ such that every point has a neighborhood parametrized by some injective[2] C^∞-map

$$\mathbb{R}^2 \overset{\text{open}}{\supset} \mathcal{U} \hookrightarrow \mathbb{R}^3.$$

With this understood, assume

$$\Sigma \subset \mathbb{R}^3$$

is a smooth surface.

1.1.1. *Distances and geodesics.* We could view Σ as a metric space by defining the distance between two points $x, y \in \Sigma$ via the Euclidean metric, but this is not necessarily the most natural thing to do. A more natural notion of distance in the surface Σ would be one that tells you something about the actual distance that an ant has to travel if it walks a path *along the surface* between x and y. If that path is parametrized by a smooth map $\gamma : [a, b] \to \mathbb{R}^3$ satisfying $\gamma([a, b]) \subset \Sigma$, $\gamma(a) = x$ and $\gamma(b) = y$, then the distance travelled is

$$\ell(\gamma) := \int_a^b |\dot{\gamma}(t)|\, dt = \int_a^b \sqrt{\langle \dot{\gamma}(t), \dot{\gamma}(t) \rangle}\, dt, \tag{1.1}$$

where $\dot{\gamma}(t)$ denotes the time derivative of $\gamma(t)$, $\langle v, w \rangle$ denotes the Euclidean inner product of two vectors $v, w \in \mathbb{R}^3$, and $|v| := \sqrt{\langle v, v \rangle}$ denotes the Euclidean norm. If we denote by $\mathcal{P}(x, y)$ the set of all smooth paths in Σ connecting x to y, then a natural notion of distance on Σ can now be defined by

$$d(x, y) := \inf_{\gamma \in \mathcal{P}(x,y)} \ell(\gamma). \tag{1.2}$$

The infimum needs to be taken since, in general, there are many distinct paths from x to y that will have different lengths. In principle we are interested in the *shortest* such path, though it is not obvious in general whether such a shortest path must exist:

---

[1] We will soon improve this definition so that surfaces do not need to be regarded as subsets of $\mathbb{R}^3$. In fact, there are some important examples of surfaces that *cannot* be embedded in $\mathbb{R}^3$; a famous example is the Klein bottle, see https://en.wikipedia.org/wiki/Klein_bottle.

[2] We will need to add a condition concerning the derivative of the map $\mathcal{U} \hookrightarrow \mathbb{R}^3$ before this becomes an adequate definition, but let's worry about that later.

Question 1.1. *Given a smooth surface Σ and two distinct points $x, y \in \Sigma$, does there exist a smooth path on Σ from x to y that has the* shortest *possible length? Is it unique?*

We will see later in this semester that the answer to both questions is always yes if x and y are close enough to each other, and the shortest path can then be characterized by a second-order ordinary differential equation. Such a path is called a **geodesic** (*Geodäte* or *geodätische Linie*), and it serves as the best possible substitute for a "straight line" on Σ, even in cases where no actual *straight* paths on Σ exist. The canonical example you should picture is the unit sphere $\Sigma := S^2 \subset \mathbb{R}^3$, whose geodesics are the so-called *great circles*, namely the subsets $S^2 \cap P$ defined via 2-dimensional linear subspaces $P \subset \mathbb{R}^3$. These are the paths that all airplanes would traverse along the Earth if there were no additional factors such as weather conditions or no-fly zones to consider.

1.1.2. *Angles, isometries, and curvature.* The fundamental piece of data that makes the above definition of distance on Σ possible is the Euclidean inner product $\langle\ ,\ \rangle$. In fact, $\langle\ ,\ \rangle$ contains strictly more information than is actually needed for defining distances on Σ; if you look again at the formula (1.1), you'll notice that it doesn't really require knowing what $\langle v, w\rangle$ is for every $v, w \in \mathbb{R}^3$, but is already well-defined if we know how to define this for every pair of vectors v, w that are *tangent* to Σ at any given point. (Indeed, $\dot{\gamma}(t) \in \mathbb{R}^3$ is always tangent to Σ at $\gamma(t)$.) In fact, it would suffice to know what $\langle v, v\rangle$ is for every individual tangent vector v, but knowing $\langle v, w\rangle$ for two distinct vectors provides some additional information that is of geometric interest: it allows us to compute the *angle* between any two tangent vectors. Indeed, the angle θ between two vectors $v, w \in \mathbb{R}^3$ can always be deduced from the formula

$$\langle v, w\rangle = |v| \cdot |w| \cdot \cos\theta.$$

We can therefore define not only the length of any smooth path along Σ, but also the angle between two smooth paths wherever they intersect. This information makes Σ into what we will later call a (2-dimensional) **Riemannian manifold** (*Riemannsche Mannigfaltigkeit*), and the restriction of the inner product to the tangent spaces on Σ, which determines all lengths and angles, is called a **Riemannian metric** (*Riemannsche Metrik*).[3]

Here is a natural question one can ask about Riemannian manifolds. Suppose $\Sigma_1, \Sigma_2 \subset \mathbb{R}^3$ are two smooth surfaces, and $\varphi : \Sigma_1 \to \Sigma_2$ is a smooth bijective map between them whose inverse is also smooth.[4] We call φ in this case a **diffeomorphism** (*Diffeomorphismus*), and say that Σ_1 and Σ_2 are **diffeomorphic** (*diffeomorph*). We say that φ is additionally an **isometry** (*Isometrie*) if it preserves all distances and angles, and in this case, Σ_1 and Σ_2 are said to be **isometric** (*isometrisch*).

Question 1.2. *Given two diffeomorphic surfaces, how can we measure whether they are isometric?*

In simple examples, it is often easy to recognize when two surfaces are diffeomorphic: an example is shown in Figures 1 and 2, where we can compare the standard unit sphere $S^2 \subset \mathbb{R}^3$ with a "nonstandard" embedding of S^2 into $\mathbb{R}^3$ that elongates a portion of the sphere into something more closely resembling a cylinder. It is surely not hard to imagine that these two surfaces in $\mathbb{R}^3$ are diffeomorphic; writing down an explicit example of a diffeomorphism would be a pain in the neck,

[3] Caution: there is a potential for confusion in this terminology, because a Riemannian metric is not a particular kind of metric in the sense of metric spaces, though it does determine one via formulas such as (1.2). A Riemannian metric carries strictly more information, since it determines angles in addition to distances.

[4] For the purposes of this discussion, you may assume that a function on a smooth surface $\Sigma \subset \mathbb{R}^3$ is smooth if it can be extended to a smooth function on a neighborhood of Σ; the latter notion is familiar from your first-year Analysis class since the neighborhood is an open subset of $\mathbb{R}^3$. We will later give an equivalent but more elegant definition of smoothness for functions on manifolds.

but we will content ourselves with the intuitive understanding that in the process of "stretching" the standard sphere into its nonstandard counterpart, one could if necessary come up with a smooth bijection between the two. The much deeper observation is that they are *not* isometric, and we will need to develop some technology before we can prove this rigorously. One of the key ideas behind the proof is shown in Figures 1 and 2: on any surface Σ, one can draw a closed piecewise-smooth path along Σ, choose a starting point p_0 on the path and a tangent vector v_0 at p_0, then translate the vector v_0 along the path via a process known as **parallel transport**. We will have to give a careful definition later of what is meant by parallel transport, but Figures 1 and 2 will hopefully give you some intuition about this. The interesting question is now: if we parallel transport the vector v_0 once around our chosen closed path, does it return to the same starting vector? As you can see in the pictures, the answer is no for the triangular path in Figure 1, but yes for the rectangular path in Figure 2. It will turn out that this observation encodes a fundamental difference between these two Riemannian manifolds: the standard sphere has positive **curvature** (*Krümmung*) at every point, but the elongated sphere does not—if fact, the surface in Figure 2 has zero curvature everywhere on the elongated region where our rectangle is drawn.

A major portion of the second half of this semester will be devoted to the precise definition of curvature and its important properties. One of these is that it completely characterizes the notion of *local flatness*:

QUESTION 1.3. *Given a smooth surface $\Sigma \subset \mathbb{R}^3$ and a point $p \in \Sigma$, does p have any neighborhood that is isometric to an open subset of the "flat" surface $\mathbb{R}^2 \times \{0\} \subset \mathbb{R}^3$?*

A surface $\Sigma \subset \mathbb{R}^3$ is called **locally flat** (*lokal flach*) if the answer to Question 1.3 is yes for every point $p \in \Sigma$. Figure 3 shows an example of a surface that is locally flat, even though it does not look flat in the picture: you know it is locally flat because you know that an ordinary piece of paper can be bent into this cylindrical shape without breaking or stretching it. This is *not* true of the standard unit sphere in $\mathbb{R}^3$. Perhaps you've never held in your hand a piece of paper that's shaped like part of a globe[5], but you can surely imagine that if you did, you could never make it *flat* without breaking or stretching it. This is another symptom of the positive curvature of the round sphere.[6] By contrast, the cylindrical surface in Figure 3 has zero curvature everywhere. The statement that a cylinder is in some sense "not curved" may seem jarring at first, but you'll get used to it: the point is that the quantity we're calling curvature should depend only on the Riemannian metric, and not on the specific way we've chosen to embed our Riemannian manifold in $\mathbb{R}^3$. If two surfaces are isometric, then their curvatures at corresponding points will always be the same.

The positive curvature of the round sphere is not unrelated to the fact that the angles of the "triangle" in Figure 1 add up to considerably more than 180 degrees. We will later also see examples of surfaces with *negative* curvature: the basic picture to have in mind is the shape of a *saddle*. In these surfaces, the angles in a triangle will add up to *less* than 180 degrees. The elongated sphere in Figure 2 has zero curvature in the shaded region, but not everywhere; since it is diffeomorphic to S^2, one could reinterpret this as the statement that S^2 admits a Riemannian metric that is locally flat in some region. That is not a deep or surprising statement, as *every* Riemannian metric on an arbitrary manifold can in fact be modified to make it flat in some small region. A more interesting question is whether it can be modified to make it locally flat *everywhere*, like the cylindrical surface in Figure 3. Let us take this opportunity to state a standard corollary of a rather deep theorem:

[5]If you know where to buy one, please let me know!

[6]This is also the mathematical reason why it is impossible to create a flat map of the Earth without distorting distances and angles in some regions.

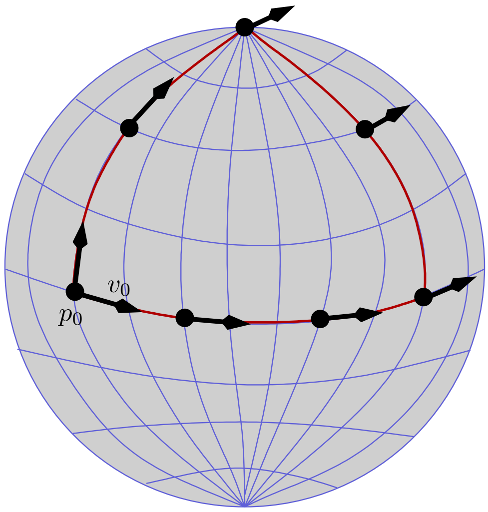

FIGURE 1. The "round" sphere $S^2 \subset \mathbb{R}^3$. Parallel transport of a vector along a closed path leads to a different vector upon return.

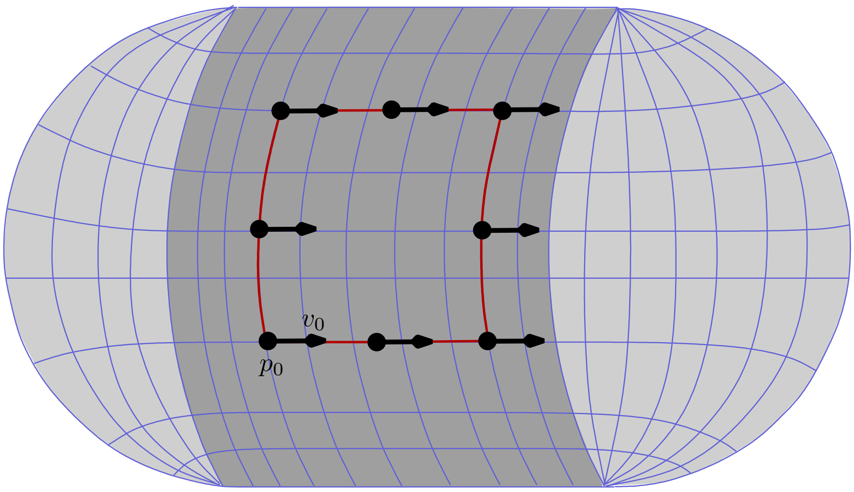

FIGURE 2. A different embedding of S^2 in $\mathbb{R}^3$, so that the darkly shaded region is locally flat. Parallel transport of a vector around a closed path in this region always leads back to the same initial vector.

THEOREM. *There is no Riemannian metric on the sphere S^2 that is everywhere locally flat.*

This will follow from the beautiful *Gauss-Bonnet theorem* for surfaces, to be proved near the end of this semester. It relates the integral of the curvature over a compact surface to a topological quantity, its *Euler characteristic*, which in the case of S^2 is positive. This is the reason why Figure 2 could not have been drawn so that *every* part of the sphere had zero curvature. We will also use a variant of this theorem to understand what the various observations above about sums of angles of triangles have to do with curvature.

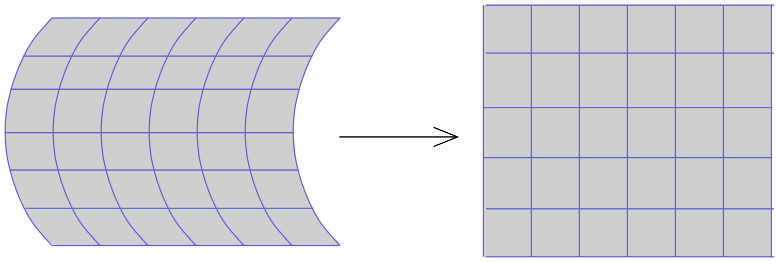

FIGURE 3. A piece of a cylinder can be flattened to a plane without changing any lengths or angles on the surface.

1.1.3. *Spacetime as a pseudo-Riemannian 4-manifold.* Differential geometry is not only about surfaces, and it also plays an important role in subjects that cannot accurately be called "pure" mathematics. This is true especially in several areas of theoretical physics, the most famous of which is Einstein's theory of gravitation, known as the general theory of relativity *(allgemeine Relativitätstheorie)*. We will not directly discuss gravitation in this course, but several of the mathematical concepts we will cover are essential for understanding Einstein's picture of the universe.

The paradigm introduced by Einstein for an understanding of space and time can be summarized as follows:

(1) There are three spatial dimensions, but time adds a fourth. Locally, an "event" occurring in a particular place at a particular time thus requires four coordinates for its description, defining a point in $\mathbb{R}^4$.
(2) The picture in item (1) is only local, i.e. it is sufficient for describing interactions between events on a small or medium scale, but one should not assume that the set of all events in the universe (known as **spacetime** or *Raumzeit*) is in bijective correspondence with $\mathbb{R}^4$. In general, spacetime could be any smooth 4-dimensional manifold.
(3) Spacetime is endowed with a (pseudo-)Riemannian metric, which determines a notion of geodesics. In the absence of forces other than gravity, all objects move along geodesics in spacetime.
(4) The presence of mass affects the curvature of spacetime and thus changes the geodesics. A precise relationship between mass and curvature is given by the Einstein equation, the fundamental field equation of general relativity.

In this paradigm, gravity is not a force: it is just a geometric effect produced by the interaction between mass and curvature. In other words, the reason a brick falls toward the Earth if you drop it is that as soon as you let go, it starts following a geodesic in spacetime, and the Earth's mass causes curvature that determines the shape of that geodesic: moving forward in time while moving closer to the Earth in space.

I should say a word about the appearance of the prefix "pseudo-" in the above paradigm, which places Einstein's theory slightly outside the realm of standard Riemannian geometry. As sketched above, a Riemannian metric on a manifold M is a choice for each point $p \in M$ of an inner product on the space of tangent vectors to M at p. As you know, an inner product $\langle \ , \ \rangle$ on a real vector space V is a positive-definite bilinear form, implying in particular that it is

- **symmetric**: $\langle v, w\rangle = \langle w, v\rangle$ for all $v, w \in V$;
- **nondegenerate**: For every $v \in V\backslash\{0\}$, there exists $w \in V$ such that $\langle v, w\rangle \neq 0$.

To define a pseudo-Riemannian metric on M, one adopts these two assumptions for the inner product $\langle\ ,\ \rangle$ on the space of tangent vectors at every point $p \in M$, but without assuming any positivity, i.e. we do not require $\langle v, v\rangle$ to be positive whenever $v \neq 0$. The classification of quadratic forms (or equivalently the spectral theorem for symmetric linear maps) implies that any n-dimensional vector space V with a symmetric nondegenerate bilinear form $\langle\ ,\ \rangle$ can be split into two orthogonal (with respect to $\langle\ ,\ \rangle$) subspaces

$$V = V_+ \oplus V_-$$

such that $\langle\ ,\ \rangle$ is positive-definite on V_+ and negative-definite on V_-. (Note that if both subspaces are nontrivial, then there always also exist nonzero vectors $v \in V$ such that $\langle v, v\rangle$ is zero—this does not contradict nondegeneracy!) The pseudo-Riemannian metrics used in general relativity have the property that on every tangent space, $\dim V_+ = 3$ and $\dim V_- = 1$.[7] Pseudo-Riemannian metrics with this property are also sometimes called **Lorentzian metrics**, and said to have **Lorentz signature**.

The canonical example of a Lorentzian inner product is what is called the **Minkowski metric** on $\mathbb{R}^4$: we define it by

$$\langle x, y\rangle = -x^0 y^0 + \sum_{j=1}^{3} x^j y^j, \tag{1.3}$$

where we are following the physicists' convention of labeling vectors $v \in \mathbb{R}^4$ by their coordinates v^μ with $\mu = 0, 1, 2, 3$. It is actually crucial for Einstein's theory that the metric on spacetime is *not* positive-definite, because the Lorentzian signature is precisely what produces qualitative physical distinctions between the three spatial dimensions and the fourth one, time. In the convention used above to write down the Minkowski metric, time is labelled as the zeroth coordinate, and is thus distinguished by the minus sign appearing in (1.3). More generally, a vector v in a vector space V with a Lorentzian inner product $\langle\ ,\ \rangle$ is called **time-like** if $\langle v, v\rangle < 0$, **space-like** if $\langle v, v\rangle > 0$, and **light-like** if $\langle v, v\rangle = 0$. With a bit of linear algebra, one can see that the set of all space-like vectors is connected, but the set of vectors that are time-like or light-like splits into two connected components, which we think of as representing motion *forward* or *backward in time*. Similarly, on a Lorentzian manifold, a geodesic can be either time-like, light-like or space-like, and in the first two categories one can distinguish between parametrizations of the geodesic that are oriented forward or backward in time, while for space-like geodesics there is no such distinction. The physical significance of these observations is the following: in general relativity, all particles with mass travel through spacetime along time-like geodesics, while particles with no mass travel along light-like geodesics—the latter are the particles that observers perceive as travelling at the speed of light. As far as we know, *nothing* travels along space-like geodesics, which is equivalent to saying that nothing travels faster than light. According to the geometry of spacetime, anything that *could* do this would also sometimes be observed to travel backward in time. Naturally, the non-existence of such particles according to the known laws of physics has not stopped physicists from giving them a name—tachyons—and they are mentioned frequently in science fiction, as a clearly necessary ingredient in time travel.

While we will probably not say anything further about general relativity in this course, we will prove some results about pseudo-Riemannian manifolds, and will try to avoid assuming that inner products are positive-definite unless that assumption is absolutely necessary.

1.1.4. *Gauge theory.* To round out this motivational introduction, I want to mention briefly another area of physics beyond general relativity where differential geometry plays a key role. The last half-century has witnessed intense and fruitful interactions between geometry and quantum

[7] Or possibly the other way around—the literature is not unanimous on this convention.

field theory (on which the theory of elementary particles is based), along with its more exotic and controversial cousin, string theory. Each of the classical fields underlying the various types of elementary particles can be described mathematically as a geometric object, namely a *section* of a *smooth fiber bundle*. The particles that mediate the electromagnetic, strong and weak nuclear forces, in particular, are described via so-called *gauge fields*, which are known to mathematicians as *connections*: these are a fundamental piece of geometric data on a fiber bundle, analogous to the Lorentzian metrics on the spacetime manifold of general relativity. This subject as a whole is known as *gauge theory*, a term which means slightly different things in the two fields: physicists understand it as the basis of their understanding of the forces of nature, while for mathematicians, it is a powerful framework for developing geometric and topological invariants based on spaces of solutions to nonlinear PDEs. In the big picture, gauge theory is both, and it has served as one of the most exciting sources of interactions between theoretical physics and pure mathematics during the past few decades. We will lay a few of the basic foundations for this subject via the study of vector bundles in the second half of this semester.

1.2. Charts and transition maps. We now begin the study of differential geometry in earnest.

The fundamental objects of study in this subject are called *smooth, finite-dimensional manifolds*. We will spend most of the first two lectures explaining the definition of this term and giving some basic examples.

We start with the intuition that a 1-dimensional manifold is what you have previously called a "curve" (*Kurve*), and a 2-dimensional manifold is a "surface" (*Fläche*). For arbitrary $n \in \mathbb{N}$, an elementary example of an n-dimensional manifold will be the so-called n**-sphere**

$$S^n := \left\{ x \in \mathbb{R}^{n+1} \;\middle|\; |x| = 1 \right\},$$

where $|\cdot|$ again denotes the Euclidean norm. The word "sphere" (*Sphäre*) on its own normally refers to the familiar case $n = 2$, though it can also refer to the general case if the value of n is clear from context. The 1-sphere has been known to you since Kindergarten under a different name: the **circle** (*Kreis*). Let us examine this example a bit more closely, and clarify in particular the following point: S^1 is defined as a subset of $\mathbb{R}^2$, so why do we consider it a "one-dimensional" object?

The answer can be explained via an intelligent choice of coordinates. Consider the standard polar coordinates (r, θ) on $\mathbb{R}^2$, which are related to the Cartesian coordinates (x, y) by

$$x = r\cos\theta, \qquad y = r\sin\theta.$$

For concreteness, we assume (and will *always* assume) the angle θ is measured in radians, so the range $\theta \in [0, 2\pi]$ describes a full rotation. In polar coordinates, S^1 is the subset $\{r = 1\} \subset \mathbb{R}^2$, thus one of the coordinates becomes irrelevant, and having one coordinate left makes S^1 a one-dimensional object.

The above discussion of polar coordinates glossed over an important point: one cannot simultaneously describe *every* point in S^1 via a unique value of the angular coordinate $\theta \in \mathbb{R}$, at least not if we want the values of θ to be unambiguously defined and continuously dependent on the points that they describe. One could e.g. require θ to take values only in a half-open interval like $[0, 2\pi)$ or $(-\pi, \pi]$: this creates a one-to-one correspondence between points on S^1 and values of the coordinate, but the function one defines in this way from S^1 to $[0, 2\pi)$ or $(-\pi, \pi]$ has a jump discontinuity at the point where the coordinate reaches either end of the allowed interval. If you want to avoid such discontinuities, then the only option is to give up on the notion of describing *all* of S^1 in a single coordinate system, and instead use multiple coordinate systems defined on different subsets. For instance, we could define two subsets of the circle by

$$\mathcal{U} := S^1 \backslash \{(1,0)\}, \qquad \mathcal{V} := S^1 \backslash \{(-1,0)\},$$

and associate to these two subsets two potentially different angular coordinates θ and ϕ respectively, each taking values in an appropriate open interval, thus defining *continuous* functions

$$\theta : \mathcal{U} \to (0, 2\pi), \qquad \phi : \mathcal{V} \to (-\pi, \pi).$$

Since $S^1 = \mathcal{U} \cup \mathcal{V}$, these two coordinate systems together can be used to describe every point in S^1. Moreover, there is a large region on which both coordinates θ and ϕ are defined: it consists of the two semi-circles $S^1_+ := \{(x,y) \in S^1 \mid y > 0\}$ and $S^1_- := \{(x,y) \in S^1 \mid y < 0\}$, and on each of these one can easily derive a relationship between θ and ϕ, namely

$$\phi = \begin{cases} \theta & \text{on } S^1_+, \\ \theta - 2\pi & \text{on } S^1_-. \end{cases} \tag{1.4}$$

The pairs $(\mathcal{U}, \theta)$ and $(\mathcal{V}, \phi)$ are our first examples of what we will call *charts* on the 1-dimensional manifold S^1, and together they form a *smooth atlas* that determines a *smooth structure* on S^1. Let us now begin giving precise definitions to these terms.

In the following, assume M is a set, and $n \geqslant 0$ is an integer. For the sake of intuition, you may picture M as a surface (in which case $n = 2$), and picture the subsets $\mathcal{U}, \mathcal{V} \subset M$ as open subsets of that surface.[8] Recall that a continuous map defined on an open subset of Euclidean space is called **smooth** (*glatt*) if it admits derivatives of all orders.

Definition 1.4. An n-dimensional **chart** (*Karte*)[9] $(\mathcal{U}, x)$ on M consists of a subset $\mathcal{U} \subset M$ and an injective map $x : \mathcal{U} \hookrightarrow \mathbb{R}^n$ whose image $x(\mathcal{U}) \subset \mathbb{R}^n$ is an open set.

Any two charts $(\mathcal{U}, x)$ and $(\mathcal{V}, y)$ determine a pair of **transition maps** (*Kartenübergänge*)

$$\begin{aligned} \mathbb{R}^n \supset x(\mathcal{U} \cap \mathcal{V}) &\xrightarrow{y \circ x^{-1}} y(\mathcal{U} \cap \mathcal{V}) \subset \mathbb{R}^n, \\ \mathbb{R}^n \supset y(\mathcal{U} \cap \mathcal{V}) &\xrightarrow{x \circ y^{-1}} x(\mathcal{U} \cap \mathcal{V}) \subset \mathbb{R}^n, \end{aligned} \tag{1.5}$$

which are inverse to each other, and are thus bijections between subsets of $\mathbb{R}^n$. We say that the two charts are C^k**-compatible** (*verträglich*) for some $k \in \mathbb{N} \cup \{0, \infty\}$ if the sets $x(\mathcal{U} \cap \mathcal{V})$ and $y(\mathcal{U} \cap \mathcal{V})$ are both open and the transition maps $y \circ x^{-1}$ and $x \circ y^{-1}$ are both of class C^k. If $k = \infty$, we say the charts are **smoothly compatible** (*glatt verträglich*).

A picture of what a pair of overlapping charts on a surface might look like is shown in Figure 4. An individual chart $(\mathcal{U}, x)$ should be understood as defining a *coordinate system* for describing all points in the subset $\mathcal{U} \subset M$, where the individual **coordinates** (*Koordinaten*) are the n real-valued functions

$$x^1, \ldots, x^n : \mathcal{U} \to \mathbb{R}$$

defined as the component functions of the map $x = (x^1, \ldots, x^n) : \mathcal{U} \to \mathbb{R}^n$. Note that in Definition 1.4, it is permissible for the domains $\mathcal{U}$ and $\mathcal{V}$ of the two charts to be disjoint, in which case

[8] Saying the word "open" presumes that M has some structure beyond merely being an arbitrary set, e.g. it could be a subset of some Euclidean space $\mathbb{R}^n$, or more generally, a metric or topological space. We will address this point properly in the next lecture, but since we have not addressed it yet, Definition 1.4 refers to $\mathcal{U}$ and $\mathcal{V}$ simply as "subsets" of M, without saying they are open. In practice, they always will be.

[9] A word of caution for German speakers: the mathematical word *Abbildung* (as in "eine injektive Abbildung von $\mathbb{R}^n$ nach $\mathbb{R}^m$") can be translated into English as either "map" or "mapping", but do not be tempted to translate "map" into mathematical German as *Karte*. In mathematical English, a "chart" and a "map" are not exactly the same thing.

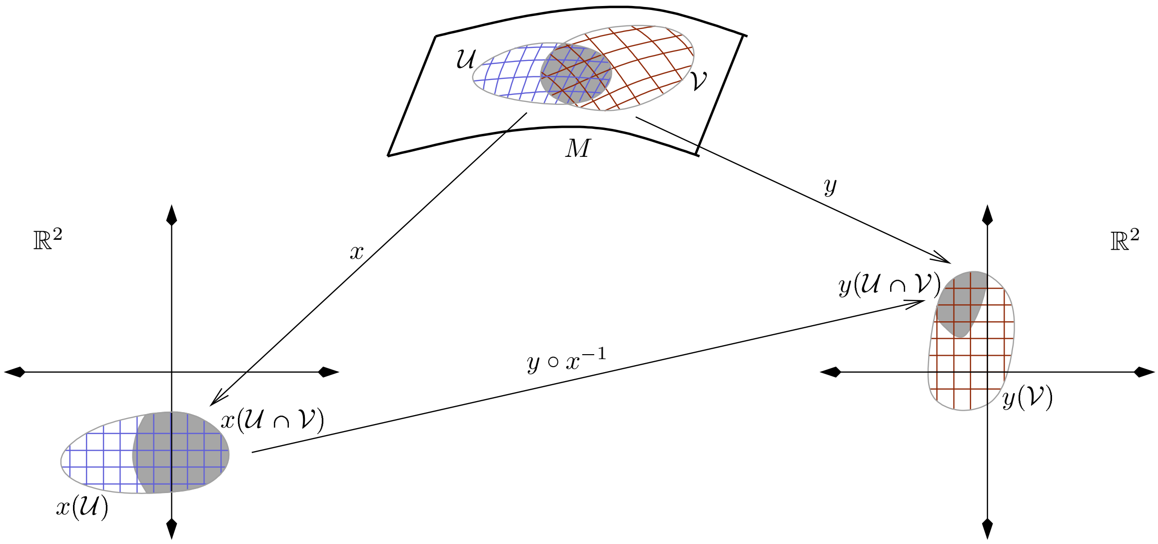

FIGURE 4. Two charts $(\mathcal{U}, x)$ and $(\mathcal{V}, y)$ on a surface M, with an associated transition map $y \circ x^{-1}$ defining a bijection between two open sets (the shaded regions) in $\mathbb{R}^2$.

the transition maps $y \circ x^{-1}$ and $x \circ y^{-1}$ are both just the trivial map from the empty set to itself. But if $\mathcal{U} \cap \mathcal{V} \neq \varnothing$, then the transition map

$$\begin{tikzcd} & & \mathcal{U} \cap \mathcal{V} \arrow[dl, "x"'] \arrow[dr, "y"] & & \\ \mathbb{R}^n \overset{\text{open}}{\supset} & x(\mathcal{U} \cap \mathcal{V}) \arrow[rr, "y \circ x^{-1}"] & & y(\mathcal{U} \cap \mathcal{V}) & \overset{\text{open}}{\subset} \mathbb{R}^n \end{tikzcd}$$

defines a coordinate transformation, e.g. for any point $p \in \mathcal{U} \cap \mathcal{V}$, $y \circ x^{-1}$ sends the vector $(x^1(p), \dots, x^n(p)) \in \mathbb{R}^n$ that represents p in "x-coordinates" to the vector that represents the same point in "y-coordinates", namely $(y^1(p), \dots, y^n(p)) \in \mathbb{R}^n$. It is often convenient in this situation to write the y-coordinates on the overlap region as functions of the x-coordinates, i.e. if we identify each point in $\mathcal{U} \cap \mathcal{V}$ with the vector in $\mathbb{R}^n$ determined by its x-coordinates, then the y-coordinates can be viewed as functions of n variables, which are naturally labelled $x^1, \dots, x^n$, producing a transformation

$$(x^1, \dots, x^n) \mapsto (y^1(x^1, \dots, x^n), \dots, (y^n(x^1, \dots, x^n)). \tag{1.6}$$

This is a slight abuse of notation, because in this expression, the variables $x^1, \dots, x^n$ are no longer interpreted as real-valued functions on $\mathcal{U} \subset M$, but simply as the usual Cartesian coordinates on the open subset $x(\mathcal{U} \cap \mathcal{V}) \subset \mathbb{R}^n$. With this understood, (1.6) is just another expression for the transition map $y \circ x^{-1}$, and the inverse transition map $x \circ y^{-1}$ can similarly be written as

$$(y^1, \dots, y^n) \mapsto (x^1(y^1, \dots, y^n), \dots, (x^n(y^1, \dots, y^n)), \tag{1.7}$$

with the variables $y^1, \dots, y^n$ now understood to represent Cartesian coordinates on $y(\mathcal{U} \cap \mathcal{V}) \subset \mathbb{R}^n$. If the two charts are C^k-compatible, then both of the transformations in (1.6) and (1.7) are of class C^k. If $k \geqslant 1$, then since the two transformations are inverse to each other, it follows that the n-by-n matrix with entries

$$\frac{\partial y^i}{\partial x^j}(x^1, \dots, x^n), \qquad i, j \in \{1, \dots, n\}$$

is invertible for every $(x^1, \dots, x^n) \in x(\mathcal{U} \cap \mathcal{V}) \subset \mathbb{R}^n$.

REMARK 1.5. You may have been accustomed to using *subscripts* $x_1, \dots, x_n$ for coordinates on $\mathbb{R}^n$ in your studies up to this point, and will thus wonder why I am instead using superscripts in all the expressions above. This is not an arbitrary choice—it is a convention that is widespread in differential geometry, and especially popular among physicists, and we will try to use it consistently throughout this course. Subscripts will at some point also appear, but they will have a different meaning.

EXAMPLE 1.6. In the discussion of the unit circle S^1 above, we defined two charts $(\mathcal{U}, \theta)$ and $(\mathcal{V}, \phi)$, with images $\theta(\mathcal{U}) = (0, 2\pi) \subset \mathbb{R}$ and $\phi(\mathcal{V}) = (-\pi, \pi) \subset \mathbb{R}$. The overlap region $\mathcal{U} \cap \mathcal{V}$ of these two charts is the union of two disjoint open sets that we denoted by S^1_+ and S^1_-, the upper and lower semicircle (disjoint from the x-axis). The transition map $\phi \circ \theta^{-1} : \theta(S^1_+ \cup S^1_-) \to \phi(S^1_+ \cup S^1_-)$ is then found by writing ϕ as a function of θ as in (1.4), which gives

$$\phi(\theta) = \begin{cases} \theta & \text{for } 0 < \theta < \pi, \\ \theta - 2\pi & \text{for } \pi < \theta < 2\pi. \end{cases}$$

Observe that while this map appears at first glance to have a jump discontinuity, its actual domain is $\theta(S^1_+ \cup S^1_-) = (0, \pi) \cup (\pi, 2\pi)$, i.e. it excludes the point π at which the discontinuity would occur. As a result, this transition map is smooth, and so is its inverse; the two charts $(\mathcal{U}, \theta)$ and $(\mathcal{V}, \phi)$ are therefore smoothly compatible.

EXERCISE 1.7. The standard **spherical coordinates** (*Kugelkoordinaten*) on $\mathbb{R}^3$ are defined via the transformation

$$(r, \theta, \phi) \mapsto (x, y, z), \qquad \begin{cases} x & := r \cos\theta \cos\phi, \\ y & := r \sin\theta \cos\phi, \\ z & := r \sin\phi, \end{cases} \tag{1.8}$$

where θ plays the role of an angle in the xy-plane, and $\phi \in [-\pi/2, \pi/2]$ is the angle between the vector $(x, y, z) \in \mathbb{R}^3$ and the xy-plane.[10] Restricting to $r = 1$, the other two coordinates (θ, ϕ) can be used to describe points on the unit sphere $S^2 \subset \mathbb{R}^3$, though there are choices to be made since θ is only defined up to multiples of 2π (and it is not defined at all at the north and south poles $p_\pm := (0, 0, \pm 1) \in S^2$, where $\phi = \pm\pi/2$.)

(a) Find two subsets $\mathcal{U}_1, \mathcal{U}_2 \subset S^2$ with $\mathcal{U}_1 \cup \mathcal{U}_2 = S^2 \backslash \{p_+, p_-\}$ such that for $i = 1, 2$, there are 2-dimensional charts of the form $(\mathcal{U}_i, \alpha_i)$ with $\alpha_i = (\theta_i, \phi_i)$, where the coordinate functions $\theta_i, \phi_i : \mathcal{U}_i \to \mathbb{R}$ are continuous and satisfy the spherical coordinate relations (1.8), and have images $\alpha_1(\mathcal{U}_1) = (0, 2\pi) \times (-\pi/2, \pi/2) \subset \mathbb{R}^2$ and $\alpha_2(\mathcal{U}_2) = (-\pi, \pi) \times (-\pi/2, \pi/2) \subset \mathbb{R}^2$.
(b) One cannot use spherical coordinates to construct a chart on S^2 that contains either of the poles $p_\pm = (0, 0, \pm 1)$. Can you think of another way to construct charts on open subsets of S^2 that contain these two points?

[10] Achtung: there are various conventions for spherical coordinates in use. I'm told that this is the standard convention learned by mathematics students in Germany. I learned a different convention as a physics student in the U.S.: $x = r \cos\phi \sin\theta$, $y = r \sin\phi \sin\theta$, $z = r \cos\theta$. Here ϕ plays the role of the angle in the xy-plane, and $\theta \in [0, \pi]$ is the angle between $(x, y, z) \in \mathbb{R}^3$ and the positive z-axis.

Hint: On any sufficiently small neighborhood of p_+ or p_- in S^2, every point has its z-coordinate determined by the x and y-coordinates.

(c) Now that you've constructed charts that cover every point on S^2, write down the associated transition maps and show that your charts are all smoothly compatible with each other.

2. Smooth manifolds

In this lecture we give the definition of the term *smooth manifold* and look at a few more examples.

2.1. Atlases and smooth structures. We concluded Lecture 1 by defining the notion of a *chart* on a set M, and C^k-compatibility between two charts. A chart $(\mathcal{U}, x)$ should be interpreted as a "local" coordinate system, which can be used to label points in the subset $\mathcal{U} \subset M$. We saw in the example of the circle S^1 that while one cannot apparently describe *all* points in S^1 via a single chart, it was easy to find two smoothly compatible charts such that every point is in at least one or the other. Exercise 1.7 similarly outlines how to cover S^2 with four charts using spherical coordinates. These were the first examples of the following general concept.

Definition 2.1. An **atlas of class C^k** for the set M (or **smooth atlas** in the case $k = \infty$) is a collection of charts $\mathcal{A} = \{(\mathcal{U}_\alpha, x_\alpha)\}_{\alpha \in I}$ that are all C^k-compatible with each other, such that $\bigcup_{\alpha \in I} \mathcal{U}_\alpha = M$.[11]

In first-year analysis, you learned what it means for a real-valued function on an open subset of $\mathbb{R}^n$ to be differentiable; it was important in that definition that the domain of the function should be *open*, as differentiation at a point p involves limits that are not well defined unless f itself is defined on some ball around p. In differential geometry, we would also like to be able to differentiate functions

$$f : M \to \mathbb{R}$$

defined on a manifold M, such as the circle S^1 or sphere S^2. This is a nontrivial problem, even in simple examples such as S^n that are given as subsets of Euclidean space, since they are not generally *open* subsets. But if M is a set equipped with an atlas, then M is covered by subsets that have coordinate systems, so for each chart $(\mathcal{U}, x)$ we can write down f "in local coordinates", meaning we identify each point $p \in \mathcal{U}$ with its coordinate vector $(x^1(p), \dots, x^n(p)) \in \mathbb{R}^n$, so that $f|_{\mathcal{U}} : \mathcal{U} \to \mathbb{R}$ becomes a function of n real variables

$$(x^1, \dots, x^n) \mapsto f(x^1, \dots, x^n), \tag{2.1}$$

with $x^1, \dots, x^n$ interpreted as the standard Cartesian coordinates on the open set $x(\mathcal{U}) \subset \mathbb{R}^n$. This is another slight abuse of notation, similar to the coordinate expressions for transition maps described in (1.6) and (1.7); in fact, the function that is literally described in (2.1) is not $f : M \to \mathbb{R}$ but rather

$$x(\mathcal{U}) \xrightarrow{f \circ x^{-1}} \mathbb{R}.$$

It now seems natural to say that f is differentiable at $p \in \mathcal{U} \subset M$ if and only if its coordinate expression $f \circ x^{-1}$ is differentiable (in the sense of first-year analysis) at the corresponding point $x(p) \in x(\mathcal{U}) \subset \mathbb{R}^n$. For this to be a reasonable definition, we need to know that it does not depend on the *choice* of the chart $(\mathcal{U}, x)$, as our atlas may indeed contain multiple distinct charts that contain the point p. This issue is precisely what the compatibility condition in Definition 1.4 was designed to settle:

[11] In this definition, I may be any set, finite, countable or uncountable. We refer to it as an **index set** since it is only used for labelling purposes and is otherwise unimportant in itself.

LEMMA 2.2. *Suppose* $(\mathcal{U}, x)$ *and* $(\mathcal{V}, y)$ *are two* C^k*-compatible charts on* M*, and* $f : M \to \mathbb{R}$ *is a function. Then for each nonnegative integer* $r \leqslant k$*, the function* $x(\mathcal{U} \cap \mathcal{V}) \xrightarrow{f \circ x^{-1}} \mathbb{R}$ *is of class* C^r *if and only if the function* $y(\mathcal{U} \cap \mathcal{V}) \xrightarrow{f \circ y^{-1}} \mathbb{R}$ *is of class* C^r*.*

PROOF. The statement follows from the chain rule, since $f \circ y^{-1} = (f \circ x^{-1}) \circ (x \circ y^{-1})$ and $f \circ x^{-1} = (f \circ y^{-1}) \circ (y \circ x^{-1})$. □

DEFINITION 2.3. For a set M with an atlas $\mathcal{A}$ of class C^k and $r \in \mathbb{N} \cup \{0, \infty\}$ with $r \leqslant k$, a function $f : M \to \mathbb{R}$ is said to be **of class** C^r if and only if the function $x(\mathcal{U}) \xrightarrow{f \circ x^{-1}} \mathbb{R}$ is of class C^r for every chart $(\mathcal{U}, x) \in \mathcal{A}$.

EXERCISE 2.4. Convince yourself that Lemma 2.2 becomes false in general if one allows $r > k$. (See also Example 2.7 below for a concrete special case.) This has the following consequence: if we want to define what it means for a function on a manifold to be of class C^k, then we need to have an atlas of class C^k *or better* to test it with. In particular, the notion of smooth functions on M cannot be defined unless M is equipped with a *smooth* atlas.

The examples of smooth atlases we saw in Lecture 1 on S^1 and S^2 were finite, and this will turn out to be a general pattern: we will see that almost all manifolds we are interested in admit finite atlases, though it is not often important to know this. On the other hand, a general atlas can be uncountably infinite, and one can always enlarge a finite atlas $\{(\mathcal{U}_\alpha, x_\alpha)\}_{\alpha \in I}$ in trivial ways, e.g. by choosing subsets $\mathcal{U}'_\alpha \subset \mathcal{U}_\alpha$ for which $x_\alpha(\mathcal{U}'_\alpha) \subset \mathbb{R}^n$ is open and adding in the restricted charts $(\mathcal{U}'_\alpha, x_\alpha|_{\mathcal{U}'_\alpha})$, which are obviously still compatible with all the others. We say that an atlas $\mathcal{A} = \{(\mathcal{U}_\alpha, x_\alpha)\}_{\alpha \in I}$ of class C^k is **maximal** if it cannot be enlarged any further without sacrificing compatibility, i.e. every chart that is C^k-compatible with all of the charts in $\mathcal{A}$ already belongs to $\mathcal{A}$.

LEMMA 2.5. *Given an atlas* $\mathcal{A} = \{(\mathcal{U}_\alpha, x_\alpha)\}_{\alpha \in I}$ *of class* C^k *on* M*, let* $\mathcal{A}'$ *denote the collection of all charts on* M *that are* C^k*-compatible with all the charts in* $\mathcal{A}$*. Then* $\mathcal{A}'$ *is a maximal atlas of class* C^k*, and it is the only one containing* $\mathcal{A}$*.*

PROOF. To show that $\mathcal{A}'$ is an atlas, we need to show that any two charts $(\mathcal{U}, x)$ and $(\mathcal{V}, y)$ that are C^k-compatible with every $(\mathcal{U}_\alpha, x_\alpha)$ are also C^k-compatible with each other. Given a point $p \in \mathcal{U} \cap \mathcal{V}$, pick $\alpha \in I$ so that $p \in \mathcal{U}_\alpha$. The set $x(\mathcal{U} \cap \mathcal{V} \cap \mathcal{U}_\alpha) \subset \mathbb{R}^n$ is then the intersection of the two open sets $x(\mathcal{U} \cap \mathcal{U}_\alpha)$ and $x(\mathcal{V} \cap \mathcal{U}_\alpha)$ and is thus an open neighborhood of $x(p)$, so on this neighborhood, the transition map $y \circ x^{-1}$ can then be written as

$$y \circ x^{-1} = (y \circ x_\alpha^{-1}) \circ (x_\alpha \circ x^{-1}),$$

which is a composition of two C^k-maps and is therefore of class C^k on the neighborhood of $x(p)$ in question. This trick works (possibly with different choices of α) for any point $p \in \mathcal{U} \cap \mathcal{V}$, and it also works for the inverse transition map $x \circ y^{-1}$, thus it implies that both of the transition maps relating x and y are everywhere of class C^k, and $\mathcal{A}'$ is therefore an atlas. It clearly also contains $\mathcal{A}$, and it is maximal, since any chart compatible with every chart in $\mathcal{A}'$ is also compatible with every chart in $\mathcal{A}$, and thus belongs to $\mathcal{A}'$ by definition. Finally, if $\mathcal{A}''$ is any other atlas containing $\mathcal{A}$, then every chart in $\mathcal{A}''$ is compatible with every chart in $\mathcal{A} \subset \mathcal{A}''$ and therefore belongs to $\mathcal{A}'$ by definition, proving $\mathcal{A}'' \subset \mathcal{A}'$. If $\mathcal{A}''$ is also maximal, it follows that $\mathcal{A}'' = \mathcal{A}'$. □

DEFINITION 2.6. For $k \in \mathbb{N} \cup \{\infty\}$, a C^k**-structure** (C^k*-Struktur*) or **differentiable structure of class** C^k (*differenzierbare Struktur von der Klasse* C^k) on a set M is a maximal atlas $\mathcal{A}$ of class C^k on M. In the case $k = \infty$, we also call this a **smooth structure** (*glatte Struktur*) on M. If M has been endowed with a C^k-structure $\mathcal{A}$, then a chart $(\mathcal{U}, x)$ on M will be referred to as a C^k**-chart** (or a **smooth chart** in the case $k = \infty$) if it belongs to the maximal atlas $\mathcal{A}$.

The maximality condition in Definition 2.6 is convenient for bookkeeping purposes (see Remark 2.8 below), but Lemma 2.5 shows that it is not a meaningful restriction. In practice, one typically specifies a smooth structure by first describing the smallest atlas one is able to construct, and then replacing it with its unique maximal extension. We will usually carry out the latter step without even mentioning it.

EXAMPLE 2.7. The following defines an atlas of class C^0 but not C^1 on $\mathbb{R}$: consider two charts $(\mathcal{U}, x)$ and $(\mathcal{V}, y)$ with

$$\mathcal{U} := (-\infty, 1), \qquad x(t) := t,$$
$$\mathcal{V} := (-1, \infty), \qquad y(t) := t^3.$$

The resulting transition maps both send $(-1,1) \to (-1,1)$ and are given by

$$y(x) = x^3, \qquad x(y) = \sqrt[3]{y},$$

so both are continuous, but $x \circ y^{-1}$ is not differentiable. This has the consequence that functions $\mathbb{R} \to \mathbb{R}$ that look differentiable in the x-coordinate might not look differentiable in the y-coordinate. An easy example is the identity map $f(t) = t$, which looks like $f(x) = x$ and is thus smooth in the x-coordinate, but its expression in the y-coordinate is $f(y) = \sqrt[3]{y}$, which fails to be differentiable at the point $0 \in y(\mathcal{V}) = (-1, \infty)$.

Note that if we enlarge both $\mathcal{U}$ and $\mathcal{V}$ to $\mathbb{R}$, then while the two charts $(\mathcal{U}, x)$ and $(\mathcal{V}, y)$ together do not determine any smooth structure on $\mathbb{R}$, each of these charts individually forms a smooth atlas—an atlas with only one chart is always smooth since it has no nontrivial transition maps whose differentiability would need to be checked. Each therefore determines a smooth structure via Lemma 2.5, and in this way, one obtains two *different* smooth structures on $\mathbb{R}$.

REMARK 2.8. The advantage of requiring maximality in Definition 2.6 is the following: if $\mathcal{A}$ and $\mathcal{A}'$ are two atlases on M for which every chart in $\mathcal{A}$ is compatible with every chart in $\mathcal{A}'$, then the two notions of differentiability for functions on M defined via these two atlases will be the same, and we would therefore prefer to think of them is defining the *same* smooth structure, even if they are different atlases, strictly speaking. In this scenario, it is easy to check that both atlases do in fact have the same maximal extension.

2.2. Some topological notions. With the concept of a smooth atlas in hand, a reasonable guess for the "right" definition of a smooth manifold would be that it is any set endowed with the additional structure of a smooth atlas. In practice, however, doing anything interesting with manifolds requires imposing one or two further restrictions on what is allowed to be a manifold and what is not.

I do not want to assume previous knowledge of topology in this course, but a few basic notions of the subject now need to be discussed before we can give the precise definition of a manifold. Most of them will play a negligible role in this course, and in fact, the intuition you already have about metric spaces is fully sufficient for understanding the definition of a manifold (cf. Remark 2.20 below)—nonetheless, you will not be able to understand *why* that definition is what it is unless we first discuss the alternatives.

Since you have seen metric spaces before, you know how to define fundamental notions such as **continuity** (*Stetigkeit*), **convergence** of a sequence to a point (*Konvergenz einer Folge gegen einen Punkt*) and **closed** sets (*abgeschlossene Teilmengen*) in metric spaces. You will also have seen important concepts such as that of a **neighborhood** (*Umgebung*) of a point $x \in X$, meaning any subset $\mathcal{U} \subset X$ that contains an open subset containing x, and probably also a **homeomorphism** (*Homöomorphismus*), which is a continuous bijection whose inverse is also continuous. One detail you may or may not already be aware of is that all of these notions can be defined without any explicit reference to a metric, so long as one knows what an "open set" is. In particular:

Proposition 2.9 (first-year analysis). *Assume X and Y are metric spaces.*

(1) A sequence $x_n \in X$ converges to a point $x \in X$ if and only if for every neighborhood $\mathcal{U} \subset X$ of x, $x_n \in \mathcal{U}$ for all sufficiently large n.
(2) A subset $\mathcal{U} \subset X$ is closed if and only if its complement $X \backslash \mathcal{U} \subset X$ is open.
(3) A map $f : X \to Y$ is continuous if and only if for every open subset $\mathcal{U} \subset Y$, $f^{-1}(\mathcal{U}) := \{x \in X \mid f(x) \in \mathcal{U}\}$ is an open subset of X.
(4) A bijective map $f : X \to Y$ is a homeomorphism if and only if it defines a bijective correspondence between the open subsets of X and the open subsets of Y, i.e. for all subsets $\mathcal{U} \subset X$, $\mathcal{U}$ is open if and only if $f(\mathcal{U}) \subset Y$ is open.

□

Exercise 2.10. If you do not already find Proposition 2.9 obvious, prove it.

Topology begins with the observation that it is sometimes convenient to define what an open set is without the aid of a metric. For this idea to be useful, we just need open sets to satisfy a few properties that are already familiar from the theory of metric spaces:

Definition 2.11. A **topology** (*Topologie*) on a set X is a collection $\mathcal{T}$ of subsets of X satisfying the following axioms:

(i) $\emptyset \in \mathcal{T}$ and $X \in \mathcal{T}$;
(ii) For every subcollection $I \subset \mathcal{T}$, $\bigcup_{\mathcal{U} \in I} \mathcal{U} \in \mathcal{T}$;
(iii) For every pair $\mathcal{U}_1, \mathcal{U}_2 \in \mathcal{T}$, $\mathcal{U}_1 \cap \mathcal{U}_2 \in \mathcal{T}$.

The pair $(X, \mathcal{T})$ is then called a **topological space** (*topologischer Raum*), and we call the sets $\mathcal{U} \in \mathcal{T}$ the **open** subsets (*offene Teilmengen*) in $(X, \mathcal{T})$.

We will usually not give an actual label to the topology when discussing a topological space, so e.g. instead of talking about $(X, \mathcal{T})$, we will talk about "the topological space X" with the understanding that a subset $\mathcal{U} \subset X$ is called "open" if and only if it belongs to the topology that has been specified on X. For topological spaces X and Y, one now takes the statements in Proposition 2.9 as *definitions* of the notions of convergence, closed subsets, continuity and homeomorphisms.

We call a topological space X **metrizable** (*metrisierbar*) if it admits a metric for which the given topology of X consists of all sets that are unions of open balls, i.e. the metrizable spaces are the topological spaces that you already saw (but without using the word "topology") when you studied metric spaces. Two things about this notion are important to understand:

(1) If X is metrizable, then the metric that defines its topology is typically far from being unique. For example, $d(x, y) := c|x - y|$ for any constant $c > 0$ defines a "nonstandard" metric on $\mathbb{R}$ that nonetheless induces the same topology as the standard one.
(2) Many topological spaces are not metrizable, and they can easily have properties that are counterintuitive. (We will see an example in a moment.)

We saw in §2.1 that an atlas of class C^k on a set M determines a natural way to define what it means for a function $f : M \to \mathbb{R}$ to be of class C^r for any $r \leqslant k$. This holds in particular for $r = 0$, so that continuity of functions can be defined in a certain sense, even though we never explicitly endowed M with a topology. But actually, we did, we just didn't notice:

PROPOSITION 2.12. *Given an atlas $\mathcal{A} = \{(\mathcal{U}_\alpha, x_\alpha)\}_{\alpha\in I}$ of class C^0 on a set M, there exists a unique topology on M such that the sets $\mathcal{U}_\alpha \subset M$ are all open and the maps x_α are all homeomorphisms onto their images.*[12] *Moreover, for every other chart $(\mathcal{U}, x)$ that is C^0-compatible with the charts in $\mathcal{A}$, $\mathcal{U} \subset M$ is also open and x is also a homeomorphism onto its image.*

PROOF. Suppose M carries a topology with the properties described, and $\mathcal{O} \subset M$ is an open subset. Then each of the sets $\mathcal{O}_\alpha := \mathcal{O} \cap \mathcal{U}_\alpha$ is open, and $\mathcal{O} = \bigcup_{\alpha\in I} \mathcal{O}_\alpha$. Since each x_α is a homeomorphism onto its image in $\mathbb{R}^n$, $x_\alpha(\mathcal{O}_\alpha)$ is then also an open subset of $\mathbb{R}^n$. Conversely, if $\mathcal{O} \subset M$ is *any* subset such the sets $\Omega_\alpha := x_\alpha(\mathcal{O} \cap \mathcal{U}_\alpha) \subset \mathbb{R}^n$ are all open, then each $\mathcal{O}_\alpha := \mathcal{O} \cap \mathcal{U}_\alpha = x_\alpha^{-1}(\Omega_\alpha) \subset M$ must also be open since x_α is a homeomorphism, and therefore so is the union $\mathcal{O} = \bigcup_{\alpha\in I} \mathcal{O}_\alpha$. This proves that a topology with the stated properties is unique: if it exists, then it is precisely the collection of all subsets $\mathcal{O} \subset M$ such that $x_\alpha(\mathcal{O} \cap \mathcal{U}_\alpha) \subset \mathbb{R}^n$ is open for every $\alpha \in I$.

To prove existence, one now has to prove that the collection of subsets of M described above satisfies the axioms of a topology, i.e. it contains M and $\varnothing$ and is closed under arbitrary unions and finite intersections. This is a straightforward exercise.

Finally, let us fix the topology on M described above and suppose $(\mathcal{U}, x)$ is another chart that is C^0-compatible with $(\mathcal{U}_\alpha, x_\alpha)$ for every $\alpha \in I$. We need to show that $\mathcal{U} \subset M$ is open and $x : \mathcal{U} \to \mathbb{R}^n$ is a homeomorphism onto its image, which is equivalent to showing that for subsets $\mathcal{O} \subset \mathcal{U}$, $\mathcal{O}$ is open in M if and only if $x(\mathcal{O})$ is open in $\mathbb{R}^n$. For this, we make use of the transition maps relating $(\mathcal{U}, x)$ and $(\mathcal{U}_\alpha, x_\alpha)$ for an arbitrary choice of $\alpha \in I$:

$$\begin{tikzcd}
 & & \mathcal{U} \cap \mathcal{U}_\alpha \arrow[dl, "x"'] \arrow[dr, "x_\alpha"] & & \\
\mathbb{R}^n \overset{\text{open}}{\supset} & x(\mathcal{U} \cap \mathcal{U}_\alpha) \arrow[rr, bend left, "x_\alpha \circ x^{-1}"] & & x_\alpha(\mathcal{U} \cap \mathcal{U}_\alpha) \arrow[ll, bend left, "x \circ x_\alpha^{-1}"] & \overset{\text{open}}{\subset} \mathbb{R}^n
\end{tikzcd}$$

By the assumption of C^0-compatibility, the two maps in the bottom row of this diagram are both continuous, and since they are inverse to each other, they are homeomorphisms, meaning they define a bijection between the open subsets of $x(\mathcal{U} \cap \mathcal{U}_\alpha)$ and $x_\alpha(\mathcal{U} \cap \mathcal{U}_\alpha)$. Now suppose $\mathcal{O} \subset M$ is open, which means $x_\alpha(\mathcal{O} \cap \mathcal{U}_\alpha) \subset x_\alpha(\mathcal{U} \cap \mathcal{U}_\alpha) \subset \mathbb{R}^n$ is open for every α. Feeding this set into the homeomorphism $x \circ x_\alpha^{-1}$ gives $x(\mathcal{O} \cap \mathcal{U}_\alpha)$, proving that the latter is an open set, and therefore so is $x(\mathcal{O}) = \bigcup_{\alpha\in I} x(\mathcal{O} \cap \mathcal{U}_\alpha)$. Conversely, if $\mathcal{O} \subset M$ is an arbitrary subset such that $x(\mathcal{O})$ is open, then for every $\alpha \in I$, $x(\mathcal{O} \cap \mathcal{U}_\alpha)$ is the intersection of two open sets $x(\mathcal{O})$ and $x(\mathcal{U} \cap \mathcal{U}_\alpha)$, and is thus also open. Feeding it into $x_\alpha \circ x^{-1}$ then shows that $x_\alpha(\mathcal{O} \cap \mathcal{U}_\alpha)$ is also open, proving that $\mathcal{O} \subset M$ is open. □

Whenever we discuss a set M with an atlas $\mathcal{A}$ from now on, we will assume that M is endowed with the topology described in Proposition 2.12.

REMARK 2.13. Notice that according to the last statement in Proposition 2.12, the topologies induced on M by $\mathcal{A}$ or any extension of $\mathcal{A}$ to a larger (e.g. maximal) atlas are the same.

REMARK 2.14. It is rarely actually necessary to apply Proposition 2.12 for defining a topology on a manifold. The much more common situation is that our manifold M comes equipped with some natural topology that is clear from the context (e.g. because M is a subset or quotient of $\mathbb{R}^n$ or some other manifold that we already understand), and when specifying an atlas $\mathcal{A} = \{(\mathcal{U}_\alpha, x_\alpha)\}_{\alpha\in I}$ for M,

[12] Recall that $x_\alpha(\mathcal{U}_\alpha)$ is an open subset of a Euclidean space $\mathbb{R}^n$, so it is understood in this statement to carry the obvious topology that it inherits from the Euclidean metric on $\mathbb{R}^n$.

we just need to check that the topology determined by the atlas is the same as the natural topology. In other words, we need to check that the sets $\mathcal{U}_\alpha$ are open and the maps $x_\alpha : \mathcal{U}_\alpha \to x_\alpha(\mathcal{U}_\alpha) \subset \mathbb{R}^n$ are all homeomorphisms with respect to the natural topology. In most situations, this will be obvious.

EXERCISE 2.15. We now have two ways of defining what it means for a function $f : M \to \mathbb{R}$ to be continuous: one is the case $k = 0$ of Definition 2.3, in terms of the atlas $\mathcal{A}$, and the other is the standard notion of continuity in topological spaces, using the topology determined by $\mathcal{A}$ according to Proposition 2.12. Convince yourself that these two definitions are equivalent.

Since the atlas identifies small neighborhoods in M with neighborhoods in Euclidean space, and the topology of Euclidean space is pleasantly familiar to us, one might intuitively expect the topology induced on M by $\mathcal{A}$ to have similarly pleasant properties. The next example shows that this intuition is wrong.

EXAMPLE 2.16. Define an equivalence relation $\sim$ on the set $\widetilde{M} := \mathbb{R} \times \{0, 1\}$ such that every element is equivalent to itself and $(t, 0) \sim (t, 1)$ for all $t \in \mathbb{R}\backslash\{0\}$, but not for $t = 0$. Let

$$M := \widetilde{M}/\sim$$

denote the set of equivalence classes. We can think of M intuitively as a "real line with two zeroes", because it mostly looks just the same as $\mathbb{R}$ (each number $t \neq 0$ corresponding to the equivalence class of $(t, 0)$ and $(t, 1)$), but $t = 0$ is an exception, where there really are *two* distinct points $[(0, 0)]$ and $[(0, 1)]$ in M. The following pair of 1-dimensional charts define a smooth atlas on M: let

$$\mathcal{U}_\alpha := \left\{[(t, 0)] \in M \;\middle|\; t \in \mathbb{R}\right\}, \qquad \mathcal{U}_\beta := \left\{[(t, 1)] \in M \;\middle|\; t \in \mathbb{R}\right\},$$

and define both $x_\alpha : \mathcal{U}_\alpha \to \mathbb{R}$ and $x_\beta : \mathcal{U}_\beta \to \mathbb{R}$ by $[(t, k)] \mapsto t$ for $k = 0, 1$. The transition maps relating these two charts are both the identity map on $\mathbb{R}\backslash\{0\}$, thus the charts are smoothly compatible, and clearly $M = \mathcal{U}_\alpha \cup \mathcal{U}_\beta$.

Now consider the sequence

$$p_j := [(1/j, 0)] \in M.$$

Does it converge? We need to think for a moment about what convergence means in the topology induced by an atlas: if $p \in \mathcal{U}_\alpha$, then since x_α is a homeomorphism onto its image, p_j will converge to p if and only if $x_\alpha(p_j)$ converges to $x_\alpha(p)$ in $\mathbb{R}$, and a moment's thought reveals that that condition holds for $p := [(0, 0)]$. However, if we use the *other* chart x_β, then since $(1/j, 0) \sim (1/j, 1)$ for every j, the same condition also holds for the point $p' := [(0, 1)] \in \mathcal{U}_\beta$, and we have thus found two *distinct* points $p \neq p'$ such that $p_j \to p$ and $p_j \to p'$.

This seems like a contradiction if you have not seen any topology before, but it is not: it merely shows that M is a much stranger topological space than our intuition about metric spaces had led us to expect. In fact, the points p and p' have the peculiar property that every neighborhood of p intersects every neighborhood of p', so even though they are distinct points, the topology of M does not "separate" them; the technical term for this is that the topology of M is not **Hausdorff**.[13]

We do not want our notion of manifolds to include pathological examples in which a sequence can converge to two distinct points at once. Among other issues, it would clearly be impossible to define a metric compatible with that notion of convergence, as the triangle inequality ensures that limits of sequences are unique in metric spaces. Since the notion of distance on manifolds is one of the main things we plan to study when we get further into this subject, we would like to have a guarantee that every manifold *admits* a metric that is compatible with its natural topology,

[13] Or, as my topology professor in grad school once put it, the points p and p' are not "housed off" from each other. The proper delivery of this joke requires a Brooklyn accent.

i.e. we will insist that all manifolds be metrizable. This condition will turn out to have many advantages beyond the study of distance, though we will rarely need to make explicit use of it: it will only become important when we discuss the construction of global geometric structures (such as Riemannian metrics) via partitions of unity.

Although it will play no significant role in this course, we need one more topological notion in order to understand the main definition: a topological space is called **separable** (*separabel*) if it contains a countable dense subset. Euclidean spaces, for example, are separable, because $\mathbb{Q}^n \subset \mathbb{R}^n$ is a countable dense subset. Every space of interest in this course will be separable, and one can often use the result of the following exercise to prove it.

EXERCISE 2.17. Show that every subset of a separable metric space (X, d) is also a separable metric space.
Hint: Given a countable dense subset $E \subset X$ and another subset $Y \subset X$, show first that every open set in X is a union of open balls of the form $B_r(x) := \{y \in X \mid d(y, x) < r\}$ for $x \in E$ and $r \in \mathbb{Q}$. (This depends on the density of E.) Then define $E_0 \subset Y$ to consist of exactly one element from each of the sets $B_r(x) \cap Y$ for $x \in E$ and $r \in \mathbb{Q}$, whenever those sets are nonempty. Show that E_0 is countable and dense in Y.

2.3. The definition of a manifold. Hopefully you now have sufficient motivation to accept the following definition.

DEFINITION 2.18. Assume $k \in \mathbb{N} \cup \{\infty\}$. A **differentiable manifold of class** C^k (*differenzierbare Mannigfaltigkeit von der Klasse* C^k) or C^k**-manifold** (C^k*-Mannigfaltigkeit*) is a set M endowed with a C^k-structure (see Definition 2.6) such that the induced topology on M is metrizable and separable. In the case $k = \infty$, we also call M a **smooth manifold** (*glatte Mannigfaltigkeit*). We say that M is n**-dimensional** and refer to M as an n**-manifold**, written

$$\dim M = n,$$

if every chart in its differentiable structure is n-dimensional.[14]

REMARK 2.19. For the purposes of this course, you are essentially free to ignore the separability condition in Definition 2.18, as nothing in our study of differential geometry will truly depend on it. An example of something that satisfies every condition in the definition except separability would be the disjoint union of *uncountably* many copies of a manifold (see §2.4.3 below for more on disjoint unions); in fact, one can show that the condition on separability in our definition is equivalent to requiring M to have at most countably many connected components. One does sometimes need to know this for important results in differential *topology*, e.g. there is a theorem guaranteeing that every smooth n-manifold M can be embedded as a smooth submanifold of $\mathbb{R}^{2n+1}$, and this would clearly contradict Exercise 2.17 if M were not separable. (This issue is related to the second countability axiom—see Remark 2.21.)

REMARK 2.20. If you prefer never to think about topological spaces, then you can read Definition 2.18 as saying that a manifold M is a separable metric space endowed with an atlas $\{(\mathcal{U}_\alpha, x_\alpha)\}_{\alpha \in I}$ for which the sets $\mathcal{U}_\alpha \subset M$ are open and the bijections $x_\alpha : \mathcal{U}_\alpha \to x_\alpha(\mathcal{U}_\alpha) \subset \mathbb{R}^n$ are continuous with continuous inverses. Calling M a "metric space" comes however with the following caveat: while the *existence* of a suitable metric on M is an important condition, the *choice* of metric on M is not considered a part of its intrinsic structure, i.e. you are free to replace it with any other metric that has the above properties with respect to the atlas. This is why we have used the word "metrizable" in Definition 2.18 instead of just calling M a "metric space".

[14]Note that in our general definition of a manifold, M might admit multiple charts of different dimensions. One can show however that each individual connected component of M is itself a manifold with a uniquely defined dimension. For this reason we will usually only consider manifolds that have a well-defined dimension.

Remark 2.21. For students who have seen some topology, the more standard definition of a manifold found in many textbooks would replace the conditions of metrizability and separability with the conditions that M is *Hausdorff* and *second countable*. This gives an equivalent definition, though proving this equivalence would require more of a digression into point-set topology than we have space for here; the details can (mostly) be found in [**Lee11**, Chapter 2].

Remark 2.22. Another reasonable guess for a good definition of a manifold would be to drop metrizability and separability from Definition 2.18 but still require M to be Hausdorff (thus excluding things like Example 2.16). It turns out that this also does not include enough conditions to rule out some pathological behavior. The issue here is that a locally Euclidean Hausdorff space may fail to be *paracompact*, in which case the construction of basic geometric objects like Riemannian metrics becomes impossible. (We will discuss paracompactness and its applications later in the course.) If you have some topological background and would like to see some examples of the kinds of pathological behavior I'm talking about, see the discussion of the *long line* and *Prüfer surface* in [**Wen18**, Lecture 18].

In this course, we will almost always consider only the case $k = \infty$ of Definition 2.18, so that we speak of *smooth* manifolds. Actually, a large portion of differential geometry still makes sense for C^1-manifolds, though the important notion of *curvature* on a Riemannian manifold depends on second derivatives of the metric, and thus only makes sense on manifolds of class C^2. In either case, one has to be very careful in every proof so as not to differentiate anything more times than is allowed, and since the most important examples of manifolds are of class C^∞, it is conventional to avoid this annoyance by restricting attention to the smooth case. There is an additional reason to allow this restriction: according to a standard theorem in differential topology (see [**Hir94**, Theorem 2.9]), every manifold of class C^1 can be made into a *smooth* manifold by removing some of the charts in its maximal C^1-atlas. In this sense, one does not lose any significant generality by ignoring manifolds that are differentiable but not smooth.

You may have noticed on the other hand that Definition 2.18 also makes sense for $k = 0$, though in this case one cannot use the word "differentiable"; manifolds of class C^0 are called **topological manifolds** (*topologische Mannigfaltigkeiten*). These really are a different beast than differentiable manifolds: for every $n \geqslant 4$, there exist topological n-manifolds that *do not admit* any differentiable structure, i.e. their topology is not compatible with any atlas of class C^k for $k \geqslant 1$. Proving such things typically requires very advanced techniques, e.g. from mathematical gauge theory, which uses nonlinear PDEs to derive topological restrictions on smooth manifolds. (The classic introduction to this subject is [**DK90**].) In any case, the study of topological manifolds as such belongs squarely to the subject of topology, not differential geometry, so we will say no more about it here.

2.4. Some basic examples.

2.4.1. *Vector spaces.* For each integer $n \geqslant 0$, $\mathbb{R}^n$ admits a canonical smooth atlas consisting of a single n-dimensional chart, namely $(\mathbb{R}^n, \mathrm{Id})$. The smoothness of this atlas is a triviality: since there is only one chart, there is only one transition map to consider, which is the identity map and is therefore smooth. The unique extension of this atlas to a maximal smooth atlas on $\mathbb{R}^n$ defines what we will call the **standard smooth structure** on $\mathbb{R}^n$. The topology induced by this atlas is the standard one, which can also be defined in terms of the standard Euclidean metric; this follows via Remark 2.14 from the observations that $\mathbb{R}^n \subset \mathbb{R}^n$ is an open subset and $\mathrm{Id} : \mathbb{R}^n \to \mathbb{R}^n$ is a homeomorphism. It follows that $\mathbb{R}^n$ with its standard smooth structure is metrizable and (in light of the countable dense subset $\mathbb{Q}^n \subset \mathbb{R}^n$) separable. We conclude that $\mathbb{R}^n$ is, in a natural way, a smooth n-dimensional manifold. Note that it is *possible* to define different smooth structures on

$\mathbb{R}^n$, as shown by Example 2.7 in the case $n = 1$, but whenever we discuss $\mathbb{R}^n$ as a manifold in this course, we will always assume unless stated otherwise that it carries its standard smooth structure.

Since every real n-dimensional vector space V is isomorphic to $\mathbb{R}^n$, one can always choose such an isomorphism $\Phi : V \to \mathbb{R}^n$ and similarly regard V as a smooth n-manifold with an atlas consisting of the global chart (V, Φ). While the choice of isomorphism Φ here is typically not canonical, the resulting smooth structure on V is, since any other choice of isomorphism $\Psi : V \to \mathbb{R}^n$ would produce a chart (V, Ψ) that is related to (V, Φ) by the transition map $\Phi \circ \Psi^{-1} : \mathbb{R}^n \to \mathbb{R}^n$. The latter is a vector space isomorphism, and thus a smooth map with a smooth inverse. In this way, we can regard every real n-dimensional vector space naturally as a smooth n-manifold.

2.4.2. *Open subsets.* If M is an n-dimensional C^k-manifold with atlas $\mathcal{A} = \{(\mathcal{U}_\alpha, x_\alpha)\}_{\alpha \in I}$, then any open subset $\mathcal{O} \subset M$ admits a natural atlas

$$\mathcal{A}_{\mathcal{O}} := \{(\mathcal{U}_\alpha \cap \mathcal{O}, x_\alpha|_{\mathcal{U}_\alpha \cap \mathcal{O}})\}_{\alpha \in I} ,$$

which is also of class C^k since its transition maps are all restrictions of transition maps from $\mathcal{A}$ to open subsets. The key point here is that since $\mathcal{O} \subset M$ is open, each $\mathcal{U}_\alpha \cap \mathcal{O}$ is an open subset of $\mathcal{U}_\alpha$ and is thus mapped homeomorphically by x_α to another open subset of $\mathbb{R}^n$, making it an n-dimensional chart on $\mathcal{O}$. This atlas endows $\mathcal{O}$ with a natural C^k-structure, and since it is a subset of a separable metrizable space, Exercise 2.17 implies that it is also separable and metrizable, and is thus an n-dimensional C^k-manifold. Combining this with §2.4.1, we can now regard every open subset of $\mathbb{R}^n$ as a smooth n-manifold in a natural way.

2.4.3. *Disjoint unions.* The **disjoint union** (*disjunkte Vereinigung*) of a collection of sets $\{M_j\}_{j \in J}$ can be defined as the set

$$\coprod_{j \in J} M_j := \left\{ (j, t) \;\middle|\; j \in J,\ t \in M_j \right\} .$$

Here J can be an arbitrary index set, finite, countable or uncountable. In the special case where J is finite, e.g. if $J = \{1, \dots, N\}$, we also use the notation

$$M_1 \sqcup \ldots \sqcup M_N := \coprod_{j=1}^{N} M_j := \coprod_{j \in \{1,\dots,N\}} M_j .$$

Identifying each of the individual sets M_j with the subset $\{j\} \times M_j \subset \coprod_{j \in J} M_j$, we can think of $\coprod_{j \in J} M_j$ as literally a union of all the sets M_j, with the caveat that for $j \neq k$, M_j and M_k are always *disjoint* as subsets of $\coprod_{j \in J} M_j$, even if as abstract sets they have elements in common. For example, the set $S^1 \sqcup S^1$ contains two copies of every point on the circle, and is thus not the same set as $S^1 \cup S^1 = S^1$. If you think of S^1 as the unit circle in $\mathbb{R}^2$, then the definition above gives $S^1 \sqcup S^1 = \{1, 2\} \times S^1 \subset \mathbb{R}^3$, so the disjoint union consists of two copies of the circle that live in disjoint planes in $\mathbb{R}^3$.

Suppose now that each of the sets M_j is a C^k-manifold with atlas $\mathcal{A}^{(j)} = \{(\mathcal{U}_\alpha^{(j)}, x_\alpha^{(j)})\}_{\alpha \in I_j}$. Regarding each set M_j as a subset of $\coprod_{j \in J} M_j$ makes each of the sets $\mathcal{U}_\alpha^{(j)}$ also into subsets of $\coprod_{j \in J} M_j$, such that $\mathcal{U}_\alpha^{(j)} \cap \mathcal{U}_\beta^{(k)} = \varnothing$ whenever $j \neq k$. It follows that the union

$$\mathcal{A} := \bigcup_{j \in J} \mathcal{A}^{(j)}$$

defines an atlas of class C^k on $\coprod_{j \in J} M_j$, whose set of transition maps is just the union of the sets of transition maps for all the atlases $\mathcal{A}^{(j)}$. (Transition maps relating two charts $(\mathcal{U}_\alpha^{(j)}, x_\alpha^{(j)})$ with $(\mathcal{U}_\beta^{(k)}, x_\beta^{(k)})$ with $j \neq k$ do not arise here since their overlap is always empty.)

It does *not* follow however that every disjoint union of a collection of C^k-manifolds is naturally a C^k-manifold—this is one of the few situations where we have to pay attention to the condition of separability. The topology induced by the atlas $\mathcal{A}$ on $\coprod_{j\in J} M_j$ is the so-called **disjoint union topology**, in which a subset $\mathcal{O} \subset \coprod_{j\in J} M_j$ is open if and only if $\mathcal{O} \cap M_j$ is an open subset of M_j for every $j \in J$. If the sets M_j are nonempty for uncountably many distinct values of $j \in J$, then no *countable* subset $E \subset \coprod_{j\in J} M_j$ can have an element in every one of the subsets M_j, and it follows that E cannot be dense, so the disjoint union cannot be separable. On the other hand, one can show (see Exercise 2.23 below) that every *finite* or *countable* disjoint union of separable metrizable spaces is also separable and metrizable. We conclude that for any $N \in \mathbb{N} \cup \{\infty\}$ and any finite or countable collection $\{M_j\}_{j=1}^N$ of C^k-manifolds, the disjoint union $\coprod_{j=1}^N M_j$ is also a C^k-manifold in a natural way. Moreover, if $\dim M_j = n$ for every j, then the disjoint union is also n-dimensional.

EXERCISE 2.23.

(a) Show that for any metric space (X, d), the formula

$$d'(x, y) := \begin{cases} d(x, y) & \text{if } d(x, y) < 1, \\ 1 & \text{if } d(x, y) \geqslant 1 \end{cases}$$

defines another metric d' on X that induces the same topology as d.

(b) Show that for any collection of metric spaces $\{(X_j, d_j)\}_{j\in J}$ with $d_j(x, y) \leqslant 1$ for all $j \in J$ and $x, y \in X_j$, the formula

$$d(x, y) := \begin{cases} d_j(x, y) & \text{if } x, y \in X_j \text{ for some } j \in J, \\ 2 & \text{if } x \in X_j \text{ and } y \in X_k \text{ for some } j, k \in J \text{ with } j \neq k \end{cases}$$

defines a metric on $\coprod_{j\in J} X_j$ that induces the disjoint union topology.

(c) Show that the metric d on $\coprod_{j\in J} X_j$ in part (b) is separable if J is a finite or countable set and all of the metric spaces (X_j, d_j) are separable.

EXERCISE 2.24. Recall that a metrizable space[15] is called **compact** (*kompakt*) if every open covering has a finite subcover. Show that a disjoint union $\coprod_{j\in J} M_j$ is compact if and only if J is finite and M_j is compact for every $j \in J$.

2.4.4. *Dimension zero.* You may not have thought about the case $n = 0$ when we defined the notion of an n-dimensional chart, but the definition in that case does make sense: $\mathbb{R}^0$ consists of a single point, and its only nontrivial open subset is itself, so if $(\mathcal{U}, x)$ is a 0-dimensional chart on M, then $\mathcal{U} \subset M$ is a single point. It follows that if M is a 0-dimensional manifold with atlas $\mathcal{A} = \{(\mathcal{U}_\alpha, x_\alpha)\}_{\alpha\in I}$, then every point of M is its own open set, implying that *every* subset of M is open. This is known as the **discrete topology**, and it is always metrizable; a suitable metric is the **discrete metric**, defined by

$$d(x, y) := \begin{cases} 0 & \text{if } x = y, \\ 1 & \text{if } x \neq y. \end{cases}$$

The only dense subset of M in this topology is M itself, so separability requires M to be finite or countable. We conclude: a 0-dimensional manifold is simply a *finite or countable discrete set*, and it is compact if and only if it is finite. Equivalently, every 0-dimensional manifold can be identified with the disjoint union of at most countably many copies of the manifold $\mathbb{R}^0$, which is a single

[15]In fact this definition is also valid for arbitrary topological spaces.

point. Notice that since every map from $\mathbb{R}^0$ to itself is trivially smooth, every atlas on a 0-manifold is automatically a smooth atlas.

2.4.5. *Dimension one.* We have seen two explicit examples thus far of 1-dimensional manifolds: $\mathbb{R}$ and S^1, where the former carries its standard smooth structure as defined in §2.4.1, and the latter has a smooth structure that we defined using two charts based on polar coordinates in Lecture 1. We can now add to this list arbitrary open subsets of each, and arbitrary finite or countable disjoint unions of such open subsets. In this entire list, the only actual *compact* examples are S^1 and its finite disjoint unions; the compactness of the circle $S^1 \subset \mathbb{R}^2$ follows from the general fact that closed and bounded subsets of Euclidean space are compact. Up to a natural notion of equivalence for smooth manifolds that we will discuss in the next lecture, it turns out that these really are the *only* examples: in particular, every compact and *connected* 1-manifold is "diffeomorphic" to S^1. Later when we discuss manifolds with boundary, we will have to add the compact interval $[0, 1]$ to the list of compact 1-manifolds up to diffeomorphism. Similarly, it turns out that every noncompact connected 1-manifold is diffeomorphic to $\mathbb{R}$. We will not prove such classification results in this course, nor make use of them, but the curious reader will find a sketch of the corresponding result about connected topological 1-manifolds up to homeomorphism in [**Wen18**, Lecture 18]. Note that this is one of the important results that becomes false if one drops the metrizability condition from the definition of a manifold; we already saw one peculiar counterexample in Example 2.16, and another is the so-called "long line", which is essentially a union of *uncountably* many compact intervals glued together at their end points (see [**Wen18**, Lecture 18] or [**Spi99a**, Appendix to Chapter 1]).

2.4.6. *Cartesian products.* Since we have no plans to discuss infinite-dimensional manifolds in this course, we will not talk about infinite products, but finite products still provide a useful way of producing new manifolds from old ones. Assume M and N are C^k-manifolds of dimensions m and n respectively, with atlases $\mathcal{A} = \{(\mathcal{U}_\alpha, x_\alpha)\}_{\alpha \in I}$ on M and $\mathcal{B} = \{(\mathcal{V}_\beta, y_\beta)\}_{\beta \in J}$ on N. For each $(\alpha, \beta) \in I \times J$, one can then define a **product chart** on $M \times N$ with domain $\mathcal{U}_\alpha \times \mathcal{V}_\beta$ by

$$\mathcal{U}_\alpha \times \mathcal{V}_\beta \to \mathbb{R}^{m \times n} : (p, q) \mapsto (x_\alpha(p), y_\beta(q)).$$

Each of the transition maps relating two product charts is just the Cartesian product of a transition map from $\mathcal{A}$ with one from $\mathcal{B}$, thus they are all of class C^k, and the collection of all product charts therefore defines an atlas of class C^k and makes $M \times N$ into a C^k-manifold of dimension $m + n$.[16] One can of course repeat this construction finitely many times to make any finite product of manifolds $M_1 \times \ldots \times M_N$ into a manifold.

An important special case of this construction is the compact smooth n-manifold known as the n-**torus**, defined by

$$\mathbb{T}^n := \underbrace{S^1 \times \ldots \times S^1}_{n}.$$

In the case $n = 1$, this is just another name for the circle, but the most popular torus is the case $n = 2$: as we've defined it, $\mathbb{T}^2$ is literally a subset of $\mathbb{R}^4$, but for visualization purposes there is also a favorite way of embedding it in $\mathbb{R}^3$, as shown in Figure 5.

[16] Note that even if $\mathcal{A}$ and $\mathcal{B}$ are maximal atlases, the set of all product charts is generally not maximal, but this is immaterial since it has a unique maximal extension.

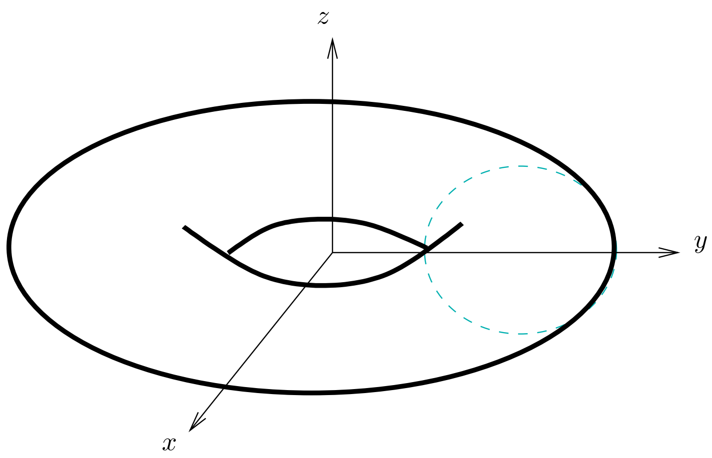

FIGURE 5. A representation of the torus $\mathbb{T}^2$ as a submanifold of $\mathbb{R}^3$.

The n-torus for $n \geqslant 3$ is less straightforward to visualize, but it is often useful to think of it[17] as the quotient of $\mathbb{R}^n$ by the lattice $\mathbb{Z}^n$, using the bijection

$$\mathbb{R}^n/\mathbb{Z}^n \to \underbrace{S^1 \times \ldots \times S^1}_{n} : [(\theta_1, \ldots, \theta_n)] \mapsto (e^{2\pi i\theta_1}, \ldots, e^{2\pi i\theta_n}),$$

where for computational convenience we have replaced $\mathbb{R}^2$ with $\mathbb{C}$ in order to describe points in the unit circle S^1 as complex exponentials. Under this identification, a point in $\mathbb{T}^n$ is represented by a vector in $\mathbb{R}^n$, with the understanding that two vectors represent the same point in the torus if and only if they differ by a vector with integer coordinates. This perspective is especially useful in the study of Fourier series, as a function $f : \mathbb{R}^n \to \mathbb{C}$ that is 1-periodic in each of the n variables can now be regarded equivalently as a function $f : \mathbb{T}^n \to \mathbb{C}$.

EXERCISE 2.25. Convince yourself that the natural smooth structure on $\underbrace{\mathbb{R} \times \ldots \times \mathbb{R}}_{n}$ derived from the standard smooth structure of $\mathbb{R}$ is the same as the standard smooth structure of $\mathbb{R}^n$.

2.4.7. *The projective plane and the Klein bottle.* We conclude with two explicit examples of surfaces (i.e. smooth 2-manifolds) that are somewhat harder to visualize, because they cannot be embedded in $\mathbb{R}^3$.[18]

The **projective plane** (*projektive Ebene*) is the set of equivalence classes

$$\mathbb{RP}^2 := S^2/\sim,$$

where the equivalence relation is defined by $p \sim p$ and $p \sim -p$ for all $p \in S^2 \subset \mathbb{R}^3$, meaning that every point p in the unit sphere gets identified with its *antipodal* point $-p$. (For more on why this might be a natural object to define, see Exercise 2.26 below.) If you have ever been on a long-haul international flight, then you are familiar with the notion of traversing a continuous path along S^2. In order to picture a continuous path on $\mathbb{RP}^2$, you should imagine that there are

[17] In fact, many sources in the literature prefer to *define* $\mathbb{T}^n$ as the quotient group $\mathbb{R}^n/\mathbb{Z}^n$, in which case its smooth structure can be derived from the standard smooth structure of $\mathbb{R}^n$ using a general result about quotients by discrete group actions.

[18] The claim that embedding them into $\mathbb{R}^3$ is *impossible* is something I expect you to find plausible, but not obvious. Proving it would require some methods from topology which we do not yet have at our disposal in this course, though we may come back to this later.

always two identical and interchangeable airplanes, containing identical copies of the same crews and passengers, constrained to fly at exact antipodal points over the Earth. If one of those airplanes flies from Shanghai to Buenos Aires while the other one flies along the antipodal path,[19] then since the two planes are completely interchangeable, they can be understood to describe a *closed loop* on $\mathbb{RP}^2$. Got it? Good.

It is relatively easy to see that $\mathbb{RP}^2$ is a smooth 2-manifold in a natural way. First, it has a natural metric, in which one can describe each point of $\mathbb{RP}^2$ as a set consisting of two points in S^2 and define the distance between two points in $\mathbb{RP}^2$ as the distance between those two sets. The fact that S^2 is separable (as a subset of the separable metric space $\mathbb{R}^3$) implies easily that $\mathbb{RP}^2$ is also separable. One can also derive a smooth atlas on $\mathbb{RP}^2$ from the one that we already constructed on S^2 in Exercise 1.7: the only issue is that some of the charts need to have their domains shrunk so that they no longer contain any pairs of antipodal points, as the coordinate map will otherwise fail to be injective, but this can easily be done.

The second example is the **Klein bottle** (*Kleinsche Flasche*), a picture of which is shown in Figure 6. The picture must be interpreted with caution, since what it shows is not really a manifold in the usual sense, but if you imagine perturbing part of it in an unseen fourth dimension so that part of the surface no longer has to pass through another part, then you get the right intuition. The picture also shows a "grid" structure similar to the coordinate grid one would obtain on $\mathbb{T}^2$ after identifying it with $\mathbb{R}^2/\mathbb{Z}^2$, but the Klein bottle is not the same thing as the torus. The latter can be identified with the quotient

$$(\mathbb{R} \times (\mathbb{R}/\mathbb{Z})) \Big/ \sim$$

by the smallest equivalence relation on $\mathbb{R} \times (\mathbb{R}/\mathbb{Z})$ such that $(s, [t]) \sim (s+1, [t])$ for all $s, t \in \mathbb{R}$. One obtains a rigorous definition of the Klein bottle from this via a reversal of orientation: instead of $(s, [t]) \sim (s+1, [t])$, one takes the smallest equivalence relation on $\mathbb{R} \times (\mathbb{R}/\mathbb{Z})$ such that

$$(s, [t]) \sim (s+1, [-t])$$

for all $s, t \in \mathbb{R}$. If you think about what grid lines of the form $\{s = \text{const}\}$ and $\{t = \text{const}\}$ look like in the set of equivalence classes defined via this relation, you will end up with something resembling Figure 6. It is not difficult to construct an atlas of smoothly compatible 2-dimensional charts on this quotient: the basic idea is to view it as a quotient of $\mathbb{R}^2$, and restrict the canonical global chart of $\mathbb{R}^2$ to neighborhoods that are sufficiently small so as to contain at most one element from every equivalence class.

EXERCISE 2.26. The projective plane is the $n = 2$ case of the **real projective n-space** (*reeller projektiver Raum*)

$$\mathbb{RP}^n := S^n / \sim,$$

where here again the equivalence relation identifies antipodal points $x \sim -x \in S^n \subset \mathbb{R}^{n+1}$. A useful interpretation of this definition comes from the observation that there is a unique line through the origin passing through each pair of points $\{x, -x\} \subset \mathbb{R}^{n+1}$. One can therefore view $\mathbb{RP}^n$ equivalently as the space of all *lines through the origin in* $\mathbb{R}^{n+1}$, which can be defined more precisely as the quotient

$$\mathbb{RP}^n = (\mathbb{R}^{n+1} \backslash \{0\}) / \sim$$

where two nontrivial vectors $v, w \in \mathbb{R}^{n+1}$ are now considered equivalent if and only if $v = \lambda w$ for some $\lambda \in \mathbb{R}$. From this perspective, it is convenient to denote points in $\mathbb{RP}^n$ via so-called

[19]According to the British science fiction TV series *Torchwood*, Buenos Aires and Shanghai are at exact antipodal points on the Earth. Wikipedia says this is true up to an error of about 400km. Let's just pretend it's true.

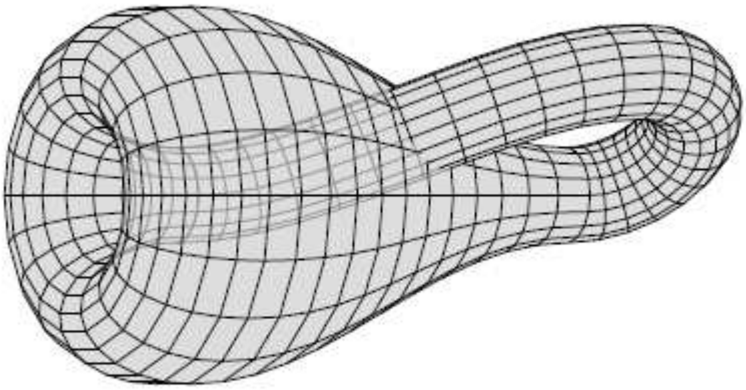

FIGURE 6. An immersion of the Klein bottle into $\mathbb{R}^3$. It is not an embedding because it intersects itself. (We will discuss the precise meanings of the words "immersion" and "embedding" in Lecture 4.)

homogeneous coordinates, in which the symbol

$$[x_0 : \ldots : x_n] \in \mathbb{RP}^n$$

means the equivalence class containing the vector $(x_0, \ldots, x_n) \in \mathbb{R}^{n+1}\backslash\{0\}$.

The homogeneous coordinates can be used to define an explicit smooth atlas on $\mathbb{RP}^n$. For $j = 0, \ldots, n$, define

$$\mathcal{U}_j := \left\{[x_0 : \ldots : x_n] \in \mathbb{RP}^n \;\middle|\; x_j \neq 0\right\}$$

and a map $\varphi_j : \mathbb{R}^n \to \mathbb{RP}^n$ by

$$\varphi_j(t_1, \ldots, t_n) := [t_1 : \ldots : t_j : 1 : t_{j+1} : \ldots : t_n].$$

Show that φ_j is an injective map onto $\mathcal{U}_j$, so $(\mathcal{U}_j, \varphi_j^{-1})$ is a chart, and compute the transition maps relating any two of the charts constructed in this way for different values of $j = 0, \ldots, n$. Show that these $n + 1$ charts together form a smooth atlas.

3. Smooth maps and tangent vectors

We have several more definitions to get through before the subject of differential geometry gets seriously underway. In this lecture we clarify what it means for a map between two manifolds to be differentiable, and what kind of object its derivative is.

3.1. Smooth maps between manifolds. We defined in §2.1 what it means for a real-valued function on a smooth manifold to be smooth (see Definition 2.3). The following is based on the same idea.

DEFINITION 3.1. Assume M and N are manifolds of dimensions m and n respectively, with differentiable structures $\mathcal{A}_M$ and $\mathcal{A}_N$ of class C^k. A continuous map $f : M \to N$ is said to be **of class** C^r for some $r \leqslant k$ (or **smooth** in the case $r = k = \infty$) if for every pair of charts $(\mathcal{U}, x) \in \mathcal{A}_M$ and $(\mathcal{V}, y) \in \mathcal{A}_N$, the map

$$\mathbb{R}^m \overset{\text{open}}{\supset} x(\mathcal{U} \cap f^{-1}(\mathcal{V})) \xrightarrow{y \circ f \circ x^{-1}} y(\mathcal{V}) \overset{\text{open}}{\subset} \mathbb{R}^n$$

is of class C^r.

In other words, a map $f : M \to N$ is of class C^r if it looks like a map of class C^r when expressed in local coordinates on both the domain and the target. The assumption $r \leqslant k$ is again crucial here, and guarantees that for any given point $p \in M$, the question of whether f is of class C^r near p does not depend on the charts one has to choose near $p \in M$ and $f(p) \in N$. Note that we had to explicitly assume f was continuous in this definition: this assumption guarantees that $f^{-1}(\mathcal{V}) \subset M$ is an open set, so that $x(\mathcal{U} \cap f^{-1}(\mathcal{V}))$ is open in $\mathbb{R}^n$, and differentiability on this domain can therefore be checked.

The set of C^k maps from M to N is often denoted by

$$C^k(M, N) = \left\{ f : M \to N \;\middle|\; f \text{ is of class } C^k \right\}.$$

One can endow this space with various natural topologies to make it into a topological (and sometimes also metrizable) space, though you should be aware that it is generally not a vector space, since N is not. On the other hand, the special case $N = \mathbb{R}$ is quite important, and is often abbreviated

$$C^k(M) := C^k(M, \mathbb{R}).$$

This *is* a vector space in a natural way, i.e. real-valued functions on a manifold M can be added and multiplied by constants.

Exercise 3.2. Show that for the standard smooth structure on $\mathbb{R}$ defined in §2.4.1, the notion of differentiability for a map $f : M \to \mathbb{R}$ as given in Definition 3.1 matches our previous definition for real-valued functions (Definition 2.3).

Up until this point I have been including non-smooth manifolds in the picture. I could continue doing this, but it would require frequently including slightly annoying extra hypotheses (like $r \leqslant k$) in statements of results, and the generality one gains by doing this does not fully compensate for the annoyance, so I will mostly assume $k = \infty$ from now on.

We can now define the natural notion of equivalence for smooth manifolds.

Definition 3.3. For two smooth manifolds M and N, a smooth map $f : M \to N$ is called a **diffeomorphism** (*Diffeomorphismus*) if it is bijective and its inverse $f^{-1} : N \to M$ is also smooth. Two smooth manifolds are called **diffeomorphic** (*diffeomorph*) if there exists a diffeomorphism between them.

Exercise 3.4. Viewing S^1 as the unit circle in $\mathbb{C}$, the quotient group $\mathbb{R}^n/\mathbb{Z}^n$ admits a natural bijection to the n-torus $\mathbb{T}^n = S^1 \times \ldots \times S^1$, given by

$$\mathbb{R}^n/\mathbb{Z}^n \to \mathbb{T}^n : [(\theta_1, \ldots, \theta_n)] \mapsto (e^{2\pi i \theta_1}, \ldots, e^{2\pi i \theta_n}).$$

For each $v \in \mathbb{R}^n$, choose a neighborhood $\widetilde{\mathcal{U}}_v \subset \mathbb{R}^n$ of v that is small enough to contain at most one element from each equivalence class in $\mathbb{R}^n/\mathbb{Z}^n$, and use this to define an n-dimensional chart $(\mathcal{U}_v, x_v)$ of the form

$$\mathcal{U}_v = \left\{ [w] \in \mathbb{R}^n/\mathbb{Z}^n \;\middle|\; w \in \widetilde{\mathcal{U}}_v \right\}, \qquad x_v([w]) = w.$$

Show that the collection of all charts of this form determines a smooth atlas on $\mathbb{R}^n/\mathbb{Z}^n$ such that the bijection to $\mathbb{T}^n$ described above is a diffeomorphism.

3.2. Tangent and cotangent spaces. Let us start this discussion with a concrete example: on the unit sphere $S^2 \subset \mathbb{R}^3$, a *tangent vector* to S^2 at a point $p \in S^2$ is by definition any vector of the form

$$\gamma'(0) \in \mathbb{R}^3,$$

where $\gamma : (-\epsilon, \epsilon) \to S^2$ is any choice of smooth path in $\mathbb{R}^3$ whose image is in S^2 and satisfies $\gamma(0) = p$. It should be easy to convince yourself that the set of all vectors of this form is a linear subspace of $\mathbb{R}^3$, namely, it is the orthogonal complement of p. We would now like to generalize this notion to an arbitrary smooth manifold, without needing to assume that is a subset of some Euclidean space.

For the rest of this subsection, assume M is a smooth manifold and $p \in M$. Having defined what a smooth map between manifolds is, we can fix the standard smooth structure on small intervals such as $(-\epsilon, \epsilon) \subset \mathbb{R}$ and talk about smooth maps $\gamma : (-\epsilon, \epsilon) \to M$. If $\gamma(0) = p \in M$, then we will refer to any such smooth map as a **path through p in M**. Note that the value of $\epsilon > 0$ here is not fixed, so it is allowed to be arbitrarily small.

Let us say that two paths α, β through p in M are **tangent** if for some some chart $(\mathcal{U}, x)$ with $p \in \mathcal{U}$,

$$\frac{d}{dt}(x \circ \alpha)\bigg|_{t=0} = \frac{d}{dt}(x \circ \beta)\bigg|_{t=0}.$$

It is easy to show that this condition does not depend on the choice of chart: indeed, if $(\mathcal{V}, y)$ is another chart with $p \in \mathcal{V}$, then for all t close enough to 0 so that $\alpha(t) \in \mathcal{U} \cap \mathcal{V}$, we have $(y \circ \alpha)(t) = (y \circ x^{-1}) \circ (x \circ \alpha)(t)$ and thus by the chain rule,

$$(y \circ \alpha)'(0) = D(y \circ x^{-1})(x(p))(x \circ \alpha)'(0), \tag{3.1}$$

where $D(y \circ x^{-1})(x(p)) : \mathbb{R}^n \to \mathbb{R}^n$ denotes the derivative of the transition map $y \circ x^{-1}$ at $x(p)$, which is an invertible linear map since $y \circ x^{-1}$ is smooth and has a smooth inverse. Since $(y \circ \beta)'(0)$ is related to $(x \circ \beta)'(0)$ in the same way, it is equal to $(y \circ \alpha)'(0)$ if and only if $(x \circ \beta)'(0) = (x \circ \alpha)'(0)$.

Definition 3.5. A **tangent vector** (*Tangentialvektor*) to M at p is an equivalence class $[\gamma]$ of paths γ through p in M, where two paths are considered equivalent if and only if they are tangent. The set of all tangent vectors to M at p is called the **tangent space** (*Tangentialraum*) to M at p, and is denoted by

$$T_pM = \left\{[\gamma] \;\middle|\; \gamma \text{ a path through } p \text{ in } M\right\}.$$

This definition of T_pM has many intuitive advantages, but it leaves several details unclear, foremost among them the fact that T_pM is a vector space. In order to see this, we'll need to make more use of coordinates.

Proposition 3.6. *The tangent space T_pM has a unique vector space structure such that for any smooth n-dimensional chart $(\mathcal{U}, x)$ with $p \in \mathcal{U}$, the map*

$$d_px : T_pM \to \mathbb{R}^n : [\gamma] \mapsto (x \circ \gamma)'(0) \tag{3.2}$$

is a vector space isomorphism. In particular, every tangent space of a smooth n-manifold is naturally an n-dimensional vector space.

Proof. The map (3.2) is a bijection by definition, so one can clearly always *choose* a chart $(\mathcal{U}, x)$ and define a vector space structure on T_pM so as to make this map an isomorphism. The point is then to show that any other choice of chart $(\mathcal{V}, y)$ would have given the same vector space structure on T_pM. This follows from the formula

$$d_py \circ (d_px)^{-1} = D(y \circ x)(x(p)) : \mathbb{R}^n \to \mathbb{R}^n,$$

which follows from (3.1) and shows that this transformation is itself a vector space isomorphism. □

Example 3.7. If M is an open subset of an n-dimensional vector space V, then the derivative $\gamma'(0)$ for a smooth path $\gamma : (-\epsilon, \epsilon) \to V$ can be defined in the classical way as a vector in V, giving rise to a canonical map

$$T_pM \to V : [\gamma] \mapsto \gamma'(0)$$

for every $p \in M$. It is a straightforward exercise to show that this map is a vector space isomorphism.

In the future, we shall always use this isomorphism to identify tangent spaces on open subsets of a vector space V with V itself, so that we do not need to talk about equivalence classes of paths. In particular, every tangent space on an open subset of $\mathbb{R}^n$ is in this way canonically identified with $\mathbb{R}^n$. We will see in §4.3 below that whenever N is a submanifold of M, one can also naturally regard T_pN for each $p \in N$ as a linear subspace of T_pM, so in the special case where N is a submanifold of $\mathbb{R}^n$, its tangent spaces will all naturally be subspaces of $\mathbb{R}^n$. This means that for the vast majority of examples we are interested in, it will not be necessary to use the original definition in terms of equivalence classes of paths for describing a tangent space.

Exercise 3.8. Show that for two smooth manifolds M, N and any two points $p \in M$ and $q \in N$, there is a canonical vector space isomorphism $T_{(p,q)}(M \times N) = T_pM \times T_qN$.

In linear algebra, it is often useful to associate to any vector space V its **dual space** (*Dualraum*), which is the space of all scalar-valued linear maps on V. Assuming V is a *real* (rather than complex) vector space, this can be denoted by

$$V^* := \mathrm{Hom}(V, \mathbb{R}),$$

where for two real vector spaces V, W in general we denote by $\mathrm{Hom}(V, W)$ the vector space of linear maps $V \to W$. When V is a tangent space T_pM on a manifold M, its dual space is called the **cotangent space** (*Kotangentialraum*) to M at p and denoted by

$$T_p^*M := \mathrm{Hom}(T_pM, \mathbb{R}).$$

Its elements are called **cotangent vectors** (*Kotangentialvektoren*), or sometimes also **covectors**.

Remark 3.9. Among physicists, covectors are often called "covariant vectors", while ordinary tangent vectors are called "contravariant vectors". I will not use this terminology.

3.3. The tangent bundle. The usefulness of the following definition will probably not be obvious to you at first glance, but it will become more apparent when we start differentiating smooth maps.

Definition 3.10. The **tangent bundle** (*Tangentialbündel*) TM of a smooth manifold M is the union of all its tangent spaces:

$$TM := \bigcup_{p \in M} T_pM.$$

The map $\pi : TM \to M$ such that $\pi^{-1}(p) = T_pM \subset TM$ for each $p \in M$ is called the **tangent projection**, and the subset in TM consisting of the zero vectors $0 \in T_pM$ for all $p \in M$ is called the **zero-section** (*Nullschnitt*) of TM. As subsets of TM, the individual tangent spaces $T_pM \subset TM$ for each $p \in M$ are sometimes referred to as the **fibers** (*Fasern*) of the tangent bundle.

Note that for distinct points $p \neq q \in M$, the tangent spaces T_pM and T_qM are by definition disjoint sets. Do not be tempted to think that the zero vector in T_pM is the same point as the zero vector in T_qM for $p \neq q$; in fact, there is a natural identification of the zero-section with M, giving rise to a natural inclusion

$$i : M \hookrightarrow TM : p \mapsto 0 \in T_pM. \tag{3.3}$$

At the level of set theory, we could just as well have used the disjoint union notation $\coprod_{p \in M} T_pM$ in Definition 3.10, but we did not do this because it would give a misleading impression about the topology and smooth structure we intend to define on TM.

Lemma 3.11. *On a manifold M, any n-dimensional chart $(\mathcal{U}, x)$ determines a $2n$-dimensional chart $(T\mathcal{U}, Tx)$ on the tangent bundle TM, where $T\mathcal{U} = \bigcup_{p \in \mathcal{U}} T_pM$ is the tangent bundle of the open* *subset $\mathcal{U} \subset M$, and $Tx : T\mathcal{U} \to \mathbb{R}^{2n}$ is defined in terms of the linear isomorphism $d_p x : T_pM \to \mathbb{R}^n$ of* (3.2) *by*

$$T\mathcal{U} \supset T_pM \ni X \mapsto (x(p), d_px(X)) \in \mathbb{R}^n \times \mathbb{R}^n = \mathbb{R}^{2n}.$$

If $(\mathcal{V}, y)$ is another chart on M, then transition maps relating the charts $(T\mathcal{V}, Ty)$ and $(T\mathcal{U}, Tx)$ on TM are given by

$$Ty \circ (Tx)^{-1}(q, v) = \left(y \circ x^{-1}(q), D(y \circ x^{-1})(q)v\right).$$

Proof. The map $Tx : T\mathcal{U} \to \mathbb{R}^{2n}$ is clearly injective, and its image is $x(\mathcal{U}) \times \mathbb{R}^n$, which is open. The stated formula for the transition map $Ty \circ (Tx)^{-1}$ follows from (3.1). □

Corollary 3.12. *For any smooth manifold M, the tangent bundle TM can be endowed naturally with the structure of a smooth manifold such that the tangent projection $\pi : TM \to M$, the inclusion $i : M \hookrightarrow TM$ of the zero-section* (3.3) *and the natural inclusions $T_pM \hookrightarrow TM$ for all $p \in M$ are smooth maps.*[20] *If $\dim M = n$, then $\dim TM = 2n$.*

Proof. We endow TM with the unique maximal smooth atlas containing all charts of the form $(T\mathcal{U}, Tx)$ determined via Lemma 3.11 from smooth charts $(\mathcal{U}, x)$ on M.

To check that $\pi : TM \to M$ is a smooth map, one can now write its coordinate expression with respect to any chart $(\mathcal{U}, x)$ on M and the corresponding chart $(T\mathcal{U}, Tx)$ on TM: the resulting map from an open subset of $\mathbb{R}^{2n}$ to $\mathbb{R}^n$ takes the form $(q, v) \mapsto q$, and is thus clearly smooth. Writing down the inclusion of the zero-section $M \hookrightarrow TM$ in similar coordinates produces $q \mapsto (q, 0)$, and for the inclusion $T_pM \hookrightarrow TM$, one obtains $v \mapsto (q, v)$. All of these maps are smooth.

I hope you find it plausible that TM with the atlas constructed above is metrizable and separable. Separability is easy to prove, e.g. one can take the union of countable dense subsets of individual fibers T_pM for all p in some countable dense subset of M, thus forming a countable dense subset of TM. The easiest way I can think of to prove metrizability is by constructing a Riemannian metric on TM, which we will do in Lecture 15. That construction will rely on the assumption that M is metrizable; we will not need to assume this about TM. □

Exercise 3.13. Find a diffeomorphism from the tangent bundle TS^1 to the product manifold $S^1 \times \mathbb{R}$.

One can similarly define a **cotangent bundle** (*Kotangentialbündel*)

$$T^*M := \bigcup_{p \in M} T_p^*M,$$

which satisfies a result analogous to Corollary 3.12. We will postpone the proof of this fact, since it follows from more general results about vector bundles to be discussed later in the course, and we will not really have use for it until then.

3.4. Tangent maps. We can now answer a question you may have wondered about: we know how to define whether a map $f : M \to N$ between manifolds is differentiable, but how does one actually *differentiate* it, i.e. what is its derivative at a point? In the special case $M \overset{\text{open}}{\subset} \mathbb{R}^m$ and $N = \mathbb{R}^n$, the answer you learned from first-year analysis is to view the derivative $Df(p)$ at a point $p \in M$ as a linear map $\mathbb{R}^m \to \mathbb{R}^n$, and according to the chain rule, it satisfies the relation

$$(f \circ \gamma)'(0) = Df(p)\gamma'(0)$$

for any smooth path γ through p. In fact, since any vector in $\mathbb{R}^m$ can be the derivative of some smooth path through p, this formula uniquely characterizes the linear map $Df(p) : \mathbb{R}^m \to \mathbb{R}^n$. It

[20] Here we are using the vector space structure of T_pM to regard it as a smooth manifold as in §2.4.1.

also admits an obvious generalization to the setting of smooth manifolds, using the fact that if $\gamma : (-\epsilon, \epsilon) \to M$ is a path through $p \in M$, then $f \circ \gamma : (-\epsilon, \epsilon) \to N$ is a path through $f(p) \in N$.

Definition 3.14. For two smooth manifolds M, N and a smooth map $f : M \to N$, the **tangent map** (*Tangentialabbildung*) of f is the map

$$Tf : TM \to TN : [\gamma] \mapsto [f \circ \gamma].$$

Its restriction to the tangent space at a specific point $p \in M$ can be denoted by

$$T_p f : T_p M \to T_{f(p)} N,$$

and is also called the **derivative** of f at p.[21]

Lemma 3.15. *The map $T_p f : T_p M \to T_{f(p)} N$ defined above for a smooth map $f : M \to N$ and a point $p \in M$ is independent of choices, and it is linear. Moreover, if $f : M \to N$ is smooth, then $Tf : TM \to TN$ is also smooth.*

Proof. All of these statements will become obvious if we write down a local coordinate expression for the map $Tf : TM \to TN$. Choose charts $(\mathcal{U}, x)$ on M and $(\mathcal{V}, y)$ on N with $p \in \mathcal{U}$ and $f(p) \in \mathcal{V}$. These give rise to charts $(T\mathcal{U}, Tx)$ on TM and $(T\mathcal{V}, Ty)$ on TN as in Lemma 3.11, so that given any $[\gamma] \in T_p M$, $Tx([\gamma]) = (x(p), (x \circ \gamma)'(0)) \in \mathbb{R}^m \times \mathbb{R}^m$, and according to the definition of Tf,

$$Ty(Tf([\gamma])) = (y(f(p)), (y \circ (f \circ \gamma))'(0)) \in \mathbb{R}^n \times \mathbb{R}^n.$$

The assumption that f is smooth means that $y \circ f \circ x^{-1}$ is smooth on its domain of definition, which is a neighborhood of $x(p)$ in $\mathbb{R}^m$. On this neighborhood, we can then write $y \circ (f \circ \gamma) = (y \circ f \circ x^{-1}) \circ (x \circ \gamma)$ and apply the chain rule to derive from the above expression,

$$Ty \circ Tf \circ (Tx)^{-1}(x(p), (x \circ \gamma)'(0)) = (y \circ f \circ x^{-1}(x(p)), D(y \circ f \circ x^{-1})(x(p))(x \circ \gamma)'(0)),$$

or if we simplify by writing $q := x(p) \in \mathbb{R}^m$ and $v := (x \circ \gamma)'(0) \in \mathbb{R}^m$,

$$Ty \circ Tf \circ (Tx)^{-1}(q, v) = (y \circ f \circ x^{-1}(q), D(y \circ f \circ x^{-1})(q)v).$$

This formula does not depend on any choice of path γ to represent the tangent vector $[\gamma] \in T_p M$, thus it proves that $Tf([\gamma])$ also does not depend on this choice, and moreover, it defines a smooth map $TM \to TN$ with a linear restriction $T_p M \to T_{f(p)} N$. □

The tangent bundle provides a more elegant language for talking about derivatives than was available in your first-year analysis course. As justification for this claim, I offer the following reformulation of the chain rule in the language of manifolds; it follows directly from the definitions of tangent spaces and tangent maps (which are in themselves crucially dependent on the chain rule from first-year analysis).

Proposition 3.16 (chain rule). *For any pair of smooth maps $f : M \to N$ and $g : N \to Q$ between smooth manifolds, $T(g \circ f) = Tg \circ Tf : TM \to TQ$.* □

Corollary 3.17. *If $f : M \to N$ is a diffeomorphism, then so is $Tf : TM \to TN$, and $(Tf)^{-1} = T(f^{-1}) : TN \to TM$.*

Proof. Observe first that the tangent map to the identity map on M is the identity map on TM. The chain rule then implies $\mathrm{Id}_{TM} = T(f \circ f^{-1}) = Tf \circ T(f^{-1})$. □

[21] You will find a variety of alternative notation in the literature for what I am calling $T_p f$, e.g. $df(p)$ and $Df(p)$ are also popular choices. In these notes, I will try to consistently reserve $Df(p)$ for the notion of derivatives defined in first-year analysis, where one only considers maps between open subsets of Euclidean spaces. The notation df will be reserved for the *differential* of a function valued in $\mathbb{R}$ or another vector space, to be defined in the next lecture.

Remark 3.18. Since $T_q\mathbb{R}^n$ is canonically isomorphic to $\mathbb{R}^n$ for every $q \in \mathbb{R}^n$, the tangent bundle $T\mathbb{R}^n$ has a canonical identification with $\mathbb{R}^n \times \mathbb{R}^n$ in which $T_q\mathbb{R}^n = \{q\} \times \mathbb{R}^n$. Under this identification, the chart $Tx : T\mathcal{U} \to \mathbb{R}^n \times \mathbb{R}^n$ on TM derived in Lemma 3.11 from a chart $x : \mathcal{U} \to \mathbb{R}^n$ on M is simply the tangent map of x.

Remark 3.19. If you are familiar with the language of categories and functors, then you might appreciate the following interpretation of Proposition 3.16. One can define a category Diff whose objects are the smooth manifolds, with morphisms $M \to N$ defined to be smooth maps, hence the isomorphisms in this category are the diffeomorphisms. The construction of the tangent bundle now gives rise to a functor $T : \mathsf{Diff} \to \mathsf{Diff}$ which sends each manifold M to TM and associates to any morphism $f : M \to N$ its tangent map $Tf : TM \to TN$. The formula $T(g \circ f) = Tg \circ Tf$ is the main step required for proving that T is a functor.

Remark 3.20. If M is a manifold of class C^k for some finite $k \in \mathbb{N}$, then the definition of tangent spaces requires a slight adjustment since the notion of *smooth* paths in M might not make sense; it is good enough however (and gives an equivalent definition) if we consider all paths $\gamma : (-\epsilon, \epsilon) \to M$ of class C^1. Inspecting the proof of Corollary 3.12 now reveals that TM is naturally a manifold of class C^{k-1}; one derivative is lost because the transition maps for TM involve derivatives of the transition maps for M. Similarly, if $f : M \to N$ is of class C^r with $1 \leqslant r \leqslant k$, then the tangent map $Tf : TM \to TN$ can be defined as a map of class C^{r-1}.

4. Submanifolds

The overarching message of this lecture will be that sometimes, understanding what is happening in a manifold is just a matter of finding the right coordinates.

4.1. Partial derivatives and differentials. There are two special situations in which the tangent map of $f : M \to N$ can be expressed in slightly more convenient forms. First, if $\mathcal{U} \subset \mathbb{R}^n$ is an open subset of Euclidean space, M is a manifold and $f : \mathcal{U} \to M$ is smooth, then f can be regarded (without needing to make a choice of coordinates) as an M-valued function of n variables, $f(x^1, \dots, x^n)$. For each point $x_0 = (x_0^1, \dots, x_0^n) \in \mathcal{U}$, f now determines n smooth paths through $f(x_0)$, namely

$$\gamma_j(t) := f(x_0^1, \dots, x_0^{j-1}, x_0^j + t, x_0^{j+1}, \dots, x_0^n), \qquad j = 1, \dots, n.$$

The equivalence classes of these paths are called the **partial derivatives** of f at x_0,

$$\partial_j f(x_0) := \frac{\partial f}{\partial x^j}(x_0) := [\gamma_j] \in T_{f(x_0)}M.$$

They are actually just particular values of the tangent map, i.e. $\partial_j f(x_0) = T_{x_0} f(e_j)$, where we are using the fact that $T_{x_0}\mathcal{U}$ is canonically isomorphic to $\mathbb{R}^n$ (see Example 3.7) and thus comes with a canonical basis $e_1, \dots, e_n$. The n tangent vectors $\partial_1 f(x_0), \dots, \partial_n f(x_0) \in T_{f(x_0)}M$ all together thus contain the same information as the tangent map $T_{x_0} f : T_{x_0}\mathcal{U} \to T_{f(x_0)}M$.

The second special situation is in some dense dual to the first: we consider a smooth function on a smooth manifold M with values in a finite-dimensional vector space V,

$$f : M \to V.$$

The most important special case of this is when $V = \mathbb{R}$, so that f is a real-valued function. Taking advantage again of the canonical isomorphisms $T_{f(p)}V = V$ from Example 3.7, we can rewrite $Tf(X) \in T_{f(p)}V$ for each $p \in M$ and $X \in T_pM$ as a vector in V, denoted by $df(X) \in V$. This associates to every smooth function $f : M \to V$ a smooth function

$$df : TM \to V,$$

called the **differential** (*Differential*) of f. We will denote its restriction to each individual tangent space T_pM for $p \in M$ by

$$d_pf : T_pM \to V.$$

In terms of equivalence classes of paths through p, a direct formula for d_pf is given by

$$d_pf([\gamma]) = (f \circ \gamma)'(0), \tag{4.1}$$

and one can deduce from Lemma 3.15 that this is independent of the choice of path γ in the equivalence class, and moreover, $d_pf : T_pM \to V$ is a linear map. In particular, for a smooth real-valued function $f : M \to \mathbb{R}$, d_pf is an element of the cotangent space at p,

$$d_pf \in T_p^*M \qquad (\text{for } f : M \to \mathbb{R}).$$

This makes the differentials df of smooth real-valued functions $f : M \to \mathbb{R}$ into our first examples of *differential forms*; we will have a lot more to say about them when we discuss integration in a few weeks.

EXAMPLE 4.1. The differentials defined above directly generalize the linear map $d_px : T_pM \to \mathbb{R}^n$ in (3.2), which can be associated to any smooth chart $(\mathcal{U}, x)$ on M and a point $p \in \mathcal{U}$. This map can also be constructed out of the differentials of the coordinate functions $x^1, \ldots, x^n : \mathcal{U} \to \mathbb{R}$; it is given by

$$d_px(X) = (d_px^1(X), \ldots, d_px^n(X)) \in \mathbb{R}^n.$$

4.2. The inverse function theorem. In the examples of manifolds we have dealt with so far, we have always had charts that were explicitly constructed, but such explicit constructions are not always convenient in more general situations. A nice tool for obtaining less explicit but often more useful constructions of charts is provided by the inverse function theorem from first-year analysis. Let us recall the statement:

THEOREM (inverse function theorem). *Suppose $\mathcal{U} \subset \mathbb{R}^n$ is open, $f : \mathcal{U} \to \mathbb{R}^n$ is a map of class C^k for some $k \in \mathbb{N} \cup \{\infty\}$, and $x_0 \in \mathcal{U}$ is a point at which the derivative $Df(x_0) : \mathbb{R}^n \to \mathbb{R}^n$ is an isomorphism. Then there exist open neighborhoods $x_0 \in \Omega \subset \mathcal{U}$ and $f(x_0) \in \Omega' \subset \mathbb{R}^n$ such that f maps Ω bijectively onto Ω' and the inverse $(f|_\Omega)^{-1} : \Omega' \to \Omega$ is also of class C^k.* □

We will now turn this standard analytical result into a pair of criteria for proving that certain maps we construct define smooth charts.

LEMMA 4.2. *Suppose M is a smooth n-manifold, $\mathcal{U} \subset \mathbb{R}^n$ is an open set, $\varphi : \mathcal{U} \to M$ is a smooth map and $x_0 \in \mathcal{U}$ is a point at which the partial derivatives $\partial_1\varphi(x_0), \ldots, \partial_n\varphi(x_0)$ form a basis of $T_{\varphi(x_0)}M$. Then there exist open neighborhoods $x_0 \in \Omega \subset \mathcal{U}$ and $p := \varphi(x_0) \in \mathcal{O} \subset M$ such that φ maps Ω bijectively onto $\mathcal{O}$ and $(\mathcal{O}, (\varphi|_\Omega)^{-1})$ defines a smooth chart on M.*

PROOF. Choose any smooth chart $(\mathcal{V}, y)$ on M with $p = \varphi(x_0) \in \mathcal{V}$, and observe that $d_py(\partial_j\varphi(x_0)) = \partial_j(y\circ\varphi)(x_0)$In for each $j = 1, \ldots, n$. Since $d_py : T_pM \to \mathbb{R}^n$ is an isomorphism, our assumption on the basis $\partial_1\varphi(x_0), \ldots, \partial_n\varphi(x_0) \in T_pM$ means that $\partial_1(y \circ \varphi)(x_0), \ldots, \partial_n(y \circ \varphi)(x_0)$ is similarly a basis of $\mathbb{R}^n$, which is equivalent to saying that the linear map $D(y \circ \varphi)(x_0) : \mathbb{R}^n \to \mathbb{R}^n$ is an isomorphism. The inverse function theorem thus provides open neighborhoods $x_0 \in \Omega \subset \mathcal{U}$ and $y(p) \in \Omega' \subset \mathbb{R}^n$ such that $y\circ\varphi$ is a diffeomorphism between Ω and Ω', implying that $\varphi = y^{-1}\circ(y\circ\varphi)$ sends Ω bijectively to an open neighborhood $\mathcal{O} := y^{-1}(\Omega')$ of p. Denoting the inverse of this bijection by $x : \mathcal{O} \to \Omega \subset \mathbb{R}^n$, the transition map $y \circ x^{-1}$ is now just $y \circ \varphi|_\Omega$, so it is smooth and has a smooth inverse. □

LEMMA 4.3. *Suppose M is a smooth n-manifold, $\mathcal{U} \subset M$ is an open set, $x^1, \ldots, x^n : \mathcal{U} \to \mathbb{R}$ are smooth functions and $p \in \mathcal{U}$ is a point such that the differentials $d_px^1, \ldots, d_px^n$ form a basis of*

T_p^*M. Then there exists an open neighborhood $p \in \mathcal{O} \subset \mathcal{U}$ such that $(\mathcal{O}, x)$ with $x := (x^1, \dots, x^n)$: $\mathcal{O} \to \mathbb{R}^n$ defines a smooth chart on M.

PROOF. Since $d_p x^1, \dots, d_p x^n$ is a basis of T_p^*M, it is dual to a unique basis $X_1, \dots, X_n$ of T_pM, meaning the two bases are related by

$$d_p x^i(X_j) = \delta^i_j := \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Define the linear map $d_p x := (d_p x^1, \dots, d_p x^n) : T_pM \to \mathbb{R}^n$ as in Example 4.1, so $d_p x$ is the tangent map $T_p x : T_pM \to T_{x(p)}\mathbb{R}^n$ after identifying $T_{x(p)}\mathbb{R}^n = \mathbb{R}^n$. Since $d_p x$ sends the basis $X_1, \dots, X_n$ to the standard basis of $\mathbb{R}^n$, it is an isomorphism. Now if $(\mathcal{V}, y)$ is any smooth chart with $p \in \mathcal{V}$, the map $x \circ y^{-1}$ is smooth on a neighborhood of p, and the chain rule gives

$$D(x \circ y^{-1})(y(p)) = d_p x \circ (d_p y)^{-1},$$

hence the latter is also an isomorphism $\mathbb{R}^n \to \mathbb{R}^n$. The inverse function theorem now provides open neighborhoods $y(p) \in \Omega \subset \mathbb{R}^n$ and $x(p) \in \Omega' \subset \mathbb{R}^n$ such that $x \circ y^{-1}$ is a diffeomorphism from Ω onto Ω', so $\mathcal{O} := y^{-1}(\Omega) = x^{-1}(\Omega')$ is then a neighborhood of p on which the restriction of x defines a chart that is smoothly compatible with $(\mathcal{V}, y)$. □

4.3. Slice charts. We have used the word "submanifold" already a few times in an informal way, e.g. the unit circle S^1 is a manifold that lives inside the manifold $\mathbb{R}^2$, so we called it a submanifold. It is now time to clarify more precisely what this word means.

The archetypal example of a submanifold is a linear subspace of a vector space, for instance

$$\mathbb{R}^\ell \times \{0\} = \left\{ (x^1, \dots, x^\ell, 0, \dots, 0) \in \mathbb{R}^n \;\middle|\; (x^1, \dots, x^\ell) \in \mathbb{R}^\ell \right\} \subset \mathbb{R}^n.$$

Basic results in linear algebra imply that *any* ℓ-dimensional subspace of an n-dimensional vector space looks like this example after a suitable linear change of coordinates. The notion of a smooth submanifold generalizes this by allowing nonlinear (but smooth) changes of coordinates.

DEFINITION 4.4. A chart $(\mathcal{U}, x)$ on an n-manifold M is called an ℓ-dimensional **slice chart** (*Bügelkarte*) for a subset $L \subset M$ if

$$L \cap \mathcal{U} = x^{-1}(\mathbb{R}^\ell \times \{0\}),$$

i.e. the points in $\mathcal{U}$ belong to L if and only if their coordinates $x^{\ell+1}, \dots, x^n$ vanish.

DEFINITION 4.5. Suppose M is a smooth n-manifold. A subset $L \subset M$ is called an **ℓ-dimensional smooth submanifold** (*Untermannigfaltigkeit*) of M if M admits a collection of smooth slice charts for L whose domains cover every point of L.

REMARK 4.6. More generally, if M is a manifold of class C^k but not necessarily smooth, one can speak of *submanifolds of class C^k*, in which the transition maps between slice charts are required to be of class C^k. Note that a C^k-manifold can also be regarded as a C^r-manifold for any $r \leqslant k$, so under this condition it makes sense to talk about C^r-submanifolds, but e.g. there is no such thing as a *smooth* submanifold of M if the latter is of class C^k for some $k < \infty$ but not equipped with a smooth structure.

EXAMPLE 4.7. The smooth structure we constructed on $S^1 \subset \mathbb{R}^2$ in Lecture 1 was obtained from polar coordinates by restricting to the unit circle $\{r = 1\}$; this gave rise to two charts $(\mathcal{U}, \theta)$ and $(\mathcal{V}, \phi)$, where θ and ϕ both had the meaning of an angle in polar coordinates, but with different ranges of values, namely $\theta(\mathcal{U}) = (0, 2\pi)$ and $\phi(\mathcal{V}) := (-\pi, \pi)$. These two coordinates were defined on open subsets of S^1, but they also have natural extensions to open subsets of $\mathbb{R}^2$, namely

$$\mathcal{U}' := \left\{ tv \in \mathbb{R}^2 \;\middle|\; v \in \mathcal{U},\ t > 0 \right\}, \qquad \mathcal{V}' := \left\{ tv \in \mathbb{R}^2 \;\middle|\; v \in \mathcal{V},\ t > 0 \right\}.$$

The radial coordinate r is defined on $\mathbb{R}^2\backslash\{0\}$ and takes all positive values; if we now set $\rho := r - 1$ so that $\{r = 1\} = \{\rho = 0\}$, we obtain a pair of smoothly compatible slice charts $(\mathcal{U}', (\theta, \rho))$ and $(\mathcal{V}', (\phi, \rho))$ for S^1 such that $S^1 \subset \mathcal{U}' \cup \mathcal{V}'$. This means that S^1 is a smooth submanifold of $\mathbb{R}^2$.

One can similarly turn the atlas for S^2 in Exercise 1.7 into a family of slice charts to prove that S^2 is a submanifold of $\mathbb{R}^3$. In practice, however, constructing slice charts by hand is not usually necessary, as we will see in §4.4 that some much more general and powerful tools for this purpose are provided by the inverse function theorem.

Let us first clarify the fact that a submanifold of a manifold is also a manifold in its own right.

PROPOSITION 4.8. *If L is an ℓ-dimensional C^k-submanifold of an n-dimensional C^k-manifold M, then L inherits naturally from M the structure of an ℓ-dimensional C^k-manifold such that the inclusion map $L \hookrightarrow M$ is of class C^k. Moreover, for each $p \in L$, the tangent space T_pL is naturally an ℓ-dimensional linear subspace of T_pM.*

PROOF. We associate to every slice chart $(\mathcal{U}, x)$ for $L \subset M$ a chart of the form $(\mathcal{U} \cap L, x_L)$ on L, where we use the coordinate projection $\pi_\ell(x^1, \ldots, x^n) := (x^1, \ldots, x^\ell)$ to define

$$x_L = \pi_\ell \circ x|_{\mathcal{U} \cap L} : \mathcal{U} \cap L \to \mathbb{R}^\ell.$$

By assumption, L can be covered by slice charts, so the collection of all charts of this form defines an atlas on L. Given two such charts $(\mathcal{U} \cap L, x_L)$ and $(\mathcal{V} \cap L, y_L)$ derived from two C^k-compatible slice charts $(x, \mathcal{U})$ and $(y, \mathcal{V})$, the transition map $y \circ x^{-1}$ preserves the subspace $\mathbb{R}^\ell \times \{0\} \subset \mathbb{R}^n$, and its restriction to the intersection of its domain with this subspace is the transition map $y_L \circ x_L^{-1}$, which is therefore of class C^k. Moreover, the fact that M is metrizable and separable implies the same for L by Exercise 2.17, thus L is a C^k-manifold. The local coordinate expression for the inclusion $i : L \hookrightarrow M$ with respect to any slice chart $(\mathcal{U}, x)$ and the associated chart $(\mathcal{U} \cap L, x_L)$ on L is $(x^1, \ldots, x^\ell) \mapsto (x^1, \ldots, x^\ell, 0, \ldots, 0)$, which is clearly smooth, thus the inclusion is of class C^k.[22] For each $p \in L$, the tangent map $T_pi : T_pL \to T_pM$ is simply the canonical inclusion $T_pL \hookrightarrow T_pM$ defined by regarding each path in L as a path in M. Since its image is a linear subspace, it gives a canonical isomorphism of T_pL to a linear subspce of T_pM. □

Whenever we speak of a submanifold $L \subset M$ from now on, we will assume that L is endowed with the differentiable structure described in Proposition 4.8, so that it can also be regarded as a manifold in its own right. We will often make use of the canonical identification of tangent spaces T_pL with subspaces of T_pM, especially in the case $M = \mathbb{R}^n$, where (in light of Example 3.7) this identification allows us to view each tangent space T_pL as a subspace of $\mathbb{R}^n$.

EXERCISE 4.9. Assume in the following that M and N are both C^k-manifolds and $f : M \to N$ is a map of class C^k. Prove:

(a) For any C^k-submanifold $L \subset M$, the restriction $f|_L : L \to N$ is also a map of class C^k.
(b) If $L \subset N$ is a C^k-submanifold such that $f(M) \subset L$, then the resulting map $f : M \to L$ is also of class C^k.

4.4. Immersions and submersions.

DEFINITION 4.10. A smooth map $f : M \to N$ is called an **immersion** at $p \in M$ if the linear map $T_pf : T_pM \to T_{f(p)}N$ is injective, and similarly, f is a **submersion** at p if $T_pf : T_pM \to$

[22] Recall that if both L and M are manifolds of class C^k but $k < \infty$, then it does not make sense to say that the inclusion $L \hookrightarrow M$ is smooth, even though it looks smooth in the particular local coordinates we chose. The point is that one could also choose different coordinates in which it would still appear to be a map of class C^k, but not necessarily C^∞.

$T_{f(p)}N$ is surjective. If one says that f is an immersion/submersion without specifying a point p, the meaning is that it is true for *all* points in M. One sometimes uses the notation

$$f : M \looparrowright N$$

to indicate when f is an immersion.

Recall that for any two finite-dimensional vector spaces V, W, the sets of linear maps $V \to W$ that are injective or surjective are open. It follows that if f is an immersion or submersion at some point $p \in M$, then this is also true on a *neighborhood* of p; equivalently, the set of points at which f is an immersion or submersion is open.

There is a good reason to single out these two particular classes of smooth maps between manifolds: it turns out that up to choices of smooth coordinates near $p \in M$ and $f(p) \in N$, all immersions look the same, and similarly for all submersions. This fact will give us a new user-friendly tool for identifying smooth submanifolds. The main tool required in its proof is the inverse function theorem, or more precisely, the two lemmas in §4.2 that used the inverse function theorem to construct charts.

THEOREM 4.11. *Assume M is a smooth m-manifold, N is a smooth n-manifold, $f : M \to N$ is a smooth map, $p \in M$ and $q = f(p) \in N$. If f is either an immersion or a submersion at p, then there exist smooth charts $(\mathcal{U}, x)$ on M with $x(p) = 0 \in \mathbb{R}^m$ and $(\mathcal{V}, y)$ on N with $y(q) = 0 \in \mathbb{R}^n$ such that the coordinate expression $y \circ f \circ x^{-1}$ for f is given by*

$$\mathbb{R}^m \ni (x^1, \dots, x^m) \mapsto \begin{cases} (x^1, \dots, x^n) \in \mathbb{R}^n & \text{if } m \geqslant n \text{ (submersion case)}, \\ (x^1, \dots, x^m, 0, \dots, 0) \in \mathbb{R}^n & \text{if } m < n \text{ (immersion case)}. \end{cases}$$

PROOF. Assume first that $T_p f : T_p M \to T_{f(p)} N$ is injective, so $n \geqslant m$, and set $\ell := n - m$. Choose a smooth chart $(\mathcal{U}, x)$ on M with $p \in \mathcal{U}$ and $x(p) = 0 \in \mathbb{R}^m$; note that the latter can be assumed without loss of generality by taking any chart with $p \in \mathcal{U}$ and composing the map $\mathcal{U} \to \mathbb{R}^n$ with a translation on $\mathbb{R}^n$ sending the image of p to the origin. With this understood, $\Omega := x(\mathcal{U}) \subset \mathbb{R}^m$ is an open neighborhood of the origin, and we observe that $F := f \circ x^{-1} : \Omega \to N$ is now a smooth map such that $F(0) = q$ and $T_0 F = T_p f \circ (d_p x)^{-1} : \mathbb{R}^m \to T_q N$ is injective. The latter is equivalent to the condition that the partial derivatives $\partial_1 F(0), \dots, \partial_m F(0) \in T_q N$ are linearly independent.

We claim that after possibly shrinking Ω to a smaller neighborhood of $0 \in \mathbb{R}^m$, and choosing $\epsilon > 0$ sufficiently small, $F : \Omega \to N$ can be extended to a smooth map

$$\widetilde{F} : \Omega \times (-\epsilon, \epsilon)^\ell \to N$$

such that $\widetilde{F}(x^1, \dots, x^m, 0, \dots, 0) = F(x^1, \dots, x^m)$ and the partial derivatives $\partial_1 \widetilde{F}, \dots, \partial_n \widetilde{F}$ at the origin form a basis of $T_q N$. This extension is not canonical, but it is also not difficult: if N were simply $\mathbb{R}^n$, we could define it by choosing any extension of the linearly independent set $\partial_1 F(0), \dots, \partial_m F(0)$ to a basis $\partial_1 F(0), \dots, \partial_m F(0), Y_{m+1}, \dots, Y_n$ of $T_q N$ and then defining

$$\widetilde{F}(x^1, \dots, x^n) := F(x^1, \dots, x^m) + \sum_{j=m+1}^n x^j Y_j.$$

This formula does not make sense in general if N is not a vector space, but one could more generally choose a chart on N near q in order to express F in local coordinates, and define the extension in this way in coordinates. Lemma 4.2 now implies that on a sufficiently small neighborhood of $0 \in \mathbb{R}^n$, $\widetilde{F}$ can be inverted to define a chart $(\mathcal{V}, y)$ on N with the stated properties.

Next suppose $T_p f : T_p M \to T_{f(p)} N$ is surjective, thus $m \geqslant n$, and we can set $\ell := m - n$. The idea now is to choose any chart $(\mathcal{V}, y)$ on N with $y(q) = 0$ and define the first n coordinates over the neighborhood $f^{-1}(\mathcal{V}) \subset M$ of p by

$$x^i := y^i \circ f, \qquad i = 1, \dots, n.$$

Writing $\hat{x} := (x^1, \dots, x^n) : f^{-1}(\mathcal{V}) \to \mathbb{R}^n$, we have $d_p\hat{x} = d_q y \circ T_p f$, thus $d_p\hat{x} : T_pM \to \mathbb{R}^n$ is surjective, which is equivalent to the condition that the n covectors $d_px^1, \dots, d_px^n \in T_p^*M$ are linearly independent.

To define the remaining ℓ coordinates on M near p, first choose an extension of the linearly-independent set $d_px^1, \dots, d_px^n$ to a basis $d_px^1, \dots, d_px^n, \Lambda^{n+1}, \dots, \Lambda^m$ of T_p^*M. For each $i = n+1, \dots, m$, we can then define a smooth function x^i on a neighborhood of p such that $x^i(p) = 0$ and $d_px^i = \Lambda^i$; this is another step that would be trivial to carry out if M were the vector space $\mathbb{R}^m$, so the idea is to choose a chart near p and write down suitable functions in local coordinates. With this done, Lemma 4.3 implies that after possibly shrinking to a smaller neighborhood $\mathcal{U} \subset M$ of p, $x = (x^1, \dots, x^m)$ becomes a smooth chart with the desired properties. □

Remark 4.12. For a continuous map $f : M \to N$ between topological manifolds, one can define f to be a *topological immersion* or *topological submersion* at $p \in M$ if there exist continuous charts near p and $q := f(p)$ in which f satisfies the coordinate formula in Theorem 4.11. Note that without having at least one continuous derivative at our disposal, there is no alternative way to characterize either of these conditions in terms of a tangent map being injective or surjective, nor is there any inverse function theorem available for proving such statements. On the other hand, Theorem 4.11 does make sense in the setting of C^k-manifolds for any $k \in \mathbb{N}$; in this case one must assume that $f : M \to N$ is of class C^k, and the resulting charts will be as well. (One should not be fooled by the fact that f will then *look* like a smooth map with respect to those charts—if $k < \infty$, it will not look smooth after arbitrary changes of C^k-coordinates.)

4.5. Embeddings and regular level sets. We now have enough technology to produce many more examples of submanifolds.

Definition 4.13. A smooth map $f : M \to N$ is called an **embedding** (*Einbettung*) if it is an injective immersion whose inverse $f(M) \xrightarrow{f^{-1}} M$ is also continuous. The notation

$$f : M \hookrightarrow N$$

is sometimes used to indicate that f is an embedding.

The typical example of an embedding is the natural inclusion $M \hookrightarrow N$ that exists whenever M is a submanifold of N. The next result states that, up to diffeomorphism, all examples are this one.

Theorem 4.14. *If $f : M \to N$ is an embedding, then its image $f(M)$ is a smooth submanifold of N.*

Proof. Suppose $q \in f(M)$. By injectivity, there is a unique point $p \in M$ such that $f(p) = q$, and Theorem 4.11 provides charts $(\mathcal{U}, x)$ on M and $(\mathcal{V}, y)$ on N with $x(p) = 0$ and $y(q) = 0$ such that $y \circ f \circ x^{-1}$ takes the form $(x^1, \dots, x^m) \mapsto (x^1, \dots, x^m, 0, \dots, 0)$. Since the inverse $f(M) \to M$ is also continuous, we are free to assume after possibly shrinking $\mathcal{V} \subset N$ to a smaller neighborhood of q that

$$f^{-1}(\mathcal{V} \cap f(M)) \subset \mathcal{U},$$

or in other words, $\mathcal{V} \cap f(M) = f(\mathcal{U})$. This proves that $(\mathcal{V}, y)$ is a slice chart for the subset $f(M)$. □

The following consequence appears in some books as an alternative definition of the notion of a submanifold:

COROLLARY 4.15. *A subset $L \subset M$ of a smooth manifold M is a smooth submanifold if and only if it admits a smooth structure for which the inclusion map $L \hookrightarrow M$ is a smooth embedding.* □

It is worth pausing a moment to consider what an immersion $f : M \looparrowright N$ can look like if it is *not* an embedding. Theorem 4.11 implies that every immersion is *locally* an embedding, i.e. for every $p \in M$, one can find a neighborhood $\mathcal{U} \subset M$ of p such that $f|_{\mathcal{U}} : \mathcal{U} \hookrightarrow N$ is an embedding and $f(\mathcal{U}) \subset N$ is therefore a submanifold. On the other hand, f may fail to be an embedding *globally* because it is not injective, meaning it has self-intersections $f(p) = f(p')$ with $p \neq p'$. The notation "$f : M \looparrowright N$" is meant to evoke this possibility by allowing the arrow to loop around and intersect itself. A classic example of a non-injective immersion is the picture of the Klein bottle in Figure 6, which shows the image of an immersion of a compact smooth 2-manifold into $\mathbb{R}^3$. Images of immersions are sometimes called **immersed submanifolds** in the literature, though I am personally not fond of this terminology,[23] so I will not use it.

For slightly subtler reasons, an injective immersion can also fail to be an embedding:

EXAMPLE 4.16. Let $N = \mathbb{R}^2$ and $M = \mathbb{R} \sqcup (0, \pi)$, and defne the immersion $f : M \looparrowright \mathbb{R}^2$ by

$$f(t) := (t, 0) \qquad \text{for } t \in \mathbb{R},$$
$$f(\theta) := (\cos\theta, \sin\theta) \qquad \text{for } \theta \in (0, \pi).$$

Omitting the points 0 and π from the interval $(0, \pi)$ makes this map an injective immersion, but the inverse $f(M) \xrightarrow{f^{-1}} M$ is discontinuous at the two points $(\pm 1, 0)$, which are precisely the points at which it fails to be a submanifold.

Turning our attention to submersions, we can now state a popular corollary of the implicit function theorem that you may have heard referred to before as the "regular value theorem".

DEFINITION 4.17. For a smooth map $f : M \to N$, $p \in M$ is called a **regular point** (*regulärer Wert*) of f if f is a submersion at p, and a **critical point** (*kritischer Wert*) otherwise. A point $q \in N$ is a **critical value** (*kritischer Wert*) of f if $q = f(p)$ for some critical point p, and q is otherwise called a **regular value** (*regulärer Wert*) of f.

THEOREM 4.18 (implicit function theorem). *For any smooth map $f : M \to N$ with regular value $q \in N$, $L := f^{-1}(q) \subset N$ is a smooth submanifold with $\dim L = \dim M - \dim N$, and its tangent space at any point $p \in L$ is $T_pL = \ker T_pf \subset T_pM$.*

PROOF. For each $p \in L = f^{-1}(q)$, f is by assumption a submersion at p, so Theorem 4.11 provides charts x near p and y near q such that $x(p)$ and $y(q)$ are both the origin in their respective Euclidean spaces and $y \circ f \circ x^{-1}$ becomes the map $(x^1, \ldots, x^m) \mapsto (x^1, \ldots, x^n)$. The zero-set of this map is a neighborhood of p in $f^{-1}(q)$ as seen in the x-coordinates, thus x is a slice chart. To see that $T_pL = \ker T_pf$, observe first that for any path γ in L through p, $f \circ \gamma$ is a constant path at $q \in N$, thus $T_pf([\gamma]) = 0 \in T_qN$, proving $T_pL \subset \ker T_pf$. The rest is dimension counting, as the surjectivity of $T_pf : T_pM \to T_qN$ implies

$$\dim T_pL = \dim L = \dim M - \dim N = \dim T_pM - \dim T_qN = \dim \ker T_pf.$$

□

[23]I have two objections to the term "immersed submanifold": first, it sounds as if it should be a type of submanifold, but it isn't. Second, one cannot always uniquely recover the manifold M from the image of an immersion $M \looparrowright N$. For example (the following is only for readers with a background in topology), a closed surface Σ_g of genus $g \geqslant 2$ admits smooth covering maps $\Sigma_h \to \Sigma_g$ by surfaces of arbitrarily large genus h (the degree of the cover will be correspondingly large). If one chooses an embedding of Σ_g into $\mathbb{R}^3$, one obtains a submanifold that is also the image of an immersion $\Sigma_h \looparrowright \mathbb{R}^3$ for arbitrarily large values of h.

Submanifolds of the form $f^{-1}(q) \subset M$ for regular values $q \in N$ are sometimes called **regular level sets** of f. In particular, a submersion $f : M \to N$ is distinguished by the property that *all* of its level sets are regular, and are thus smooth submanifolds.

4.6. Examples. We now have a *very* easy way of proving that simple examples like the unit spheres $S^n \subset \mathbb{R}^{n+1}$ really are smooth submanifolds.

EXAMPLE 4.19. Define $f : \mathbb{R}^{n+1} \to \mathbb{R}$ in terms of the standard Euclidean inner product by $f(x) = |x|^2 = \langle x, x \rangle$. This is a smooth map, with differential at any point $x \in \mathbb{R}^{n+1}$ given by $d_x f(v) = 2\langle x, v \rangle$, so it is a submersion everywhere except at the origin. This makes $S^n = f^{-1}(1)$ into a smooth submanifold of dimension $(n+1) - 1 = n$, so in particular, S^n inherits a natural smooth structure for which the inclusion $S^n \hookrightarrow \mathbb{R}^{n+1}$ is a smooth embedding. The kernel of $d_x f$ at a point $x \in S^n$ is the orthogonal complement of x, hence

$$T_x S^n = x^\perp \subset \mathbb{R}^{n+1}.$$

EXAMPLE 4.20. The smooth map $f : \mathbb{R}^2 \to \mathbb{R} : (x, y) \mapsto xy$ has only one critical point, at $(x, y) = (0, 0)$, thus $f^{-1}(t)$ is a smooth submanifold (a hyperbola) for every $t \neq 0$, and so is $f^{-1}(0) \backslash \{(0, 0)\}$, but $f^{-1}(0)$ fails to be a submanifold at the origin.

EXERCISE 4.21. Identifying the torus $\mathbb{T}^2$ with $\mathbb{R}^2/\mathbb{Z}^2$ via Exercise 3.4, find an explicit formula for an embedding $\mathbb{T}^2 \hookrightarrow \mathbb{R}^3$ whose image looks like Figure 5.

For the next set of exercises, the symbol $\mathbb{F}$ always denotes either the real numbers $\mathbb{R}$ or complex numbers $\mathbb{C}$, and we denote the vector space of m-by-n matrices over $\mathbb{F}$ by

$$\mathbb{F}^{m \times n} := \{m\text{-by-}n \text{ matrices over } \mathbb{F}\}.$$

If $\mathbb{F} = \mathbb{R}$, this is a real vector space of dimension mn. In the case $\mathbb{F} = \mathbb{C}$, it is a complex vector space of this same dimension, which means it can also be regarded as a *real* vector space of dimension $2mn$. (Indeed, if V is any complex vector space with complex basis $v_1, \ldots, v_k$, then a basis of V as a *real* vector space is given by $v_1, iv_1, \ldots, v_k, iv_k$.) Since they are vector spaces, $\mathbb{R}^{m \times n}$ and $\mathbb{C}^{m \times n}$ carry natural smooth structures and are thus smooth manifolds of dimensions mn and $2mn$ respectively. For $m = n$, there is a distinguished open subset

$$\mathrm{GL}(n, \mathbb{F}) = \left\{ \mathbf{A} \in \mathbb{F}^{n \times n} \;\middle|\; \mathbf{A} \text{ is invertible} \right\},$$

which is therefore also naturally a smooth manifold of dimension n^2 or (in the complex case) $2n^2$. That $\mathrm{GL}(n, \mathbb{F}) \subset \mathbb{F}^{n \times n}$ is open can be deduced easily from the observation that the determinant

$$\det : \mathbb{F}^{n \times n} \to \mathbb{F}$$

defines a continuous function for which $\mathrm{GL}(n, \mathbb{F}) = \det^{-1}(\mathbb{F} \backslash \{0\})$. In fact, $\det(\mathbf{A})$ is a polynomial in the entries of $\mathbf{A}$, which are all linear functions of $\mathbf{A}$, thus $\det : \mathbb{F}^{n \times n} \to \mathbb{F}$ is a smooth real- or complex-valued function. By Cramer's rule, the function

$$\mathrm{GL}(n, \mathbb{F}) \to \mathrm{GL}(n, \mathbb{F}) : \mathbf{A} \mapsto \mathbf{A}^{-1}$$

is also smooth.

EXERCISE 4.22. The n-dimensional **orthogonal group** $\mathrm{O}(n) \subset \mathbb{R}^{n \times n}$ is the set of all real n-by-n matrices $\mathbf{A}$ with the property

$$\mathbf{A}^T \mathbf{A} = \mathbb{1},$$

where $\mathbb{1}$ is the n-by-n identity matrix and $\mathbf{A}^T$ denotes the *transpose* of $\mathbf{A}$, i.e. if $\mathbf{A}$ has entries A_{ij}, then the corresponding entries of $\mathbf{A}^T$ are A_{ji}. This is precisely the set of all linear transformations $\mathbb{R}^n \to \mathbb{R}^n$ which preserve the Euclidean inner product, which means geometrically that they preserve lengths of vectors and angles between them. We will show in this exercise that $\mathrm{O}(n)$ is a smooth submanifold of $\mathbb{R}^{n \times n}$.

(a) Define the linear subspace consisting of all symmetric matrices,
$$\Sigma(n) := \left\{ \mathbf{A} \in \mathbb{R}^{n\times n} \;\middle|\; \mathbf{A} = \mathbf{A}^T \right\} \subset \mathbb{R}^{n\times n}.$$
There is a map
$$f : \mathbb{R}^{n\times n} \to \Sigma(n) : \mathbf{A} \mapsto \mathbf{A}^T\mathbf{A},$$
such that the orthogonal group is the level set $\mathrm{O}(n) = f^{-1}(\mathbb{1})$. The entries of $f(\mathbf{A})$ are quadratic functions of the entries of $\mathbf{A}$, thus f is clearly a smooth map. Show that its derivative at any $\mathbf{A} \in \mathbb{R}^{n\times n}$ is the linear map
$$Df(\mathbf{A}) : \mathbb{R}^{n\times n} \to \Sigma(n) : \mathbf{H} \mapsto \mathbf{A}^T\mathbf{H} + \mathbf{H}^T\mathbf{A}.$$
Hint: In theory you can do this by computing all the partial derivatives of f with respect to the entries of $\mathbf{A}$, but it's much, much *easier to use the definition of the derivative, i.e. regarding $\mathbb{R}^{n\times n}$ and $\Sigma(n)$ simply as vector spaces, show that a "remainder" formula of the form*
$$f(\mathbf{A}+\mathbf{H}) = f(\mathbf{A}) + Df(\mathbf{A})\mathbf{H} + R(\mathbf{H}) \cdot |\mathbf{H}|$$
with $\lim_{\mathbf{H}\to 0} R(\mathbf{H}) = 0$ is satisfied. One useful thing you may want to assume: for a reasonable choice of norm on $\mathbb{R}^{n\times n}$, matrix products satisfy $|\mathbf{AB}| \leqslant |\mathbf{A}||\mathbf{B}|$.

(b) Show that $Df(\mathbf{A})$ is surjective if $\mathbf{A} \in \mathrm{O}(n)$. In fact, you won't even need to assume $\mathbf{A} \in \mathrm{O}(n)$, but it *is* useful to assume that $\mathbf{A}$ is invertible (which is automatically true for orthogonal matrices). It is also *crucial* that the target space is $\Sigma(n)$ rather than the entirety of $\mathbb{R}^{n\times n}$—$Df(\mathbf{A})$ is certainly *not* surjective onto $\mathbb{R}^{n\times n}$.

(c) It follows now from the implicit function theorem that $\mathrm{O}(n)$ is a smooth submanifold of $\mathbb{R}^{n\times n}$. What is its dimension? (For a sanity check I will tell you: $\dim \mathrm{O}(2) = 1$ and $\dim \mathrm{O}(3) = 3$.)

(d) Show that $T_{\mathbb{1}}\,\mathrm{O}(n) \subset T_{\mathbb{1}}\mathbb{R}^{n\times n} = \mathbb{R}^{n\times n}$ is the space of all *antisymmetric* matrices $\mathbf{H}$, i.e. those which satisfy $\mathbf{H}^T = -\mathbf{H}$.

Exercise 4.23. The complex analogue of Exercise 4.22 involves the **unitary group**
$$\mathrm{U}(n) = \left\{ \mathbf{A} \in \mathbb{C}^{n\times n} \;\middle|\; \mathbf{A}^\dagger\mathbf{A} = \mathbb{1} \right\},$$
where $\mathbf{A}^\dagger$ denotes the Hermitian adjoint of $\mathbf{A}$, defined as the complex conjugate of its transpose. Prove that $\mathrm{U}(n)$ is a smooth submanifold of $\mathbb{C}^{n\times n}$, compute its dimension, and show
$$T_{\mathbb{1}}\,\mathrm{U}(n) = \left\{ \mathbf{H} \in \mathbb{C}^{n\times n} \;\middle|\; \mathbf{H}^\dagger = -\mathbf{H} \right\}.$$

Exercise 4.24. The **special linear group** over $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ is defined by
$$\mathrm{SL}(n, \mathbb{F}) = \left\{ \mathbf{A} \in \mathbb{F}^{n\times n} \;\middle|\; \det(\mathbf{A}) = 1 \right\}.$$

(a) Show that the derivative of $\det : \mathbb{F}^{n\times n} \to \mathbb{F}$ at $\mathbb{1}$ is given by the **trace** (*Spur*):
$$D(\det)(\mathbb{1})\mathbf{H} = \mathrm{tr}(\mathbf{H}).$$
Hint: Write $\mathbf{H}$ in terms of n column vectors as $\begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{pmatrix}$, so
$$\det(\mathbb{1} + t\mathbf{H}) = \det\begin{pmatrix} \mathbf{e}_1 + t\mathbf{v}_1 & \cdots & \mathbf{e}_n + t\mathbf{v}_n \end{pmatrix},$$
where $\mathbf{e}_1, \ldots, \mathbf{e}_n$ denotes the standard basis of $\mathbb{F}^n$. Differentiate this expression with respect to t at $t = 0$, using the fact that the determinant of a matrix is a multilinear function of its columns.

(b) Use the relation $\det(\mathbf{AB}) = \det(\mathbf{A}) \cdot \det(\mathbf{B})$ to generalize the formula in part (a) to
$$D(\det)(\mathbf{A})\mathbf{H} = \det(\mathbf{A}) \cdot \mathrm{tr}(\mathbf{A}^{-1}\mathbf{H}) \qquad \text{for any} \qquad \mathbf{A} \in \mathrm{GL}(n, \mathbb{F}).$$

(c) Prove that $\mathrm{SL}(n,\mathbb{F})$ is a smooth submanifold of $\mathbb{F}^{n\times n}$, compute its dimension, and show
$$T_{\mathbb{1}}\,\mathrm{SL}(n,\mathbb{F}) = \left\{\mathbf{H}\in\mathbb{F}^{n\times n}\;\middle|\;\mathrm{tr}(\mathbf{H})=0\right\}.$$
(d) Consider the set of *non-invertible* n-by-n matrices,
$$M := \left\{\mathbf{A}\in\mathbb{F}^{n\times n}\;\middle|\;\det(\mathbf{A})=0\right\}.$$
Is 0 a regular value of $\det:\mathbb{F}^{n\times n}\to\mathbb{F}$? Is M a submanifold of $\mathbb{F}^{n\times n}$?
Hint: Clearly M contains the trivial matrix $0\in\mathbb{F}^{n\times n}$. If M is a submanifold, what can you say about the tangent space $T_0M\subset\mathbb{F}^{n\times n}$? In how many different directions can you find smooth paths $\gamma:(-\epsilon,\epsilon)\to\mathbb{F}^{n\times n}$ through 0 that are contained in M?

Exercise 4.25. The **special orthogonal** and **special unitary** groups are defined as
$$\mathrm{SO}(n)=\mathrm{O}(n)\cap\mathrm{SL}(n,\mathbb{R}),\qquad\text{and}\qquad \mathrm{SU}(n)=\mathrm{U}(n)\cap\mathrm{SL}(n,\mathbb{C})$$
respectively. Prove:

(a) $\mathrm{SO}(n)$ is an open (and also closed) subset of $\mathrm{O}(n)$, hence it is a smooth submanifold with the same dimension and $T_{\mathbb{1}}\,\mathrm{SO}(n)=T_{\mathbb{1}}\,\mathrm{O}(n)$.
(b) $\mathrm{SU}(n)$ is a smooth submanifold of $\mathrm{U}(n)$ with $\dim\mathrm{SU}(n)=\dim\mathrm{U}(n)-1$, and
$$T_{\mathbb{1}}\,\mathrm{SU}(n) = \left\{\mathbf{H}\in\mathbb{C}^{n\times n}\;\middle|\;\mathbf{H}^\dagger=-\mathbf{H}\text{ and }\mathrm{tr}(\mathbf{H})=0\right\}.$$
Hint: Use Exercise 4.9 to show that the determinant defines a smooth map $\det:\mathrm{U}(n)\to S^1$, where S^1 in this case denotes the unit circle in $\mathbb{C}$. Prove that 1 is a regular value of this map.

Finally, we consider an interesting space of matrices that does not form a group, but is nonetheless a manifold.

Exercise 4.26. For $\mathbb{F}\in\{\mathbb{R},\mathbb{C}\}$ and nonnegative integers m,n and $r\leqslant\min\{m,n\}$, let
$$V_r(m,n,\mathbb{F}) := \left\{\mathbf{A}\in\mathbb{F}^{m\times n}\;\middle|\;\mathrm{rank}(\mathbf{A})=r\right\}.$$
By the standard formula relating ranks and kernels, $V_r(m,n,\mathbb{F})$ is the set of all m-by-n matrices $\mathbf{A}$ over $\mathbb{F}$ such that $\dim_{\mathbb{F}}\ker\mathbf{A}=n-r$, and the latter condition is also equivalent to $\dim_{\mathbb{F}}\mathrm{coker}\,\mathbf{A}=m-r$, where the **cokernel** of $\mathbf{A}$ is defined from its image $\mathrm{im}(\mathbf{A})\subset\mathbb{F}^m$ as the quotient space $\mathbb{F}^m/\mathrm{im}(\mathbf{A})$.

Given any $\mathbf{M}_0\in V_r(m,n,\mathbb{F})$, one can find splittings $\mathbb{F}^n=V\oplus K$ and $\mathbb{F}^m=W\oplus C$ such that $K=\ker\mathbf{M}_0$ and $W=\mathrm{im}\,\mathbf{M}_0$. Regarding any other matrix $\mathbf{M}\in\mathbb{F}^{m\times n}$ as a linear map $\mathbb{F}^n\to\mathbb{F}^m$, these splittings of $\mathbb{F}^n$ and $\mathbb{F}^m$ give rise to a block decomposition
$$\mathbf{M}=\begin{pmatrix}\mathbf{A}(\mathbf{M}) & \mathbf{B}(\mathbf{M})\\ \mathbf{C}(\mathbf{M}) & \mathbf{D}(\mathbf{M})\end{pmatrix}:V\oplus K\to W\oplus C,$$
thus defining linear (and therefore smooth) maps $\mathbf{A}:\mathbb{F}^{m\times n}\to\mathrm{Hom}(V,W)$, $\mathbf{B}:\mathbb{F}^{m\times n}\to\mathrm{Hom}(K,W)$, $\mathbf{C}:\mathbb{F}^{m\times n}\to\mathrm{Hom}(V,C)$ and $\mathbf{D}:\mathbb{F}^{m\times n}\to\mathrm{Hom}(K,C)$. By construction, the functions $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$ all vanish at $\mathbf{M}_0$, while $\mathbf{A}(\mathbf{M}_0):V\to W$ is invertible. Observe that the invertible maps in $\mathrm{Hom}(V,W)$ form an open subset; this is true for the same reason that $\mathrm{GL}(n,\mathbb{F})$ is an open subset of $\mathbb{F}^{n\times n}$. We can therefore fix an open neighborhood $\mathcal{O}\subset\mathbb{F}^{m\times n}$ of $\mathbf{M}_0$ such that $\mathbf{A}(\mathbf{M}):V\to W$ is invertible for all $\mathbf{M}\in\mathcal{O}$, and use this to define two smooth maps $\Phi:\mathcal{O}\to\mathrm{Hom}(K,C)$ and $\Psi:\mathcal{O}\to\mathbb{F}^{n\times n}$ by
$$\Phi(\mathbf{M}):=\mathbf{D}(\mathbf{M})-\mathbf{C}(\mathbf{M})\mathbf{A}(\mathbf{M})^{-1}\mathbf{B}(\mathbf{M}),\qquad\text{and}\qquad\Psi(\mathbf{M}):=\begin{pmatrix}\mathbb{1} & -\mathbf{A}(\mathbf{M})^{-1}\mathbf{B}(\mathbf{M})\\ 0 & \mathbb{1}\end{pmatrix},$$
where in the latter expression we are regarding $\Psi(\mathbf{M})$ as a linear map $\mathbb{F}^n\to\mathbb{F}^n$ and writing its block decomposition with respect to the splitting $\mathbb{F}^n=V\oplus K$.

(a) Show that $\Psi(\mathbf{M}) \in \mathbb{F}^{n \times n}$ is invertible for every $\mathbf{M} \in \mathcal{O}$.
(b) Show that for every $\mathbf{M} \in \mathcal{O}$, the kernel of the matrix product $\mathbf{M}\Psi(\mathbf{M}) : \mathbb{F}^n \to \mathbb{F}^m$ is $\{0\} \oplus \ker \Phi(\mathbf{M}) \subset V \oplus K = \mathbb{F}^n$.
(c) Deduce from parts (a) and (b) that $\mathcal{O} \cap V_r(m,n,\mathbb{F}) = \Phi^{-1}(0)$.
Hint: What is the largest dimension that $\ker \mathbf{M}$ *can have for* $\mathbf{M} \in \mathcal{O}$*?*
(d) Show that $\mathbf{M}_0$ is a regular point of Φ, and deduce from this that $V_r(m,n,\mathbb{F}) \subset \mathbb{F}^{m \times n}$ is a smooth submanifold with

$$T_{\mathbf{M}} V_r(m,n,\mathbb{F}) = \left\{ \mathbf{H} \in \mathbb{F}^{m \times n} \;\middle|\; \mathbf{H}(\ker \mathbf{M}) \subset \operatorname{im} \mathbf{M} \right\}$$

for every $\mathbf{M} \in V_r(m,n,\mathbb{F})$, and

$$\dim V_r(m,n,\mathbb{R}) = mn - (m-r)(n-r), \qquad \dim V_r(m,n,\mathbb{C}) = 2 \dim V_r(m,n,\mathbb{R}).$$

(e) A matrix $\mathbf{M} \in \mathbb{F}^{m \times n}$ is said to have **maximal rank** if its rank is $\min\{m,n\}$, which means it is either injective or surjective. Deduce from the result of part (d) that the set of maximal rank matrices is open and dense in $\mathbb{F}^{m \times n}$.

The result of this exercise produces what is called a ***stratification*** *of* $\mathbb{F}^{m \times n}$*, meaning that it decomposes* $\mathbb{F}^{m \times n}$ *into a collection of smooth submanifolds of various dimensions such that every matrix belongs to exactly one of them.*

5. Vector fields

A **vector field** (*Vektorfeld*) X on a smooth manifold M associates to every point $p \in M$ a vector in the corresponding tangent space,

$$X(p) \in T_p M.$$

For example, on $S^2 \subset \mathbb{R}^3$, the tangent space $T_p S^2$ is the orthogonal complement of the vector $p \in S^2 \subset \mathbb{R}^3$, thus a vector field associates to each such point another vector that is orthogonal to it. We say that a vector field X is **smooth** if the map

$$M \to TM : p \mapsto X(p)$$

is smooth. The set of all smooth vector fields on M forms a vector space, which we will denote by

$$\mathfrak{X}(M) := \left\{ X \in C^\infty(M, TM) \;\middle|\; X(p) \in T_p M \text{ for every } p \in M \right\}.$$

As with real-valued functions, one can define the **support** (*Träger*) of a vector field X as the closure in M of the set $\{p \in M \mid X(p) \neq 0\}$.

5.1. The flow of a vector field. The most important fact about vector fields on manifolds is that they determine dynamical systems. For a smooth path $\gamma : (a,b) \to M$, the derivative

$$\dot{\gamma}(t) := \frac{d\gamma}{dt}(t) \in T_{\gamma(t)} M$$

can be defined for each $t \in (a,b)$ as a special case of our definition of *partial* derivatives in §3.4. In important special cases such as when M is a submanifold of $\mathbb{R}^n$, $\dot{\gamma}(t)$ means exactly what you think it should; more generally, it is the equivalence class $[\gamma_t]$ represented by the reparametrized path $\gamma_t(s) := \gamma(t+s)$ that passes through $\gamma(t)$ at $s = 0$. Given $X \in \mathfrak{X}(M)$, a path $\gamma : (a,b) \to M$ is called a **flow line** or **orbit** of X if it satisfies

$$\dot{\gamma}(t) = X(\gamma(t)).$$

The following fundamental result translates most of the basic existence/uniqueness theory for ordinary differential equations into the language of differential geometry.

THEOREM 5.1. *For any smooth vector field $X \in \mathfrak{X}(M)$ on a manifold M, there exists a unique open subset $\mathcal{O} \subset \mathbb{R} \times M$ and smooth map*

$$\mathcal{O} \to M : (t, p) \mapsto \varphi_X^t(p),$$

*called the **flow** (Fluss) of X, such that for every $p \in M$, the set*

$$\ell_p := \{ t \in \mathbb{R} \mid (t, p) \in \mathcal{O} \} \subset \mathbb{R}$$

is an open interval containing 0, and

$$\gamma_p : \ell_p \to M : t \mapsto \varphi_X^t(p)$$

is the maximal solution to the initial value problem

$$\dot{\gamma}(t) = X(\gamma(t)), \qquad \gamma(0) = p.$$

Moreover, if X has compact support, then $\mathcal{O} = \mathbb{R} \times M$.

PROOF. For the most part, this result is proved by choosing local coordinates so as to rewrite the initial value problem in $\mathbb{R}^n$ and then applying standard results from the theory of ODEs. We will merely add a few observations in order to see how this works. First, given $p_0 \in M$, choose a smooth chart $(\mathcal{U}, x)$ with $p_0 \in \mathcal{U}$, which gives rise to a smooth chart $(T\mathcal{U}, Tx)$ on TM. The smoothness of X means that $p \mapsto Tx(X(p)) = (x(p), d_p x(X(p)))$ is a smooth function $\mathcal{U} \to \mathbb{R}^{2n}$, thus in particular, so is the function

$$\Phi : \mathcal{U} \to \mathbb{R}^n : p \mapsto d_p x(X(p)).$$

A path $\gamma : (-\epsilon, \epsilon) \to \mathcal{U}$ with $\gamma(0) = p_0$ will now satisfy $\dot{\gamma}(t) = X(\gamma(t))$ if and only if

$$(x \circ \gamma)'(t) = d_{\gamma(t)} x(\dot{\gamma}(t)) = d_{\gamma(t)} x(X(\gamma(t))),$$

meaning that $\alpha := x \circ \gamma : (-\epsilon, \epsilon) \to x(\mathcal{U}) \subset \mathbb{R}^n$ must be a solution to the initial value problem

$$\dot{\alpha}(t) = F(\alpha(t)), \qquad \alpha(0) = x(p_0), \tag{5.1}$$

where we define $F : x(\mathcal{U}) \to \mathbb{R}^n$ by

$$F(q) := d_{x^{-1}(q)} x(X(x^{-1}(q))) = \Phi \circ x^{-1}(q).$$

This last expression shows that F is a smooth function, so in particular it is Lipschitz, and the Picard-Lindelöf theorem therefore applies, telling us that a solution $\alpha : (-\epsilon, \epsilon) \to x(\mathcal{U})$ to (5.1) exists for some $\epsilon > 0$ and is unique. Since F is smooth, this solution also depends smoothly on the initial point $x(p_0)$. Replacing α with $\gamma = x^{-1} \circ \alpha : (-\epsilon, \epsilon) \to \mathcal{U}$, we similarly obtain existence and uniqueness of a solution to $\dot{\gamma}(t) = X(\gamma(t))$ with $\gamma(0) = p_0$, along with smooth dependence on the point p_0. This uniquely defines the flow map $(t, p) \mapsto \varphi_X^t(p)$ for all (t, p) in some neighborhood of $\{0\} \times M \subset \mathbb{R} \times M$.

It remains to establish that the flow map has a unique extension to a maximal domain which is an open subset $\mathcal{O} \subset \mathbb{R} \times M$, and is all of $\mathbb{R} \times M$ if X has compact support. This follows via the same tricks that are used to prove the corresponding statement in $\mathbb{R}^n$, e.g. whenever a flow line $\gamma : [0, T] \to M$ with $\gamma(0) = p_0$ exists, one can find a finite partition $0 = t_0 < t_1 < \ldots < t_{N-1} < t_N = T$ such that the subintervals $[t_{j-1}, t_j]$ are each sufficiently small for $\gamma([t_{j-1}, t_j])$ to lie within the domain of a single chart. One can then make use of the formula

$$\gamma(T) = \varphi_X^T(p_0) = \varphi_X^{t_N - t_{N-1}} \circ \ldots \circ \varphi_X^{t_2 - t_1} \circ \varphi_X^{t_1}(p_0),$$

in which each map in the composition is already known to be smooth and defined on an open neighborhood of the relevant point as long as the increments $t_j - t_{j-1}$ are small enough. This establishes that $\mathcal{O} \subset \mathbb{R} \times M$ is open and $(t, p) \mapsto \varphi_X^t(p)$ is smooth. Finally, if the support $K \subset M$ of X is a compact subset, then clearly every flow line through a point $p_0 \in M \backslash K$ is constant, so that

$(t, p_0) \in \mathcal{O}$ for all $t \in \mathbb{R}$. For the same reason, uniqueness of solutions implies that a flow line with initial value at a point $p_0 \in K$ can never escape from K; if it did, then it would become constant outside of K, and must therefore have *always* been a constant path outside of K. We claim now that for every $p_0 \in K$, the maximal solution to $\dot{\gamma}(t) = X(\gamma(t))$ with $\gamma(0) = p_0$ is defined for all $t \in \mathbb{R}$. If not, then suppose $\gamma : (a, b) \to M$ is the maximal solution and either $a > -\infty$ or $b < \infty$; for concreteness we will assume the latter since there is no substantial difference between the two cases. Then (a, b) contains a sequence t_j with $t_j \to b$, and after restricting to a subsequence, the compactness of K implies that we can assume $\gamma(t_j)$ converges to some point $p_1 \in K$. But solutions to the initial value problem starting at points near p_1 also exist and are unique on some sufficiently small interval, so for j large enough, $\gamma(t_j)$ must eventually lie on one of these solutions. The only way to have $\gamma(t_j) \to p_1$ is then if γ eventually matches (up to parametrization) the unique flow line through p_1, in which case it must reach that point at time $t = b$ and can be continued past it; this contradicts the assumption that γ could not be extended beyond the interval (a, b). □

We say that $X \in \mathfrak{X}(M)$ admits a **global flow** if the domain $\mathcal{O} \subset \mathbb{R} \times M$ of the flow map $(t, p) \mapsto \varphi_X^t(p)$ is $\mathbb{R} \times M$. This can sometimes be true even if X does not have compact support, e.g. it is easy to show that every C^0-bounded smooth vector field on $\mathbb{R}^n$ has a global flow. (There are also easy counterexamples if X is allowed to be unbounded, such as $X(x) := x^2$ on $\mathbb{R}$.) In the general case, φ_X^t defines for each $t \in \mathbb{R}$ a smooth map $\mathcal{O}_X^t \to M$ on the open set

$$\mathcal{O}_X^t := \left\{ p \in M \;\middle|\; (t, p) \in \mathcal{O} \right\},$$

and in fact, φ_X^t is a diffeomorphism from $\mathcal{O}_X^t$ to $\mathcal{O}_X^{-t}$, with inverse

$$(\varphi_X^t)^{-1} = \varphi_X^{-t}.$$

In particular, if the flow is global, then $\mathcal{O}_X^t = M$ for each $t \in \mathbb{R}$, and φ_X^t is therefore a diffeomorphism from M to itself. It is also possible however to have $\mathcal{O}_X^t = \varnothing$ for $t \neq 0$, though this cannot happen when t is close to 0. Indeed, it follows directly from the definition that

$$\mathcal{O}_X^s \supset \mathcal{O}_X^t \qquad \text{whenever} \qquad 0 \leqslant s \leqslant t \qquad \text{or} \qquad t \leqslant s \leqslant 0,$$

and short-time existence of solutions also implies

$$\mathcal{O}_X^0 = \bigcup_{t>0} \mathcal{O}_X^t = \bigcup_{t<0} \mathcal{O}_X^t = M.$$

The most important properties of the flow are perhaps

$$\varphi_X^0 = \mathrm{Id}, \qquad \text{and} \qquad \varphi_X^{s+t} = \varphi_X^s \circ \varphi_X^t \quad \text{on} \quad \mathcal{O}_X^s \cap \mathcal{O}_X^t \cap \mathcal{O}_X^{s+t} \text{ for every } s, t \in \mathbb{R},$$

which follow from the uniqueness of solutions to the initial value problem. Whenever the flow is global, this means that the map $t \mapsto \varphi_X^t$ defines a group homomorphism from $\mathbb{R}$ to the group $\mathrm{Diff}(M)$ of diffeomorphisms $M \to M$. This is, in practice, the single easiest way to produce a diffeomorphism on a manifold: one need not write it down explicitly, but can instead often write down an appropriate vector field more-or-less explicitly and deduce the existence of a suitable diffeomorphism via its flow. The following exercise is a demonstration of this technique:

Exercise 5.2. A manifold M is called **connected** (*zusammenhängend*)[24] if for every pair of points $p, q \in M$, there exists a continuous path $\gamma : [0, 1] \to M$ from $\gamma(0) = p$ to $\gamma(1) = q$. Show that under this assumption, there exists a diffeomorphism $\varphi : M \to M$ that is the identity map outside of a compact subset and satisfies $\varphi(p) = q$.

[24] If you know some topology, you may notice that what we are defining here is actually the notion of a **path-connected** space, and connectedness (without mentioning paths) usually means something else. However, every manifold is *locally* path-connected, so a general theorem from point-set topology (see [Wen18, Theorem 7.17]) implies that connectedness and path-connectedness on a manifold are equivalent conditions.

Hint: You should first convince yourself that the path $\gamma : [0,1] \to M$ can be assumed to be a smooth embedding without loss of generality. (This is obvious if γ happens to lie in the domain of a chart $(\mathcal{U}, x)$ such that $x(\mathcal{U}) \subset \mathbb{R}^n$ is convex, and notice that $\gamma([0,1]) \subset M$ can always be covered by finitely many such charts.) Then choose a vector field that has a flow line containing this path.

REMARK 5.3. If the vector field X is not smooth but is of class C^k for some $k \in \mathbb{N}$, then the proof of Theorem 5.1 above can be adapted to produce a flow map $(t,p) \mapsto \varphi_X^t(p)$ that is also of class C^k. As you may recall from your analysis courses, all bets are off if X is continuous but not C^1: in this case local solutions exist but may not be unique, so the flow cannot be defined.

5.2. Pullbacks and pushforwards. A diffeomorphism

$$\psi : M \to N$$

between two manifolds can be viewed as a way of "translating" all geometric data from M into equivalent geometric data on N or vice versa. The exact mechanism for the translation depends on the kind of data we are talking about: for points $p \in M$, the translation in N is simply $\psi(p) \in N$. For a function $f \in C^\infty(M)$, the equivalent data on N is a function

$$\psi_* f \in C^\infty(N)$$

that has the same value at the equivalent point $\psi(p)$ that f has at the original point p, thus

$$\psi_* f \circ \psi = f, \qquad \text{or equivalently} \qquad \psi_* f = f \circ \psi^{-1}.$$

We call $\psi_* f$ the **pushforward** of f via the diffeomorphism ψ. This process is invertible: one can associate to any $f \in C^\infty(N)$ a **pullback**

$$\psi^* f \in C^\infty(M)$$

via ψ, which takes the same value at p that f takes at $\psi(p)$; the definition is thus

$$\psi^* f = f \circ \psi.$$

To do the same trick with tangent vectors, we need to recall that the tangent map of a diffeomorphism $\psi : M \to N$ is also a diffeomorphism $T\psi : TM \to TN$, one which sends T_pM isomorphically to $T_{\psi(p)}N$ for each $p \in M$. This gives the natural way of "translating" tangent vectors between M and N, so for each $X \in TM$ and $Y \in TN$, we denote

$$\psi_* X := T\psi(X) \in TN, \qquad \psi^* Y := T\psi^{-1}(Y) \in TM.$$

The pushforward of a vector field $X \in \mathfrak{X}(M)$ should then be a vector field

$$\psi_* X \in \mathfrak{X}(N)$$

whose value at $\psi(p)$ for each $p \in M$ is the corresponding translation of the tangent vector $X(p)$, namely $\psi_*(X(p))$. This gives

$$(\psi_* X) \circ \psi = T\psi \circ X, \qquad \text{or equivalently} \qquad \psi_* X = T\psi \circ X \circ \psi^{-1}.$$

The pullback of a vector field $Y \in \mathfrak{X}(N)$ is obtained by inverting this procedure, thus

$$\psi^* Y := T\psi^{-1} \circ Y \circ \psi \in \mathfrak{X}(M).$$

PROPOSITION 5.4. *Suppose $\psi : M \to N$ is a diffeomorphism, $X \in \mathfrak{X}(N)$ is a vector field, and $t \in \mathbb{R}$. Then a point $p \in M$ is in the domain of the flow $\varphi^t_{\psi^* X}$ if and only if $\psi(p)$ belongs to the domain of φ^t_X, and $\psi \circ \varphi^t_{\psi^* X} = \varphi^t_X \circ \psi$.*

Proof. The result follows from the observation that ψ provides a natural bijective correspondence between the flow lines of X on N and flow lines of ψ^*X on M. Indeed, suppose $a < 0 < b$ and $\gamma : (a,b) \to N$ is a flow line of X, satisfying $\dot{\gamma}(t) = X(\gamma(t))$ and $\gamma(0) = q := \psi(p)$. Then $\alpha := \psi^{-1} \circ \gamma : (a,b) \to M$ satisfies $\alpha(0) = p$ and

$$\dot{\alpha}(t) = T\psi^{-1}\big(\dot{\gamma}(t)\big) = T\psi^{-1}\big(X(\gamma(t))\big) = T\psi^{-1} \circ X \circ \psi(\alpha(t)) = (\psi^*X)(\alpha(t)).$$

Conversely, the same computation implies that if α is a flow line of ψ^*X, then $\gamma := \psi \circ \alpha$ is a flow line of X. $\square$

Exercise 5.5. For two diffeomorphisms $\psi : M \to N$ and $\varphi : N \to Q$, prove the following relations:

(a) $(\varphi \circ \psi)_* f = \varphi_*(\psi_* f) \in C^\infty(Q)$ for $f \in C^\infty(M)$.
(b) $(\varphi \circ \psi)^* g = \psi^*(\varphi^* g) \in C^\infty(M)$ for $g \in C^\infty(Q)$.
(c) $(\varphi \circ \psi)_* X = \varphi_*(\psi_* X) \in \mathfrak{X}(Q)$ for $X \in \mathfrak{X}(M)$.
(d) $(\varphi \circ \psi)^* Y = \psi^*(\varphi^* Y) \in \mathfrak{X}(M)$ for $Y \in \mathfrak{X}(Q)$.

We will see later that when $\psi : M \to N$ is a diffeomorphism, pullbacks and pushforwards can be defined for any meaningful geometric data one might want to consider on M or N. A special case that arises quite often is where $M = N$ and $\psi : M \to M$ is defined by the flow of a vector field; we will see an example of this in the next lecture when we discuss the Lie derivative of a vector field. It will also be important to know that for certain types (but not all types) of data, either the pushforward *or* the pullback (but not both) can be defined via arbitrary smooth maps $\psi : M \to N$, not only for diffeomorphisms. One example of this is already apparent: for $f \in C^\infty(N)$, the pullback

$$\psi^* f := f \circ \psi \in C^\infty(M)$$

makes sense for any smooth map $\psi : M \to N$, so M and N need not be diffeomorphic. One cannot similarly define pushforwards of functions in this context, since ψ^{-1} might not be defined. We will see many more examples of this phenomenon when we discuss tensors and differential forms.

5.3. Derivations. For real-valued functions $f : M \to \mathbb{R}$, there is no natural notion of "partial derivatives" of f, unless M happens to be an open subset of $\mathbb{R}^n$. It is still natural however to talk about the **directional derivative** (*Richtungsableitung*) of f at a point $p \in M$ with respect to a tangent vector $X \in T_pM$: this is known as the **Lie derivative** (*Lie-Ableitung*) $\mathcal{L}_X f \in \mathbb{R}$ of f with respect to X, and can be evaluated using the differential df, i.e.

$$\mathcal{L}_X f := df(X).$$

If X is not just a tangent vector at a single point but a smooth vector field, then the Lie derivative defines another smooth function $M \to \mathbb{R}$, written

$$(\mathcal{L}_X f)(p) = df(X(p)).$$

The differential operator $\mathcal{L}_X$ associated to any $X \in \mathfrak{X}(M)$ thus defines a map

$$\mathcal{L}_X : C^\infty(M) \to C^\infty(M),$$

and one can check using the usual rules of differentiation that this map is linear:

$$\mathcal{L}_X(f+g) = \mathcal{L}_X f + \mathcal{L}_X g, \qquad \mathcal{L}_X(cf) = c\mathcal{L}_X f, \qquad \text{for all } f, g \in C^\infty(M),\ c \in \mathbb{R}.$$

Moreover, the product rule for differentiation translates into the following so-called **Leibniz rule**:

$$\mathcal{L}_X(fg) = (\mathcal{L}_X f)g + f\mathcal{L}_X g.$$

This formula motivates a short digression on algebras and Lie algebras.

DEFINITION 5.6. An **algebra** is a vector space $\mathcal{A}$ that is endowed with the additional structure of a bilinear multiplication operation

$$\mathcal{A} \times \mathcal{A} \to \mathcal{A} : (x, y) \mapsto xy$$

that is also associative, i.e. $(xy)z = x(yz)$ for all $x, y, z \in \mathcal{A}$.[25] A **derivation** on $\mathcal{A}$ is a linear map $L : \mathcal{A} \to \mathcal{A}$ that satisfies the Leibniz rule

$$L(xy) = (Lx)y + x(Ly) \qquad \text{for all } x, y \in \mathcal{A}.$$

An algebra endowed with a derivation is called a **differential algebra** (*Differentialalgebra*).

DEFINITION 5.7. A **Lie algebra** (*Lie-Algebra*) is a vector space V that is endowed with the additional structure of a bilinear operation

$$[\cdot, \cdot] : V \times V \to V,$$

its so-called **Lie bracket** (*Lie-Klammer*), which satisfies:

- **antisymmetry**: $[u, v] = -[v, u]$ for all $u, v \in V$;
- the **Jacobi identity**: $[u, [v, w]] + [v, [w, u]] + [w, [u, v]] = 0$ for all $u, v, w \in V$.

EXERCISE 5.8. Show that on any algebra $\mathcal{A}$, the space $\mathcal{D}$ of all derivations on $\mathcal{A}$ can be made into a Lie algebra by defining the bracket

$$[L_1, L_2] := L_1 \circ L_2 - L_2 \circ L_1.$$

In this course, the most important example of an algebra is the space of smooth real-valued functions $C^\infty(M)$ on a manifold M, in which multiplication is defined pointwise by $(fg)(p) := f(p)g(p)$. The previous remarks show that for any smooth vector field $X \in \mathfrak{X}(M)$, the associated Lie derivative operator $\mathcal{L}_X$ defines a derivation on $C^\infty(M)$. A somewhat less obvious class of examples comes from the observation in Exercise 5.8 that the **commutator bracket** of any two derivations is also a derivation, so in particular, any pair of vector fields $X, Y \in \mathfrak{X}(M)$ gives rise to a derivation on $C^\infty(M)$ defined by

$$[\mathcal{L}_X, \mathcal{L}_Y]f = \mathcal{L}_X\mathcal{L}_Y f - \mathcal{L}_Y\mathcal{L}_X f.$$

One says that the vector fields X and Y **commute** (*kommutieren*) whenever this bracket vanishes. This will turn out to be an important condition, but its meaning will take some effort to unpack. We first need to make the surprising and useful observation that the examples we have seen so far of derivations on $C^\infty(M)$ are the *only* examples that exist:

THEOREM 5.9. *Every derivation $L : C^\infty(M) \to C^\infty(M)$ is of the form $L = \mathcal{L}_X$ for some (unique) smooth vector field $X \in \mathfrak{X}(M)$.*

PROOF. The uniqueness of X is clear, since different vector fields define different derivations. The proof of existence follows from a series of claims.

***Claim 1**: If $f : M \to \mathbb{R}$ is a constant function, then $Lf = 0$ for every derivation L on $C^\infty(M)$.*

Indeed, if f is constant, then multiplication of an arbitrary function $g \in C^\infty(M)$ by f is the same as scalar multiplication, so linearity implies $L(fg) = f\,Lg$, and combining this with the Leibniz rule gives $(Lf)g = 0$. Plugging in the function $g \equiv 1$, we conclude $Lf \equiv 0$.

[25] If you're into algebra, you may notice that the definition of an algebra is quite similar to that of a ring. The difference is that while a ring is also an abelian group with respect to its "+" operation and has a distributive product operation, it does not generally come with any notion of scalar multiplication and is thus not a vector space. One can however define the notion of an algebra more generally, so that it is a module over a commutative ring R instead of a vector space. The case where R is a field then agrees with the definition we've given, but one can also speak of an algebra over $\mathbb{Z}$, which is the same thing as a ring since modules over $\mathbb{Z}$ are the same thing as abelian groups.

***Claim 2**: The stated result is true in the special case where M is a convex open subset of Euclidean space, $\Omega \subset \mathbb{R}^n$.*

This is the heart of the proof, and it depends on an important fact in first-year analysis that follows from the fundamental theorem of calculus. Assume $\Omega \subset \mathbb{R}^n$ is open and convex, and fix a point $x_0 = (x_0^1, \dots, x_0^n) \in \Omega$. For any other point $x = (x^1, \dots, x^n) \in \Omega$, the convexity of Ω implies that it contains the line segment between x_0 and x, so using the fundamental theorem of calculus and the chain rule, we find that any smooth function $f : \Omega \to \mathbb{R}$ satisfies

$$
\begin{aligned}
f(x) &= f(x_0) + \int_0^1 \frac{d}{d\tau} f(x_0 + \tau(x - x_0))\, d\tau = f(x_0) + \int_0^1 Df(x_0 + \tau(x - x_0))(x - x_0)\, d\tau \\
&= f(x_0) + \sum_{j=1}^n \left(\int_0^1 \partial_j f(x_0 + \tau(x - x_0))\, d\tau \right) (x^j - x_0^j) =: f(x_0) + \sum_{j=1}^n h_j(x)(x^j - x_0^j),
\end{aligned} \tag{5.2}
$$

where we've defined smooth functions $h_j : \Omega \to \mathbb{R}$ by $h_j(x) := \int_0^1 \partial_j f(x_0 + \tau(x - x_0))\, d\tau$. To make use of this formula, we can regard each of the coordinates $x^1, \dots, x^n$ as smooth real-valued functions on Ω and associate to these the smooth functions

$$X^j := L(x^j) \in C^\infty(\Omega), \qquad j = 1, \dots, n.$$

Linearity and the Leibniz rule, together with Claim 1, now produce from (5.2) the formula $Lf(x) = \sum_{j=1}^n \left[Lh_j(x) \cdot (x^j - x_0^j) + h_j(x) X^j(x) \right]$, so in particular,

$$Lf(x_0) = \sum_{j=1}^n h_j(x_0) X^j(x_0) = \sum_{j=1}^n X^j(x_0) \partial_j f(x_0).$$

The definition of the functions $X^j \in C^\infty(\Omega)$ did not depend on the choice of point $x_0 \in \Omega$, thus this formula is valid for every such point, giving an equality of functions

$$Lf = \sum_{j=1}^n X^j \partial_j f = \mathcal{L}_X f \qquad \text{on } \Omega,$$

where we define the smooth vector field $X \in \mathfrak{X}(\Omega)$ by $X(x) = (X^1(x), \dots, X^n(x)) \in \mathbb{R}^n = T_x\Omega$.

***Claim 3**: If the theorem holds for a particular manifold M, then it also holds for every manifold that is diffeomorphic to M.*

Assume $\psi : N \to M$ is a diffeomorphism between two manifolds, and the theorem is already known to hold for M. Any derivation L on $C^\infty(N)$ then determines a "pushforward" derivation $\psi_* L$ on $C^\infty(M)$ via the formula

$$(\psi_* L) f := L(f \circ \psi) \circ \psi^{-1}. \tag{5.3}$$

By assumption, the latter is $\mathcal{L}_X$ for some vector field $X \in \mathfrak{X}(M)$, and it is reasonable to guess that L will therefore correspond to the pullback vector field $\psi^* X \in \mathfrak{X}(N)$ as defined in §5.2. Let's check this: $\psi^* X$ is defined by

$$\psi^* X(p) = T\psi^{-1}(X(\psi(p))).$$

For $g \in C^\infty(N)$ and $p \in N$, we define $f := g \circ \psi^{-1} \in C^\infty(M)$ and use (5.3) to write

$$
\begin{aligned}
(Lg)(p) &= L(f \circ \psi)(p) = [(\psi_* L) f](\psi(p)) = (\mathcal{L}_X f)(\psi(p)) = df(X(\psi(p))) \\
&= d(g \circ \psi^{-1})(X(\psi(p))) = dg \circ T\psi^{-1}(X(\psi(p))) = dg(\psi^* X(p)) = \mathcal{L}_{\psi^* X} g(p),
\end{aligned}
$$

so the guess is correct!

For the remaining claims, assume M is a fixed manifold and $L : C^\infty(M) \to C^\infty(M)$ is a derivation.

***Claim 4**: If $f \in C^\infty(M)$ vanishes on a neighborhood of some point $p \in M$, then $Lf(p) = 0$.*

To see this, suppose $\mathcal{U} \subset M$ is a neighborhood of p on which $f \in C^\infty(M)$ vanishes, and choose any $g \in C^\infty(M)$ so that $g(p) = 1$ but g has compact support in $\mathcal{U}$.[26] Then $fg = 0$, thus $0 = (Lf)g + f(Lg)$, and evaluating the right hand side at p gives $0 = Lf(p) \cdot g(p) = Lf(p)$.

In light of linearity, a corollary of Claim 4 is that for any $f \in C^\infty(M)$, the value of $Lf(p)$ at any given point $p \in M$ depends only on the values of f on an arbitrarily small neighborhood of p.

***Claim 5**: For any open subset $\mathcal{U} \subset M$, L determines a unique derivation $L_{\mathcal{U}} : C^\infty(\mathcal{U}) \to C^\infty(\mathcal{U})$ such that for every $f \in C^\infty(M)$, $L_{\mathcal{U}}(f|_{\mathcal{U}}) = (Lf)|_{\mathcal{U}}$.*

This follows from the observation at the end of Claim 4 that $Lf(p)$ depends on f only in a neighborhood of p. Indeed, for any $f \in C^\infty(\mathcal{U})$, there is a unique function $L_{\mathcal{U}} f \in C^\infty(\mathcal{U})$ characterized by the property that for each $p \in \mathcal{U}$ and $f_p \in C^\infty(M)$ with $f_p \equiv f$ near p, $L_p f \equiv L_{\mathcal{U}} f$ near p. It is straightforward to verify that $L_{\mathcal{U}}$ defined in this way is a derivation.

Conclusion: Choose an open cover $M = \bigcup_{\alpha \in I} \mathcal{U}_\alpha$ such that for every α, there is a chart $(\mathcal{U}_\alpha, x_\alpha)$ whose image $x(\mathcal{U}_\alpha) \subset \mathbb{R}^n$ is convex. Claims 2 and 3 imply that the theorem holds for each of the open subsets $\mathcal{U}_\alpha \subset M$, thus for the derivation L_α determined on $C^\infty(\mathcal{U}_\alpha)$ by Claim 5, we have $L_\alpha = \mathcal{L}_{X_\alpha}$ for some vector field $X_\alpha \in \mathfrak{X}(\mathcal{U}_\alpha)$. We claim that for every pair $\alpha, \beta \in I$, X_α and X_β match on $\mathcal{U}_\alpha \cap \mathcal{U}_\beta$. Indeed, if $X_\alpha(p) \neq X_\beta(p)$ for some point p, then we can find a function $f \in C^\infty(M)$ with compact support in $\mathcal{U}_\alpha \cap \mathcal{U}_\beta$ such that $\mathcal{L}_{X_\alpha} f(p) \neq \mathcal{L}_{X_\beta} f(p)$, which is a contradiction since $L_\alpha(f|_{\mathcal{U}_\alpha})$ and $L_\beta(f|_{\mathcal{U}_\beta})$ should both have the same restriction as Lf on $\mathcal{U}_\alpha \cap \mathcal{U}_\beta$. The claim now implies that the vector fields X_α can be patched together to form a smooth vector field $X \in \mathfrak{X}(M)$, and in light of Claim 4, the relation $Lf = \mathcal{L}_X f$ now follows on each $\mathcal{U}_\alpha$ from $L(f|_{\mathcal{U}_\alpha}) = \mathcal{L}_{X_\alpha}(f|_{\mathcal{U}_\alpha})$. □

REMARK 5.10. In light of Theorem 5.9, it is common in differential geometry to blur the distinction between smooth vector fields on M and derivations on $C^\infty(M)$, and many books even use exactly the same notation for both, thus writing

$$Xf := \mathcal{L}_X f \in C^\infty(M)$$

so as to view the vector field $X \in \mathfrak{X}(M)$ as a differential operator acting on the function $f \in C^\infty(M)$. I personally prefer not to do this, and will thus continue writing $\mathcal{L}_X$ to distinguish the derivation defined by a vector field $X \in \mathfrak{X}(M)$ from the vector field itself; the sole exception to this will be the coordinate vector fields discussed in the next subsection. Many authors would probably call this practice overly pedantic, and I cannot say with confidence that they are wrong.

EXERCISE 5.11. For a diffeomorphism $\psi : M \to N$, vector field $X \in \mathfrak{X}(M)$ and function $f \in C^\infty(M)$, prove $\mathcal{L}_{\psi_* X}(\psi_* f) = \psi_*(\mathcal{L}_X f) \in C^\infty(N)$.

5.4. Coordinate vector fields. Given a smooth chart $(\mathcal{U}, x)$ on a manifold M, the coordinate functions $x^1, \ldots, x^n : \mathcal{U} \to \mathbb{R}$ define a natural family of derivations on $C^\infty(\mathcal{U})$, namely the n partial derivative operators

$$\partial_j := \frac{\partial}{\partial x^j} : C^\infty(\mathcal{U}) \to C^\infty(\mathcal{U}), \qquad j = 1, \ldots, n,$$

which are defined by writing any function $f \in C^\infty(\mathcal{U})$ in its local coordinate representation $(x^1, \ldots, x^n) \mapsto f(x^1, \ldots, x^n)$ and differentiating the resulting function of n variables as one would in first-year analysis. The more precise way to say this is that for each $f \in C^\infty(\mathcal{U})$ and $p \in \mathcal{U}$, the function $\partial_j f \in C^\infty(\mathcal{U})$ is given by

$$(\partial_j f)(p) := \partial_j (f \circ x^{-1})(x(p)),$$

[26] Such a function can be constructed in local coordinates our of functions of the form $\mathbb{R}^n \to [0,1] : x \mapsto \beta(|x|^2)$, where $\beta : \mathbb{R} \to [0,1]$ is a smooth function with $\beta(t) = 0$ for all $t \geqslant \epsilon > 0$ and $\beta(0) = 1$. The construction of β is an easy exercise once you've seen examples like $h(t) := e^{-1/t^2}$, a smooth function on $(0, \infty)$ admitting a smooth extension to $\mathbb{R}$ that vanishes on $(-\infty, 0]$.

where the right-hand side is a perfectly ordinary partial derivative of a real-valued function of n real variables. The fact that the operators $\partial_1, \ldots, \partial_n$ define derivations on $C^\infty(\mathcal{U})$ follows immediately from the usual product rule. The corresponding vector fields in $\mathfrak{X}(\mathcal{U})$ are also easy to identify: they come from the standard basis $e_1, \ldots, e_n$ of $\mathbb{R}^n$ as transferred over to $\mathcal{U}$ by the chart, i.e. the derivation ∂_j corresponds to the vector field

$$v_j(p) := (d_p x)^{-1}(e_j), \qquad p \in \mathcal{U}.$$

Since this notation is bit clumsy, it has become conventional in differential geometry to use the notation

$$\partial_1, \ldots, \partial_n \quad \text{or equivalently} \quad \frac{\partial}{\partial x^1}, \ldots, \frac{\partial}{\partial x^n} \in \mathfrak{X}(\mathcal{U})$$

not just for the derivations but also for the corresponding vector fields on $\mathcal{U}$, and I will follow that convention in these notes, in spite of what I said in Remark 5.10 above. We call these the **coordinate vector fields** determined on $\mathcal{U}$ by the chart $(\mathcal{U}, x)$. Two issues are very important to understand:

(1) The vector fields $\frac{\partial}{\partial x^j}$ are *only* defined on $\mathcal{U} \subset M$; it does not make sense to write down formulas involving ∂_j everywhere on M unless $(\mathcal{U}, x)$ happens to be a *global* chart, meaning $\mathcal{U} = M$.
(2) For each individual $j \in \{1, \ldots, n\}$, the vector field $\frac{\partial}{\partial x^j}$ depends not only on the coordinate function $x^j : \mathcal{U} \to \mathbb{R}$ but on all n of the coordinates $x^1, \ldots, x^n$. Indeed, the vector $\frac{\partial}{\partial x^j}$ points in the unique direction where x^j increases but all the other coordinates are constant. The issue is easy to see in simple examples, e.g. using the standard polar coordinates (r, θ) and Cartesian coordinates (x, y) on suitable regions in $\mathbb{R}^2$, one can define both (r, θ) and (r, y) as smooth charts on the open right half-plane $\{x > 0\} \subset \mathbb{R}^2$. But the partial derivative operator $\frac{\partial}{\partial r}$ has different meanings in these two coordinate systems, because differentiating in a direction where r increases but θ is constant does not typically give the same result as differentiating in a direction where r increases but y is constant.

6. The Lie algebra of vector fields

We saw in the last lecture that there is a natural equivalence between the space of smooth vector fields $\mathfrak{X}(M)$ on a smooth manifold M and the space of all derivations $L : C^\infty(M) \to C^\infty(M)$ on the algebra of smooth functions. It was also observed in Exercise 5.8 that the latter has a natural Lie algebra structure defined via the commutator bracket

$$[L_1, L_2] := L_1 L_2 - L_2 L_1,$$

which is antisymmetric and satisfies the Jacobi identity (see Definition 5.7). Lie algebras are a large topic that we will discuss in more detail next semester; if you have not seen them at all before, then I would not expect you to have any intuition as to why a bilinear bracket satisfying antisymmetry and the Jacobi identity might be an interesting or useful object to study. But we will see a first example of the answer to that question in this lecture: the Lie algebra structure on the space of vector fields characterizes the commutativity (or lack thereof) of their respective flows. This will be easily the deepest result we have proved so far in this course, and it will serve as a foundation for several later results involving curvature and integrability.

6.1. Components and the summation convention. Recall that any smooth chart $(\mathcal{U}, x = (x^1, \ldots, x^n))$ on a manifold M defines an associated set of *coordinate vector fields* $\frac{\partial}{\partial x^1}, \ldots, \frac{\partial}{\partial x^n} \in$

$\mathfrak{X}(\mathcal{U})$. These form a basis of T_pM at each point $p \in \mathcal{U}$, so any $X \in \mathfrak{X}(M)$ restricted to $\mathcal{U} \subset M$ can now be written uniquely in the form

$$X = \sum_{i=1}^{n} X^i \partial_i = \sum_{i=1}^{n} X^i \frac{\partial}{\partial x^i} \tag{6.1}$$

for uniquely defined smooth functions $X^1, \ldots, X^n \in C^\infty(\mathcal{U})$, called the **components** of X with respect to the chart $(\mathcal{U}, x)$. This observation will be useful for computations, but it becomes more so if we can make the notation a bit less cumbersome. Einstein introduced a nice trick for this, which is known as the **Einstein summation convention**: the trick is to omit the summation symbol, but assume that whenever a matching pair of "upper" and "lower" indices appears, a summation of that index over all coordinates (in this case from 1 to n) is implied. Using this convention, (6.1) becomes

$$X = X^i \partial_i = X^i \frac{\partial}{\partial x^i},$$

where the convention is also to interpret the upper index in $\frac{\partial}{\partial x^i}$ as a lower index because it appears in the denominator. (I advise you not to search for any deeper meaning behind this—just take it as a definition for now, and you will see presently why it is useful.) The simplicity of this expression in comparison with (6.1) is perhaps not so dramatic, but the Einstein convention becomes especially useful in situations where multiple indices need to be summed over at the same time, which will happen a lot once we start talking about tensors next week.

Let us derive a coordinate transformation formula: suppose $(\widetilde{\mathcal{U}}, \widetilde{x})$ is a second chart with $\mathcal{U} \cap \widetilde{\mathcal{U}} \neq \varnothing$, and the components of X in these alternative coordinates over $\widetilde{\mathcal{U}}$ are denoted by $\widetilde{X}^i$, so $X = \widetilde{X}^i \frac{\partial}{\partial \widetilde{x}^i}$ on $\widetilde{\mathcal{U}}$. How do the components X^i and $\widetilde{X}^i$ relate to each other on the region $\mathcal{U} \cap \widetilde{\mathcal{U}}$ where their domains overlap?

To answer this, we start with the observation that for any $f \in C^\infty(\mathcal{U} \cap \widetilde{\mathcal{U}})$, the chain rule relates the partial derivatives of f with respect to the two different coordinate systems by

$$\frac{\partial f}{\partial x^i} = \frac{\partial f}{\partial \widetilde{x}^j} \frac{\partial \widetilde{x}^j}{\partial x^i}, \tag{6.2}$$

where the Einstein convention gives an implied summation $\sum_{j=1}^n$ on the right hand side. This formula is hopefully familiar to you from analysis, at least when applied to functions on open subsets of $\mathbb{R}^n$; in the present setting, the partial derivatives on both sides are interpreted as derivations applied to smooth functions on $\mathcal{U} \cap \widetilde{\mathcal{U}} \subset M$, but these have been defined in terms of ordinary partial derivatives of functions on $\mathbb{R}^n$. In that context, the left hand side is the ith component of the *gradient* ∇f of f in coordinates $(x^1, \ldots, x^n)$, interpreted as a row vector, while the right hand side is the ith component of the product of the row vector $\widetilde{\nabla} f$ (the gradient of f is coordinates $(\widetilde{x}^1, \ldots, \widetilde{x}^n)$ with the Jacobian matrix $\frac{\partial \widetilde{x}}{\partial x}$ of the transition map $(x^1, \ldots, x^n) \mapsto (\widetilde{x}^1(x^1, \ldots, x^n), \ldots, \widetilde{x}^n(x^1, \ldots, x^n))$. Equation (6.2) is thus equivalent to the relation

$$D(f \circ x^{-1})(x(p)) = D(f \circ \widetilde{x}^{-1})(\widetilde{x}(p)) \circ D(\widetilde{x} \circ x^{-1})(x(p)),$$

which follows directly from the chain rule. Now, the function f was not actually important in this discussion at all: what we are really interested in is a formula relating derivations, namely

$$\frac{\partial}{\partial x^i} = \frac{\partial \widetilde{x}^j}{\partial x^i} \frac{\partial}{\partial \widetilde{x}^j}, \tag{6.3}$$

which can now equally well be interpreted as a formula for the coordinate vector field $\frac{\partial}{\partial x^i}$ as a linear combination of the other set of coordinate vector fields $\frac{\partial}{\partial \widetilde{x}^j}$ where they overlap. This implies

$$X = X^i \frac{\partial}{\partial x^i} = X^i \frac{\partial \widetilde{x}^j}{\partial x^i} \frac{\partial}{\partial \widetilde{x}^j} = \widetilde{X}^j \frac{\partial}{\partial \widetilde{x}^j},$$

from which we derive (after interchanging the indices i and j just for good measure) the transformation formula

$$\widetilde{X}^i = \frac{\partial \widetilde{x}^i}{\partial x^j} X^j. \tag{6.4}$$

You may agree that if we'd had to write summation symbols in all of these expressions, we would be slightly more tired now. Notice that this formula has an easy interpretation in terms of matrix-vector multiplication: if we package the components together into $\mathbb{R}^n$-valued functions $\xi := (X^1, \ldots, X^n) : \mathcal{U} \to \mathbb{R}^n$ and $\widetilde{\xi} := (\widetilde{X}^1, \ldots, \widetilde{X}^n) : \widetilde{\mathcal{U}} \to \mathbb{R}^n$, then (6.4) relates these two functions to each other via multiplication with the Jacobian matrix $\frac{\partial \widetilde{x}}{\partial x}$:

$$\widetilde{\xi} = \frac{\partial \widetilde{x}}{\partial x} \xi.$$

The Einstein convention has nothing intrinsically to do with differential geometry—it is actually just linear algebra. Once you get used to it, you may begin to wish you had always been doing linear algebra this way.

We will use the Einstein convention consistently throughout the rest of this course, and only include explicitly written summation symbols in situations where their omission might cause confusion.

REMARK 6.1. Using the summation convention requires being very careful and consistent about the distinction between upper and lower indices: coordinates and components of vector fields are *always* written with upper indices, while partial derivative operators (and their associated coordinate vector fields) always carry lower indices. Forgetting these conventions can cause grave confusion and should be avoided at all costs. Unfortunately, not all differential geometry books written by mathematicians are completely consistent about this, though books by physicists are—Einstein was one of them, after all, so his mathematical innovations are taken as gospel.

6.2. The Lie bracket. The **Lie bracket** (*Lie-Klammer*) of two vector fields $X, Y \in \mathfrak{X}(M)$ on a manifold M is defined to be the unique vector field

$$[X, Y] \in \mathfrak{X}(M) \qquad \text{such that} \qquad \mathcal{L}_{[X,Y]} = \mathcal{L}_X \mathcal{L}_Y - \mathcal{L}_Y \mathcal{L}_X.$$

This definition makes sense as a consequence of Exercise 5.8 and Theorem 5.9. In particular, we say that X and Y **commute** if $[X, Y] \equiv 0$.

EXERCISE 6.2. Suppose $(\mathcal{U}, x)$ is a chart on M and we express two vector fields $X, Y \in \mathfrak{X}(M)$ over $\mathcal{U}$ in this chart as $X = X^i \partial_i$ and $Y = Y^i \partial_i$.

(a) Show that the components $[X, Y]^i$ of $[X, Y]$ with respect to the same chart are given by

$$[X, Y]^i = X^j \frac{\partial Y^i}{\partial x^j} - Y^j \frac{\partial X^i}{\partial x^j}. \tag{6.5}$$

(b) Use the coordinate transformation formulas (6.3) and (6.4) to give a direct computational proof (without using the result of part (a)) that the vector field defined on $\mathcal{U}$ via the right hand side of (6.5) depends only on $X, Y \in \mathfrak{X}(\mathcal{U})$ and not on the choice of chart $(\mathcal{U}, x)$. In other words, show that for any other chart $(\widetilde{\mathcal{U}}, \widetilde{x})$,

$$\left(X^j \frac{\partial Y^i}{\partial x^j} - Y^j \frac{\partial X^i}{\partial x^j} \right) \frac{\partial}{\partial x^i} = \left(\widetilde{X}^j \frac{\partial \widetilde{Y}^i}{\partial \widetilde{x}^j} - \widetilde{Y}^j \frac{\partial \widetilde{X}^i}{\partial \widetilde{x}^j} \right) \frac{\partial}{\partial \widetilde{x}^i} \quad \text{on} \quad \mathcal{U} \cap \widetilde{\mathcal{U}}.$$

Hint: The matrices with entries $\frac{\partial \tilde{x}^i}{\partial x^j}$ and $\frac{\partial x^i}{\partial \tilde{x}^j}$ are Jacobi matrices for transformations that are inverse to each other, thus they satisfy

$$\frac{\partial \tilde{x}^i}{\partial x^j}\frac{\partial x^j}{\partial \tilde{x}^k} = \delta^i_k := \begin{cases} 1 & \text{if } i = k, \\ 0 & \text{if } i \neq k. \end{cases}$$

REMARK 6.3. Physicists like being able to do explicit computations, so they tend to emphasize coordinate-based formulas in this subject much more than mathematicians do. For example, some physics books take the formula (6.5) as a *definition* of the Lie bracket $[X, Y]$, without first talking about commutators of derivations. The price for doing this is that one must prove that switching to a different local coordinate system would not change the definition, i.e. one must do Exercise 6.2(b). The exercise is tedious, but I recommend doing it exactly once in your life, as it may give you some useful insight into the way that physicists do mathematics, and in any case, it is never bad to get better at explicit computations. As a cautionary tale, I also recommend convincing yourself that the simpler formula

$$X^j \frac{\partial Y^i}{\partial x^j}\frac{\partial}{\partial x^i} = \tilde{X}^j \frac{\partial \tilde{Y}^i}{\partial \tilde{x}^j}\frac{\partial}{\partial \tilde{x}^i} \quad \text{on} \quad \mathcal{U} \cap \tilde{\mathcal{U}}$$

is *false* in general, thus one cannot define a vector field $Z = Z^i\partial_i$ by $Z^i := X^j\partial_j Y^i$ and expect the definition to be independent of the choice of coordinates.

EXERCISE 6.4. For $X, Y \in \mathfrak{X}(M)$ and $f \in C^\infty(M)$, give two proofs of the formulas

$$[fX, Y] = f[X, Y] - (\mathcal{L}_Y f)X, \qquad [X, fY] = f[X, Y] + (\mathcal{L}_X f)Y,$$

using different methods:

(a) Directly from the definition of the Lie bracket via Theorem 5.9;
(b) Using the coordinate formula (6.5).

EXERCISE 6.5. For a diffeomorphism $\psi : M \to N$ and two vector fields $X, Y \in \mathfrak{X}(M)$, prove $\psi_*[X, Y] = [\psi_* X, \psi_* Y] \in \mathfrak{X}(N)$.

EXAMPLE 6.6. The coordinate vector fields $\partial_1, \ldots, \partial_n$ defined from any chart on an open subset all commute with each other. One can deduce this either from the fact that $\partial_i\partial_j f = \partial_j\partial_i f$ for all smooth functions f,[27] or as a trivial application of the formula in Exercise 6.2.

My goal for the rest of this lecture is to explain not just what the Lie bracket of two vector fields *is*, but what it *means*. The discussion starts with the following observation related to Example 6.6 above. Consider the manifold $M = \mathbb{R}^n$ with the standard Cartesian coordinates $x^1, \ldots, x^n$ regarded as a global chart on M; this chart is actually just the identity map $\mathbb{R}^n \to \mathbb{R}^n$. The resulting coordinate vector fields $\partial_1, \ldots, \partial_n$ produce the standard basis of the tangent space $T_p\mathbb{R}^n = \mathbb{R}^n$ at every point $p \in \mathbb{R}^n$. It is easy to write down the flow of ∂_j for each $j = 1, \ldots, n$: it is

$$\varphi^t_{\partial_j}(x^1, \ldots, x^n) = (x^1, \ldots, x^{j-1}, x^j + t, x^{j+1}, \ldots, x^n).$$

We see from this that for any two $i, j \in \{1, \ldots, n\}$ and $s, t \in \mathbb{R}$, the corresponding flows commute:

$$\varphi^s_{\partial_i} \circ \varphi^t_{\partial_j} = \varphi^t_{\partial_j} \circ \varphi^s_{\partial_i}.$$

This is a generalization of the basic observation that if you start from some point (x, y) in the plane $\mathbb{R}^2$, move a distance s to the right and then a distance t upward, you'll end up at the same point as if you had made those two moves in the reverse order, namely $(x+s, y+t)$. In other words,

[27]And since this is not an analysis course, there is no need to worry about the fact that $\partial_i\partial_j f = \partial_j\partial_i f$ does not generally hold for functions whose second-order derivatives exist but are discontinuous. With very few exceptions, all functions that we choose to worry about in the remainder of this course will be of class C^∞.

the two paths, each consisting of two straight line segments, combine to form a closed rectangle. This observation is not as trivial as it may seem: in particular, it becomes false in general if you replace ∂_i and ∂_j by different vector fields, e.g. in the example of $\mathbb{R}^2$, one could replace the "horizontal" coordinate vector field ∂_1 with one that still points in the x-direction but flows at different speeds along the lower and upper segments of the rectangle, in which case the rectangle fails to close up. There is no reason in general why the flows of two vector fields should always commute. They do commute in the case of coordinate vector fields on $\mathbb{R}^n$, and it follows easily that flows of coordinate vector fields determined by a chart $(\mathcal{U}, x)$ on a manifold M will generally commute as long as one keeps s and t close enough to 0 so that the flow lines do not escape from $\mathcal{U}$. But pairs of coordinate vector fields are special, and one symptom of this is the fact that their Lie brackets vanish. We will show in §6.4 that this is a general phenomenon: in particular, for any two vector fields $X, Y \in \mathfrak{X}(M)$ whose flows exist globally, one has $\varphi_X^s \circ \varphi_Y^t = \varphi_Y^t \circ \varphi_X^s$ for all $s, t \in \mathbb{R}$ if and only if $[X, Y] \equiv 0$.

6.3. The Lie derivative of a vector field. Before we can prove a result on commuting flows, we need a short digression to address the following question: What might it mean to differentiate a vector field $Y \in \mathfrak{X}(M)$ at a point $p \in M$ in the direction $X \in T_pM$? A naive attempt to define this would proceed as follows: choose any smooth path $\gamma : (-\epsilon, \epsilon) \to M$ with $\gamma(0) = p$ and $\dot{\gamma}(0) = X$, and set

$$\mathcal{L}_X Y(p) \overset{?}{:=} \left.\frac{d}{dt} Y(\gamma(t))\right|_{t=0} = \lim_{t \to 0} \frac{Y(\gamma(t)) - Y(p)}{t} \; ?$$

If Y were a real-valued function instead of a vector field, then we would be on solid ground with this definition, but for a vector field the right hand side does not make sense: outside of the uninteresting special case where γ is a constant path, $Y(\gamma(t)) \in T_{\gamma(t)}M$ and $Y(p) \in T_pM$ generally belong to different vector spaces, so there is no well-defined way of subtracting one from the other.

A solution to this conundrum arises if one allows X to be a vector *field* on M, rather than just a single tangent vector. In this case, the flow of X gives a natural choice of the path

$$\gamma(t) = \varphi_X^t(p),$$

which is defined for t in a sufficiently small interval $(-\epsilon, \epsilon)$ even if the flow does not globally exist. More importantly, the tangent map of the flow gives rise to natural isomorphisms,

$$T_p\varphi_X^t : T_pM \to T_{\varphi_X^t(p)}M = T_{\gamma(t)}M$$

for t close to 0, which gives us a way of identifying with each other the distinct tangent spaces in which $Y(p)$ and $Y(\gamma(t))$ live. Since the inverse of $T\varphi_X^t$ is $T\varphi_X^{-t}$, it now makes sense to define the **Lie derivative** (*Lie-Ableitung*) of $Y \in \mathfrak{X}(M)$ with respect to $X \in \mathfrak{X}(M)$ as the vector field

$$\mathcal{L}_X Y \in \mathfrak{X}(M), \qquad \mathcal{L}_X Y(p) := \left.\frac{d}{dt} T\varphi_X^{-t}\big(Y(\varphi_X^t(p))\big)\right|_{t=0} = \lim_{t \to 0} \frac{T\varphi_X^{-t}\big(Y(\varphi_X^t(p))\big) - Y(p)}{t}.$$

Recalling the definition of the *pullback* of a vector field in §5.2, we can abbreviate this formula as

$$\mathcal{L}_X Y = \left.\frac{d}{dt} (\varphi_X^t)^* Y\right|_{t=0}.$$

It turns out that $\mathcal{L}_X Y$ is just a new perspective on the Lie bracket:

PROPOSITION 6.7. *For any $X, Y \in \mathfrak{X}(M)$, $\mathcal{L}_X Y = [X, Y]$.*

PROOF. We need to show that for every $f \in C^\infty(M)$,

$$\mathcal{L}_{\mathcal{L}_X Y} f = \mathcal{L}_X \mathcal{L}_Y f - \mathcal{L}_Y \mathcal{L}_X f. \tag{6.6}$$

In the following, when writing expressions such as $\varphi_X^t(p)$, we always assume that t is close enough to 0 for this flow to be defined. With this understood, we claim that

$$f \circ \varphi_X^t = f + t g_t$$

for some smooth family of smooth real-valued functions g_t on M with $g_0 = \mathcal{L}_X f \in C^\infty(M)$.[28] This follows from the fundamental theorem of calculus: for $p \in M$ and $t \in \mathbb{R}$ close to 0, we write

$$\begin{aligned} f(\varphi_X^t(p)) - f(p) &= \int_0^1 \frac{d}{ds} f(\varphi_X^{st}(p))\, ds = \int_0^1 df\left(\partial_s \varphi_X^{st}(p)\right) ds \\ &= \int_0^1 df\left(tX(\varphi_X^{st}(p))\right) ds = t \int_0^1 df\left(X(\varphi_X^{st}(p))\right) ds, \end{aligned}$$

define $g_t(p)$ to be the integral on the right, and compute

$$g_0(p) = \int_0^1 df\left(X(\varphi_X^0(p))\right) ds = \int_0^1 df(X(p))\, ds = df(X(p)) = \mathcal{L}_X f(p),$$

proving the claim. Using this formula, we find

$$\begin{aligned} df\left([(\varphi_X^t)^* Y](p)\right) &= df\left(T\varphi_X^{-t}(Y(\varphi_X^t(p)))\right) = d(f \circ \varphi_X^{-t})(Y(\varphi_X^t(p))) \\ &= d(f - t g_t)\left(Y(\varphi_X^t(p))\right) = df\left(Y(\varphi_X^t(p))\right) - t\, dg_t\left(Y(\varphi_X^t(p))\right) \\ &= \mathcal{L}_Y f(\varphi_X^t(p)) - t\, \mathcal{L}_Y g_t(\varphi_X^t(p)). \end{aligned}$$

If we now differentiate this relation with respect to t and set $t = 0$, the left hand side becomes $df(\mathcal{L}_X Y(p)) = \mathcal{L}_{\mathcal{L}_X Y} f(p)$, while the right hand side becomes

$$d(\mathcal{L}_Y f)(X(p)) - \mathcal{L}_Y g_0(p) = \mathcal{L}_X \mathcal{L}_Y f(p) - \mathcal{L}_Y \mathcal{L}_X f(p),$$

proving (6.6). □

REMARK 6.8. The formula $\mathcal{L}_X Y = [X, Y]$ reveals that the Lie derivative of a vector field does not quite admit the interpretation we were hoping for: if $\mathcal{L}_X Y(p)$ were merely the directional derivative of $Y \in \mathfrak{X}(M)$ at p in the direction of $X \in T_p M$, then it should only depend on Y and the specific value $X(p)$, but as we see in (6.5), $[X, Y](p)$ also depends on the *first derivatives* of X at p in coordinates, not just on its value. We will see later that a straightforward directional derivative of anything more complicated than a real-valued function cannot typically be defined without making additional choices, e.g. the definition of $\mathcal{L}_X Y(p)$ requires extending $X(p)$ to a vector field that takes that value at p, and the resulting derivative depends on that choice. We will see a different and in some sense simpler way to define directional derivatives of vector fields when we study *connections* later in the semester, but a connection is also a choice that is not canonically defined in general.

6.4. Commuting flows. We can now discuss the relationship between the Lie bracket $[X, Y]$ and the question of whether the flows of X and Y commute. To understand the statement, recall from §5.1 that for each $X \in \mathfrak{X}(M)$ and $s \in \mathbb{R}$, the flow defines a diffeomorphism

$$\varphi_X^s : \mathcal{O}_X^s \to \mathcal{O}_X^{-s}$$

between two open subsets $\mathcal{O}_X^s, \mathcal{O}_X^{-s} \subset M$, which may in general be empty, but are guaranteed to be nonempty if s is close enough to 0; in fact, we have $\mathcal{O}_X^0 = \bigcup_{s>0} \mathcal{O}_X^s = \bigcup_{s<0} \mathcal{O}_X^s = M$.

[28]Saying that g_t is a "smooth family" of functions on M means literally that the function $(t, p) \mapsto g_t(p)$ for (t, p) in some open subset of $\mathbb{R} \times M$ is smooth. A slightly subtle point here is that we do not need the function $g_t : M \to M$ to be well-defined *everywhere* on M for some $t \neq 0$; for our purposes, it will suffice if $g_t(p)$ is defined for all (t, p) in some *neighborhood* of the set $\{0\} \times M$. If M is not compact, it may happen that the domain of $(t, p) \mapsto g_t(p)$ does not contain any set of the form $\{t\} \times M$ for $t \neq 0$, but is still an open neighborhood of $\{0\} \times M$.

For another vector field $Y \in \mathfrak{X}(M)$ and another $t \in \mathbb{R}$, the composition $\varphi_Y^t \circ \varphi_X^s$ is defined on $(\varphi_X^s)^{-1}(\mathcal{O}_Y^t) \subset M$, which is also open and could be empty, but is definitely not empty if both $|s|$ and $|t|$ are sufficiently small. The domain of $\varphi_X^s \circ \varphi_Y^t$ may be a different open subset of M, but is also guaranteed to overlap the domain of $\varphi_Y^t \circ \varphi_X^s$ if $|s|$ and $|t|$ are sufficiently small; in fact for every $p \in M$, there exists ϵ such that both $\varphi_X^s \circ \varphi_Y^t(p)$ and $\varphi_Y^t \circ \varphi_X^s(p)$ are defined whenever $|s|, |t| < \epsilon$.

THEOREM 6.9. *For two smooth vector fields $X, Y \in \mathfrak{X}(M)$ on a manifold M, the following conditions are equivalent:*

(i) $[X, Y] \equiv 0$*;*

(ii) Suppose $p \in M$ and $s, t \in \mathbb{R}$ are such that $\varphi_X^\sigma \circ \varphi_Y^\tau(p)$ is defined for all σ between 0 and s and all τ between 0 and t. Then $\varphi_Y^\tau \circ \varphi_X^\sigma(p)$ is also defined for all such σ and τ, and it equals $\varphi_X^\sigma \circ \varphi_Y^\tau(p)$. In particular, if X and Y both have global flows, then they define commuting diffeomorphisms

$$\varphi_X^s \circ \varphi_Y^t = \varphi_Y^t \circ \varphi_X^s \in \mathrm{Diff}(M)$$

for all $s, t \in \mathbb{R}$.

PROOF. We prove first that (ii) $\Rightarrow$ (i), so suppose X and Y are two vector fields whose flows commute in the sense described in the statement. For each $p \in M$, one can find a neighborhood $\mathcal{U} \subset \mathbb{R}^2$ of $(0,0)$ small enough so that the smooth map

$$\alpha : \mathcal{U} \to M : (s,t) \mapsto \varphi_X^s \circ \varphi_Y^t(p) = \varphi_Y^t \circ \varphi_X^s(p)$$

is well-defined via either of the compositions on the right hand side. This map satisfies $\partial_s \alpha(s,t) = X(\alpha(s,t))$ and $\partial_t \alpha(s,t) = Y(\alpha(s,t))$, where the proof of the first identity requires the first version of the composition, and the second requires the second. Given $f \in C^\infty(M)$, we now define $g := f \circ \alpha : \mathcal{U} \to \mathbb{R}$ and observe that

$$\mathcal{L}_X f(\alpha(s,t)) = \partial_s g(s,t) \qquad \text{and} \qquad \mathcal{L}_Y f(\alpha(s,t)) = \partial_t g(s,t),$$

and similarly,

$$\mathcal{L}_X \mathcal{L}_Y f(\alpha(s,t)) = \partial_s \partial_t g(s,t) = \partial_t \partial_s g(s,t) = \mathcal{L}_Y \mathcal{L}_X f(\alpha(s,t)).$$

This proves in particular that $(\mathcal{L}_X \mathcal{L}_Y - \mathcal{L}_Y \mathcal{L}_X) f(p) = 0$, hence $[X,Y](p) = 0$ for all $p \in M$.

To prove (i) $\Rightarrow$ (ii), assume $[X,Y] \equiv 0$, and fix $p \in M$ and $s, t \in \mathbb{R}$ satisfying the condition specified in (ii). Then for each σ in the interval between 0 and s, φ_X^σ defines a diffeomorphism

$$M \overset{\text{open}}{\supset} \mathcal{O}_X^\sigma \xrightarrow{\varphi_X^\sigma} \mathcal{O}_X^{-\sigma} \overset{\text{open}}{\subset} M$$

whose domain and target satisfy $\mathcal{O}_X^\sigma \supset \mathcal{O}_X^s$ and $\mathcal{O}_X^{-\sigma} \supset \mathcal{O}_X^{-s}$ respectively, and moreover, the flow line $\gamma(\tau) := \varphi_Y^\tau(p)$ exists and has image in $\mathcal{O}_X^s$ for τ in the interval between 0 and t. The main step in the proof will be to show that for every σ between 0 and s, the pullback of the vector field Y from $\mathcal{O}_X^{-\sigma}$ to $\mathcal{O}_X^\sigma$ via φ_X^σ matches Y itself on $\mathcal{O}_X^s$, i.e.

$$Y = (\varphi_X^\sigma)^* Y \quad \text{on} \quad \mathcal{O}_X^s. \tag{6.7}$$

Assuming this for the moment, it then follows from Proposition 5.4 and (6.7) that the path $\tau \mapsto \varphi_X^\sigma \circ \gamma(\tau)$ for τ between 0 and t is also a flow line of Y, namely the unique one beginning at $\varphi_X^\sigma(p)$, which proves

$$\varphi_Y^\tau(\varphi_X^\sigma(p)) = \varphi_X^\sigma(\gamma(\tau)) = \varphi_X^\sigma(\varphi_Y^\tau(p)).$$

It remains only to prove (6.7). Since the statement is clearly true for $\sigma = 0$, it will suffice to prove that the derivative of the family of vector fields $(\varphi_X^\sigma)^* Y$ with respect to the parameter σ vanishes at every point on $\mathcal{O}_X^s$ for all σ between 0 and s. To see this, we use the identities $[X,Y] = \mathcal{L}_X Y = 0$ and $\varphi_X^{\sigma+\tau} = \varphi_X^\tau \circ \varphi_X^\sigma$, which gives $(\varphi_X^{\sigma+\tau})^* = (\varphi_X^\sigma)^*(\varphi_X^\tau)^*$ by Exercise 5.5.

In the following, we will only need the latter relation for values of $\tau \in \mathbb{R}$ that are arbitrarily close to 0, thus we will be free to assume that any given point in the domain of φ_X^σ is also in the domain of $\varphi_X^{\sigma+\tau}$. Working over the open set $\mathcal{O}_X^s$, we now compute,

$$\begin{aligned}\frac{d}{d\sigma}(\varphi_X^\sigma)^*Y &= \frac{d}{d\tau}(\varphi_X^{\sigma+\tau})^*Y\bigg|_{\tau=0} = \frac{d}{d\tau}(\varphi_X^\sigma)^*(\varphi_X^\tau)^*Y\bigg|_{\tau=0} = (\varphi_X^\sigma)^*\left(\frac{d}{d\tau}(\varphi_X^\tau)^*Y\bigg|_{\tau=0}\right)\\ &= (\varphi_X^\sigma)^*(\mathcal{L}_XY) = 0.\end{aligned}$$

□

7. Tensors

It will turn out that many types of "geometric structure" on manifolds can be expressed in terms of multilinear maps on tangent and cotangent spaces, known collectively as *tensor fields.* Before beginning with the contents of this lecture, I should remind you that the Einstein summation convention (see §5.4) is in effect from now on—we are going to be needing it a lot. We will also need the following convenient notational device: for any pair of indices $i, j \in \{1, \ldots, n\}$, we define

$$\delta^{ij} = \delta_{ij} = \delta^i_j := \begin{cases}1 & \text{if } i = j,\\ 0 & \text{if } i \neq j.\end{cases}$$

The choice of whether each index is an upper or lower index will depend on the context, but the meaning will always be the same. So for example, if $\mathbf{A} \in \mathrm{GL}(n, \mathbb{R})$ is a matrix with entries $A^i{}_j$, the matrix-multiplication relation $\mathbf{A}\mathbf{A}^{-1} = \mathbb{1}$ becomes

$$A^i{}_j(A^{-1})^j{}_k = \delta^i_k.$$

Here it is very important to remember that by the summation convention, the symbol "$\sum_{j=1}^n$" has been omitted from the left hand side; we chose to write the first index of $A^i{}_j$ as an upper index and the second as a lower index mainly so that this use of the summation convention would work. Here is another example that already came up in our discussion of vector fields (cf. Exercise 6.2): if $(\mathcal{U}, x)$ and $(\widetilde{\mathcal{U}}, \widetilde{x})$ are two overlapping charts on a manifold M, then at every point in $\mathcal{U} \cap \widetilde{\mathcal{U}}$, the matrices with entries $\frac{\partial \widetilde{x}^i}{\partial x^j}$ and $\frac{\partial x^i}{\partial \widetilde{x}^j}$ are inverse to each other, as they are Jacobi matrices of inverse transition maps, thus

$$\frac{\partial \widetilde{x}^i}{\partial x^j}\frac{\partial x^j}{\partial \widetilde{x}^k} = \delta^i_k.$$

Other versions of δ will sometimes arise with the indices placed in various ways in order to make the summation convention work. This symbol is known as the **Kronecker delta**, and maybe it would have been called something different if it had been invented in the age of Covid-19, but here we are.

7.1. Motivational examples. In order to motivate the idea of a tensor field on a manifold, it's best to start with a few examples that are already somewhat familiar.

7.1.1. *One-forms.* Any smooth function $f : M \to \mathbb{R}$ has a differential

$$df : TM \to \mathbb{R},$$

whose restriction to each individual tangent space T_pM is a linear map $T_pM \to \mathbb{R}$ and thus an element of the cotangent space T_p^*M. In this sense, df is analogous to a vector field, but instead of associating a tangent vector $X(p) \in T_pM$ to every point $p \in M$, it associates a cotangent vector $d_pf \in T_p^*M$, thus defining a map

$$M \to T^*M : p \mapsto d_pf.$$

In general, a map

$$\lambda : TM \to \mathbb{R}$$

whose restriction to each individual tangent space is linear is called a **1-form** on M, or sometimes also a **dual vector field** or **covector field**. For each $p \in M$, it is common to denote the restriction $\lambda|_{T_pM} : T_pM \to \mathbb{R}$ by

$$\lambda_p \in T_p^*M = \mathrm{Hom}(T_pM, \mathbb{R}),$$

hence one can equivalently view a 1-form λ as associating to each point $p \in M$ a cotangent vector $\lambda_p \in T_p^*M$. For the special case where λ is the differential of a function f, we have been writing $d_pf \in T_p^*M$ for the restriction to T_pM, but the notation $(df)_p$ would also be sensible, and is preferred by many authors.[29]

Since we have not yet endowed the cotangent bundle T^*M with a smooth structure, we need to put some thought into defining what it means for a 1-form to be "smooth". The easiest way to do this is by writing it in local coordinates. Any chart $(\mathcal{U}, x)$ on M gives rise to coordinate functions $x^i : \mathcal{U} \to \mathbb{R}$ for $i = 1, \dots, n$, whose differentials dx^i are 1-forms on $\mathcal{U}$.

Proposition 7.1. *For each $p \in \mathcal{U}$, every element $\lambda \in T_p^*M$ can be expressed as a linear combination $\lambda = \lambda_i\, d_px^i$ for unique real numbers $\lambda_1, \dots, \lambda_n \in \mathbb{R}$. In other words, the differentials $d_px^1, \dots, d_px^n$ form a basis of T_p^*M.*

Proof. What's actually happening here is that $d_px^1, \dots, d_px^n$ is the dual basis to the basis of coordinate vector fields $\partial_1, \dots, \partial_n$ defined by the chart $(\mathcal{U}, x)$ at p; indeed, for each $i, j \in \{1, \dots, n\}$,

$$dx^i(\partial_j) = dx^i\left(\frac{\partial}{\partial x^j}\right) = \mathcal{L}_{\frac{\partial}{\partial x^j}} x^i = \frac{\partial x^i}{\partial x^j} = \delta^i_j.$$

The coefficients λ_i are thus given by $\lambda_i = \lambda(\partial_i)$. □

The 1-forms $dx^1, \dots, dx^n$ on $\mathcal{U}$ defined by a chart $(\mathcal{U}, x)$ are known as the **coordinate differentials**, and Proposition 7.1 implies that every 1-form λ can be written over the region $\mathcal{U}$ as

$$\lambda = \lambda_i\, dx^i,$$

where its uniquely determined **component** functions $\lambda_i : \mathcal{U} \to \mathbb{R}$ are given by

$$\lambda_i(p) := \lambda\left(\frac{\partial}{\partial x^i}(p)\right), \qquad p \in \mathcal{U}.$$

For example, the component functions of the differential df are precisely the partial derivatives of f, namely $(df)_i = df(\partial_i) = \partial_i f : \mathcal{U} \to \mathbb{R}$, giving rise to the formula

$$df = \partial_i f\, dx^i \qquad \text{on } \mathcal{U},$$

which was understood for at least two centuries in terms of "infinitessimal quantities" before it was given a mathematically rigorous meaning in terms of 1-forms.

Remark 7.2. Notice that while components of vector fields are written with upper indices, components of 1-forms get lower indices. This is necessary in order for the summation convention to work properly, since coordinate differentials come with upper indices.

[29] Or if one prefers to think of df as a function $M \to T^*M$, one can write $df(p)$ instead of d_pf or $(df)_p$. I have done that in some of my research papers, but will avoid it in these notes for the sake of consistency, as we have defined df as a function $TM \to \mathbb{R}$ rather than $M \to T^*M$.

EXERCISE 7.3. Suppose $(\mathcal{U}, x)$ and $(\widetilde{\mathcal{U}}, \widetilde{x})$ are two smooth charts with $\mathcal{U} \cap \widetilde{\mathcal{U}} \neq \varnothing$, so any 1-form λ can be written as both $\lambda_i \, dx^i$ and $\widetilde{\lambda}_i \, d\widetilde{x}^i$ in the overlap region. Prove the following coordinate transformation formulas on $\mathcal{U} \cap \widetilde{\mathcal{U}}$, analogous to the formulas (6.3) and (6.4) for vector fields:

$$dx^i = \frac{\partial x^i}{\partial \widetilde{x}^j} \, d\widetilde{x}^j \qquad \text{and} \qquad \widetilde{\lambda}_i = \lambda_j \frac{\partial x^j}{\partial \widetilde{x}^i}. \tag{7.1}$$

The formula (7.1) shows that if a 1-form has smooth component functions with respect to any given chart, its component functions in any other chart defined on the same domain will also be smooth, due to the fact that transition maps (and therefore also their derivatives $\frac{\partial x^i}{\partial \widetilde{x}^j}$) are smooth. The following definition therefore makes sense.

DEFINITION 7.4. A 1-form on M is said to be **smooth** if and only if its component functions with respect to every chart are smooth. The set of all smooth 1-forms on M forms a vector space, which we denote by

$$\Omega^1(M) := \{\text{smooth 1-forms on } M\}.$$

EXERCISE 7.5. Show that a 1-form λ on M is smooth if and only if the function $M \to \mathbb{R} : p \mapsto \lambda(X(p))$ is smooth for every smooth vector field $X \in \mathfrak{X}(M)$.

From now on, we will assume that all 1-forms we consider are smooth unless stated otherwise.

7.1.2. *Vector fields.* Recall that every finite-dimensional vector space V is naturally isomorphic to the dual of its dual space, with a canonical isomorphism $\Phi : V \to V^{**}$ given by

$$\Phi(v)\lambda := \lambda(v).$$

If we choose to, we can therefore also think of every tangent space T_pM as a dual space, namely $(T_p^*M)^*$, meaning that every vector field $X \in \mathfrak{X}(M)$ can equivalently be viewed as associating to each $p \in M$ a linear map $\tau_p : T_p^*M \to \mathbb{R}$, defined by $\tau_p(\lambda) := \lambda(X(p))$. I'm sure you can imagine why we didn't define vector fields this way in the first place, but we could have done so if we'd wanted to. From this perspective, the notion of smoothness for a vector field can also be characterized analogously to Exercise 7.5:

EXERCISE 7.6. Show that a vector field X on M is smooth if and only if the function $M \to \mathbb{R} : p \mapsto \lambda(X(p))$ is smooth for every smooth 1-form $\lambda \in \Omega^1(M)$.

7.1.3. *Riemannian metrics.* A Riemannian metric g on a manifold M associates to every point $p \in M$ an inner product g_p on T_pM, so in particular, g_p is a bilinear map

$$g_p : T_pM \times T_pM \to \mathbb{R}$$

that is also symmetric and positive-definite. We can think of g itself as a function

$$g : TM \oplus TM \to \mathbb{R},$$

where $TM \oplus TM := \bigcup_{p \in M}(T_pM \times T_pM)$. As a provisional notion of smoothness for Riemannian metrics, we can define g to be **smooth** if and only if the function

$$M \to \mathbb{R} : p \mapsto g(X(p), Y(p))$$

is smooth for every pair of smooth vector fields $X, Y \in \mathfrak{X}(M)$. Under this condition, g is an example of something we will shortly define as a "smooth covariant tensor field of rank 2" on M.

7.1.4. *Almost complex structures.* Here is an example you may not have heard of before. One can make any $2n$-dimensional real vector space V into an n-dimensional complex vector space by choosing a linear map $J : V \to V$ with $J^2 = -\mathbb{1}$ and defining complex scalar multiplication on V by $(a + ib)v := av + bJv$. Such a linear map J is therefore called a **complex structure** on V. It is sometimes useful to introduce such a structure on the tangent spaces of an even-dimensional manifold M. An **almost complex structure** (*fast komplexe Struktur*) on M is a map

$$J : TM \to TM$$

whose restriction to each individual tangent space is a complex structure $J_p : T_pM \to T_pM$. We can define J to be **smooth** if and only if the vector field $p \mapsto JX(p)$ is smooth for all smooth vector fields $X \in \mathfrak{X}(M)$. The following lemma gives an alternative algebraic way of understanding what an almost complex structure is.

LEMMA 7.7. *For a finite-dimensional real vector space V, let $\mathrm{End}(V) = \mathrm{Hom}(V, V)$ denote the vector space of all linear maps $V \to V$, $V^* = \mathrm{Hom}(V, \mathbb{R})$ the dual space of V, and $\mathrm{Hom}(V^* \otimes V, \mathbb{R})$ the vector space of all bilinear maps $V^* \times V \to \mathbb{R}$. There exists a canonical isomorphism*

$$\Phi : \mathrm{End}(V) \to \mathrm{Hom}(V^* \otimes V, \mathbb{R}), \qquad \Phi(A)(\lambda, v) := \lambda(Av).$$

PROOF. It is easy to check that Φ is a linear injection, and if $\dim V = n$, then $\dim \mathrm{End}(V) = \dim \mathrm{Hom}(V^* \otimes V, \mathbb{R}) = n^2$, thus Φ is also surjective. $\square$

For an almost complex structure J on M, Lemma 7.7 allows us to view $J_p : T_pM \to T_pM$ equivalently as a bilinear map $T_p^*M \times T_pM \to \mathbb{R}$, and from this perspective, one can check that J is smooth (according to our previous definition) if and only if the function $M \to \mathbb{R} : p \mapsto J(\lambda_p, X(p))$ is smooth for all choices of smooth vector field $X \in \mathfrak{X}(M)$ and smooth 1-form $\lambda \in \Omega^1(M)$.

7.2. Tensor fields in general. We now describe a more general notion that encompasses all of the examples in §7.1 as special cases.

Recall that for vector spaces $V_1, \dots, V_n$ and W, a map

$$T : V_1 \times \ldots \times V_n \to W$$

is called **multilinear** if it is linear with respect to each variable individually, i.e. for every $i = 1, \dots, n$ and every fixed tuple of vectors $v_j \in V_j$ for $j = 1, \dots, i-1, i+1, \dots, n$, the map

$$V_i \to W : v_i \mapsto T(v_1, \dots, v_n)$$

is linear. Observe that the space of all multilinear maps $V_1 \times \ldots \times V_n \to W$ is naturally also a finite-dimensional vector space. We will sometimes denote it by[30]

$$\mathrm{Hom}(V_1 \otimes \ldots \otimes V_n, W).$$

DEFINITION 7.8. For integers $k, \ell \geqslant 0$ with $k + \ell > 0$ and a finite-dimensional real vector space V, we will denote by V_ℓ^k the vector space of multilinear maps

$$\underbrace{V^* \times \ldots \times V^*}_{k} \times \underbrace{V \times \ldots \times V}_{\ell} \to \mathbb{R},$$

where V^* as usual denotes the dual space $\mathrm{Hom}(V, \mathbb{R})$. In the case $k = \ell = 0$, we define $V_0^0 = \mathbb{R}$.

[30]We will not make use of the abstract algebraic notion of the tensor product of vector spaces in this lecture, but readers already familiar with that notion may want to pause and consider why our definition of the symbol "$\mathrm{Hom}(V_1 \otimes \ldots \otimes V_n, W)$" is equivalent to the one they've seen before. It is important that we are explicitly assuming all vector spaces to be finite dimensional in this discussion; if we did not assume this, then some more serious digressions into the meaning of the symbol "$\otimes$" would be necessary.

REMARK 7.9. To motivate the convention $V_0^0 = \mathbb{R}$, you can imagine perhaps that a "real-valued multilinear function of *zero* variables" is the same thing as a real number. If that doesn't convince you, the convention will at least begin to seem more natural when we discuss tensor products (cf. Remark 7.19).

DEFINITION 7.10. For a smooth manifold M and integers $k, \ell \geqslant 0$, a **tensor field** (*Tensorfeld*) S of type (k, ℓ) associates to each point $p \in M$ an element

$$S_p \in (T_pM)^k_\ell.$$

If $k + \ell > 0$, then the tensor field S is said to be **smooth** if and only if the function $M \to \mathbb{R} : p \mapsto S_p(\lambda^1_p, \dots, \lambda^k_p, X_1(p), \dots, X_\ell(p))$ is smooth for every tuple of smooth vector fields $X_1, \dots, X_\ell \in \mathfrak{X}(M)$ and smooth 1-forms $\lambda^1, \dots, \lambda^k \in \Omega^1(M)$. We will denote the vector space of smooth tensor fields by

$$\Gamma(T^k_\ell M) := \{\text{smooth tensor fields of type } (k, \ell)\}.$$

For $k = \ell = 0$, a tensor field is just a real-valued function on M, so we define $\Gamma(T^0_0 M) := C^\infty(M)$.

The **support** (*Träger*) of a tensor field $S \in \Gamma(T^k_\ell M)$ is defined as the closure in M of the set $\{p \in M \mid S_p \neq 0\}$.

EXAMPLE 7.11. A smooth 1-form is equivalently a smooth tensor field of type $(0, 1)$:

$$\Omega^1(M) = \Gamma(T^0_1 M).$$

Just as 1-forms $\lambda \in \Omega^1(M)$ are regarded as functions $TM \to \mathbb{R}$, it will often be useful to regard a tensor field $S \in \Gamma(T^k_\ell M)$ in the case $k + \ell > 0$ as a function

$$S : T^*M^{\oplus k} \oplus TM^{\oplus \ell} \to \mathbb{R},$$

where we introduce the notation

$$T^*M^{\oplus k} \oplus TM^{\oplus \ell} := \bigcup_{p \in M} \left(\underbrace{T^*_pM \times \dots \times T^*_pM}_{k} \times \underbrace{T_pM \times \dots \times T_pM}_{\ell} \right).$$

The key property of S is then that its restriction S_p to $T^*_pM \times \dots \times T^*_pM \times T_pM \times \dots \times T_pM \subset T^*M^{\oplus k} \oplus TM^{\oplus \ell}$ for each $p \in M$ is a multilinear map.

In the setting of smooth manifolds, the term "tensor field" is often abbreviated simply as **tensor**. The terminology for tensors of type (k, ℓ) can also vary among different sources, e.g. one sometimes says that a tensor $S \in \Gamma(T^k_\ell M)$ is **contravariant of rank** k and **covariant of rank** ℓ. The latter terminology is especially favored among physicists.

EXAMPLE 7.12. Under the canonical isomorphism identifying each tangent space T_pM with $\mathrm{Hom}(T^*_pM, \mathbb{R})$, a smooth vector field becomes the same thing as a smooth tensor field of type $(1, 0)$, hence

$$\mathfrak{X}(M) = \Gamma(T^1_0 M).$$

Here the function $T^*M \to \mathbb{R}$ corresponding to a given vector field $X \in \mathfrak{X}(M)$ sends $\lambda \in T^*_pM$ to $\lambda(X(p))$.

EXAMPLE 7.13. Every Riemannian metric (see §7.1.3) is an example of a tensor field of type $(0, 2)$.

EXAMPLE 7.14. Every almost complex structure (see §7.1.4) is an example of a tensor field of type $(1, 1)$.

Exercise 7.15. Generalize Lemma 7.7 to show the following: for any finite-dimensional real vector spaces $V_1, \dots, V_n, W$, there exists a canonical isomorphism

$$\mathrm{Hom}(V_1 \otimes \dots \otimes V_n, W) \xrightarrow{\Phi} \mathrm{Hom}(W^* \otimes V_1 \otimes \dots \otimes V_n, \mathbb{R}),$$
$$\Phi(A)(\lambda, v_1, \dots, v_n) := \lambda(A(v_1, \dots, v_n)).$$

Example 7.16. For arbitrary integers $\ell \geqslant 1$, Exercise 7.15 identifies any tensor field S of type $(1, \ell)$ with a map

$$\bigcup_{p \in M} \Big(\underbrace{T_pM \times \dots \times T_pM}_{\ell} \Big) =: TM^{\oplus \ell} \xrightarrow{\widehat{S}} TM$$

whose restriction $\widehat{S}_p$ to $T_pM \times \dots \times T_pM$ for each $p \in M$ is a multilinear map $T_pM \times \dots \times T_pM \to T_pM$. The precise correspondence between S and $\widehat{S}$ is given by

$$S(\lambda, X_1, \dots, X_\ell) = \lambda\big(\widehat{S}(X_1, \dots, X_\ell)\big),$$

and it is straightforward to show that S is smooth if and only if $\widehat{S}(X_1, \dots, X_\ell)$ defines a smooth vector field for all choices of smooth vector fields $X_1, \dots, X_\ell \in \mathfrak{X}(M)$. The case $\ell = 0$ also fits into this picture if one adopts the perspective that a "T_pM-valued function of zero variables" just means an element of T_pM: this reproduces the observation in Example 7.12 that tensor fields of type $(1, 0)$ are equivalent to vector fields.

Remark 7.17. The alternative perspective on tensors of type $(1, \ell)$ in Example 7.16 will generally be quite useful, and from now on we will typically use the same notation for the objects that are called S and $\widehat{S}$ in that example. We have already adopted this convention in our discussion of vector fields and almost complex structures as tensors of type $(1, 0)$ and $(1, 1)$ respectively.

Definition 7.18. For $S \in \Gamma(T^k_\ell M)$ and $T \in \Gamma(T^r_s M)$, the **tensor product** (*Tensorprodukt*) of S and T is the tensor field $S \otimes T \in \Gamma(T^{k+r}_{\ell+s} M)$ defined at each point $p \in M$ by

$$(S \otimes T)_p(\lambda^1, \dots, \lambda^k, \mu^1, \dots, \mu^r, X_1, \dots, X_\ell, Y_1, \dots, Y_s) :=$$
$$S_p(\lambda^1, \dots, \lambda^k, X_1, \dots, X_\ell) \cdot T_p(\mu^1, \dots, \mu^r, Y_1, \dots, Y_s).$$

Remark 7.19. For $f \in C^\infty(M) = \Gamma(T^0_0 M)$, the tensor product of f with $S \in \Gamma(T^k_\ell M)$ is just the ordinary point-wise product of S with a scalar-valued function, i.e. $(f \otimes S)_p = (S \otimes f)_p = f(p) S_p$.

7.3. Coordinate representations. We've seen that a chart $(\mathcal{U}, x)$ on M gives rise to coordinate vector fields $\partial_1, \dots, \partial_n \in \mathfrak{X}(\mathcal{U})$ and coordinate differentials $dx^1, \dots, dx^n \in \Omega^1(\mathcal{U})$ which define bases of T_pM and T_p^*M respectively at each point $p \in \mathcal{U}$. Regarding vector fields as tensors of type $(1, 0)$, it turns out that a natural basis of $(T_pM)^k_\ell$ can then be constructed by taking all possible tensor products of k coordinate vector fields with ℓ coordinate differentials. Indeed:

Proposition 7.20. *Given a chart $(\mathcal{U}, x)$ on an n-manifold M, every tensor field S of type (k, ℓ) can be written uniquely over $\mathcal{U}$ as*

$$S = S^{i_1 \dots i_k}{}_{j_1 \dots j_\ell} \frac{\partial}{\partial x^{i_1}} \otimes \dots \otimes \frac{\partial}{\partial x^{i_k}} \otimes dx^{j_1} \otimes \dots \otimes dx^{j_\ell}, \tag{7.2}$$

where the $n^{k+\ell}$ ***component*** *functions $S^{i_1 \dots i_k}{}_{j_1 \dots j_\ell} : \mathcal{U} \to \mathbb{R}$ are given by*

$$S^{i_1 \dots i_k}{}_{j_1 \dots j_\ell} := S(dx^{i_1}, \dots, dx^{i_k}, \partial_{j_1}, \dots, \partial_{j_\ell}).$$

Remark 7.21. Writing down (7.2) without the Einstein summation convention would have required inserting the symbols

$$\sum_{i_1=1}^{n} \cdots \sum_{i_k=1}^{n} \sum_{j_1=1}^{n} \cdots \sum_{j_\ell=1}^{n}$$

just to the right of the equal sign, so the right hand side is actually a sum of $n^{k+\ell}$ terms.

Proof of Proposition 7.20. Any ℓ vector fields can be written over $\mathcal{U}$ as $X_a = X_a^i\, \partial_i$ for $a = 1, \dots, \ell$ with unique component functions $X_a^i : \mathcal{U} \to \mathbb{R}$, and similarly, any k 1-forms can be written as $\lambda^b = \lambda^b_j\, dx^j$ with unique components $\lambda^b_j : \mathcal{U} \to \mathbb{R}$. By multilinearity, we then have

$$\begin{aligned} S(\lambda^1, \dots, \lambda^k, X_1, \dots, X_\ell) &= S(\lambda^1_{i_1}\, dx^{i_1}, \dots, \lambda^k_{i_k}\, dx^{i_k}, X_1^{j_1}\, \partial_{j_1}, \dots, X_\ell^{j_\ell}\, \partial_{j_\ell}) \\ &= S^{i_1 \dots i_k}{}_{j_1 \dots j_\ell}\, \lambda^1_{i_1} \dots \lambda^k_{i_k} X_1^{j_1} \dots X_\ell^{j_\ell}. \end{aligned} \tag{7.3}$$

It is straightforward to check that the tensor field on the right hand side of (7.2) gives the same result when evaluated on the same tuple of vector fields and 1-forms. □

Exercise 7.22. Show that a tensor field of type (k, ℓ) is smooth if and only if for every smooth chart, the corresponding component functions are all smooth.

Exercise 7.23. Show that in local coordinates, the components of two tensor fields $S \in \Gamma(T^k_\ell M)$, $T \in \Gamma(T^r_s M)$ and their tensor product $S \otimes T \in \Gamma(T^{k+r}_{\ell+s} M)$ are related by

$$(S \otimes T)^{i_1 \dots i_k a_1 \dots a_r}{}_{j_1 \dots j_\ell b_1 \dots b_s} = S^{i_1 \dots i_k}{}_{j_1 \dots j_\ell}\, T^{a_1 \dots a_r}{}_{b_1 \dots b_s}.$$

Exercise 7.24. Suppose $(\mathcal{U}, x)$ and $(\widetilde{\mathcal{U}}, \widetilde{x})$ are two smooth charts with $\mathcal{U} \cap \widetilde{\mathcal{U}} \ne \varnothing$, and denote the component functions of a tensor field $S \in \Gamma(T^k_\ell M)$ with respect to each chart by $S^{i_1 \dots i_k}{}_{j_1 \dots j_\ell}$ and $\widetilde{S}^{i_1 \dots i_k}{}_{j_1 \dots j_\ell}$ respectively. Prove that on the overlap region $\mathcal{U} \cap \widetilde{\mathcal{U}}$,

$$\widetilde{S}^{i_1 \dots i_k}{}_{j_1 \dots j_\ell} = \frac{\partial \widetilde{x}^{i_1}}{\partial x^{a_1}} \cdots \frac{\partial \widetilde{x}^{i_k}}{\partial x^{a_k}} S^{a_1 \dots a_k}{}_{b_1 \dots b_\ell} \frac{\partial x^{b_1}}{\partial \widetilde{x}^{j_1}} \cdots \frac{\partial x^{b_\ell}}{\partial \widetilde{x}^{j_\ell}}. \tag{7.4}$$

Hint: Use (6.3) *and* (7.1).

Remark 7.25. We have been writing all tensor fields so far as functions that take covectors $\lambda^1, \dots, \lambda^k$ followed by vectors $X_1, \dots, X_\ell$, but in some circumstances, one may want to be more flexible with the ordering, so that e.g. a tensor of type $(1, 2)$ could be written as a multilinear function

$$TM \oplus T^*M \oplus TM \to \mathbb{R} : (X, \lambda, Y) \mapsto S(X, \lambda, Y).$$

The component functions of such a tensor would then be written as $S_i{}^j{}_k$, with evaluation on $X = X^i\, \partial_i$, $\lambda = \lambda^j\, dx^j$ and $Y = Y^k\, \partial_k$ defined by the rule

$$S(X, \lambda, Y) = S_i{}^j{}_k X^i \lambda_j Y^k.$$

Example 7.26. Suppose $J : TM \to TM$ is an almost complex structure, so $J_p : T_pM \to T_pM$ is a linear map satisfying $J_p^2 = -\mathbb{1}$ for every $p \in M$. As we've seen, J can be regarded as a tensor field of type $(1, 1)$ and thus defines a function $T^*M \oplus TM \to \mathbb{R}$, with component functions with respect to a chart $(\mathcal{U}, x)$ written as

$$J^i{}_j = J(dx^i, \partial_j) := dx^i(J\partial_j), \qquad i, j \in \{1, \dots, n\}.$$

In this line, the second expression views J_p as a bilinear map $T_p^*M \times T_pM \to \mathbb{R}$, while the third views it as a linear map $T_pM \to T_pM$. This means that for two tangent vectors $X = X^i\, \partial_i$ and $Y = Y^i\, \partial_i$ at a point $p \in \mathcal{U}$, we have

$$JX = Y \qquad \Longleftrightarrow \qquad Y^i = dx^i(Y) = dx^i(JX) = dx^i(J(X^j\, \partial_j)) = X^j\, dx^i(J\partial_j) = J^i{}_j X^j,$$

so in other words, the linear map $J_p : T_pM \to T_pM$ is represented in coordinates by matrix-vector multiplication: the n-by-n matrix with entries $J^i{}_j$ gets multiplied by the n-dimensional row vector with entries X^j to produce the row vector with entries $(JX)^i$. The condition $J^2 = -\mathbb{1}$ can thus be expressed in local coordinates on $\mathcal{U}$ as

$$J^i{}_j J^j{}_k \equiv -\delta^i_k \qquad \text{on } \mathcal{U}.$$

From this perspective, the transformation formula (7.4) also ends up looking like something familiar from linear algebra: the component functions $J^i{}_j$ and $\widetilde{J}^i{}_j$ for two overlapping charts $(\mathcal{U}, x)$ and $(\widetilde{\mathcal{U}}, \widetilde{x})$ are related by

$$\widetilde{J}^i{}_j = \frac{\partial \widetilde{x}^i}{\partial x^k} J^k{}_\ell \frac{\partial x^\ell}{\partial \widetilde{x}^j}.$$

In terms of matrices, this just says

$$\widetilde{\mathbf{J}} = \left(\frac{\partial \widetilde{x}}{\partial x}\right) \mathbf{J} \left(\frac{\partial \widetilde{x}}{\partial x}\right)^{-1},$$

where $\mathbf{J}$ and $\widetilde{\mathbf{J}}$ denote the n-by-n matrices with entries $J^i{}_j$ and $\widetilde{J}^i{}_j$ respectively, while $\frac{\partial \widetilde{x}}{\partial x}$ is the n-by-n Jacobian matrix with entries $\frac{\partial \widetilde{x}^i}{\partial x^j}$.

8. Derivatives of tensors and differential forms

We motivate this lecture with the following question: for a smooth tensor field $S \in \Gamma(T^k_\ell M)$, can one define a "directional derivative" of S at a point $p \in M$ in the direction $X \in T_pM$? We considered this question for the special case of vector fields $Y \in \mathfrak{X}(M) = \Gamma(T^1_0 M)$ in §6.3, and the answer we came up with there was not entirely satisfactory: a vector field Y can be differentiated with respect to another vector field X, producing the Lie derivative $\mathcal{L}_X Y \in \mathfrak{X}(M)$, but $\mathcal{L}_X Y(p)$ depends on X as a vector field, not just on the value $X(p)$ (see Remark 6.8). Naively, one might hope for instance that if $S \in \Gamma(T^k_\ell M)$ has components $S^{i_1 \dots i_k}{}_{j_1 \dots j_\ell}$ with respect to some chart $(\mathcal{U}, x)$, then one could define a tensor "dS" of type $(k, \ell + 1)$ whose components are

$$\text{“}(dS)^{i_1 \dots i_k}{}_{j_0 \dots j_\ell} = \partial_{j_0} S^{i_1 \dots i_k}{}_{j_1 \dots j_\ell}\text{”}, \tag{8.1}$$

so that for any $p \in M$ and $X \in T_pM$, the multilinear map $(dS)(\dots, X, \dots) : (T^*_p M)^{\times k} \times (T_pM)^{\times \ell} \to \mathbb{R}$ could be interpreted as the derivative of S in the direction X. But I put that expression in quotation marks because, indeed, it doesn't work: outside of the special case $k = \ell = 0$ where the objects we are differentiating are just real-valued functions, one cannot define from $S \in \Gamma(T^k_\ell M)$ any tensor field $dS \in \Gamma(T^k_{\ell+1} M)$ whose components are given in *all* choices of local coordinates by (8.1). (Exercise 8.1(b) below asks you to prove this in the case $(k, \ell) = (0, 1)$.) In other words, the formula (8.1) is not *coordinate invariant*.

Before discussing directional derivatives further, we should talk about a sticky issue that arose in the previous paragraph: what *practical* methods do we have for writing down the definition of a tensor field? What we attempted above could be called the *physicists' method*: it starts by choosing a chart $(\mathcal{U}, x)$ and writing down a formula for the component functions of the tensor with respect to those local coordinates. That is fine if one only needs a tensor field defined on the subset $\mathcal{U} \subset M$, but the hope of course is that the formula we write down might be valid in *arbitrary* local coordinates, in which case it gives a well-defined tensor field everywhere on M. The important step is therefore to check, using the transformation formula (7.4), that the definition we've written is coordinate invariant, and that is what fails in the case of (8.1). On the other hand, sometimes it succeeds, for instance:

EXERCISE 8.1. Prove:

(a) For any $\lambda \in \Gamma(T_1^0 M)$, there exists a tensor field $S \in \Gamma(T_2^0 M)$ whose components S_{ij} with respect to arbitrary charts $(\mathcal{U}, x)$ are related to the corresponding components λ_i of λ by
$$S_{ij} = \partial_i \lambda_j - \partial_j \lambda_i.$$

(b) For general choices of λ, one cannot similarly define $S \in \Gamma(T_2^0 M)$ so that its relation to λ in arbitrary local coordinates is $S_{ij} = \partial_i \lambda_j$.

Physicists like to summarize the result of Exercise 8.1(a) by saying that the expression $\partial_i \lambda_j - \partial_j \lambda_i$ "defines a tensor" of type $(0,2)$. In fact, many textbooks on general relativity give a definition of tensors that is cosmetically quite different from ours: without mentioning multilinear maps, they define a tensor S of type (k,ℓ) as an association to each chart $(\mathcal{U}, x)$ of a collection of real-valued functions $S^{i_1 \dots i_k}{}_{j_1 \dots j_\ell} : \mathcal{U} \to \mathbb{R}$ that satisfy the transformation formula (7.4). There are good theoretical reasons why mathematicians do not usually give that as the definition of a tensor field, and contrary to what many physicists may tell you, it is also not true that defining a tensor or computing something from it always requires choosing local coordinates.

8.1. C^∞-linearity. Here is a trick for writing down tensor fields that mathematicians tend to prefer, because it does not require local coordinates. For example, let us regard a tensor field S of type $(1,\ell)$ as associating to each point $p \in M$ an ℓ-fold multilinear map $S_p : T_pM \times \ldots \times T_pM \to T_pM$, as described in Example 7.16. It therefore also defines a multilinear map

$$S : \underbrace{\mathfrak{X}(M) \times \ldots \times \mathfrak{X}(M)}_{\ell} \to \mathfrak{X}(M), \tag{8.2}$$

by interpreting $S(X_1, \ldots, X_\ell)$ for any tuple of smooth vector fields $X_1, \ldots, X_\ell$ as the vector field
$$p \mapsto S_p(X_1(p), \ldots, X_\ell(p)).$$
We already know one important concrete example of multilinear map of this type: the Lie bracket is a bilinear map
$$[\cdot, \cdot] : \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M).$$
But does the Lie bracket therefore define a tensor field of type $(1,2)$? It would be surprising if this were true, because being a tensor field would imply that the value $[X,Y](p)$ for each $p \in M$ depends only on the values $X(p)$ and $Y(p)$, whereas we saw in Exercise 6.2 that in local coordinates, $[X,Y](p)$ also depends on the first derivatives of X and Y at p. An easy way to make this intuition more precise is via the following observation: if S is a tensor field, then the map in (8.2) is not just multilinear, it also satisfies

$$S(X_1, \ldots, X_{j-1}, fX_j, X_{j+1}, \ldots, X_\ell) = fS(X_1, \ldots, X_\ell) \quad \text{for all } f \in C^\infty(M) \tag{8.3}$$

for every $j = 1, \ldots, \ell$. The key point here is that the function f does not need to be constant, so this is a much stronger statement than just saying that (8.2) respects scalar multiplication (as every multilinear map must). A multilinear map on the space of vector fields is said to be **C^∞-linear** in its jth argument if it satisfies (8.3). In general, the notion of C^∞-linearity can be defined for multilinear maps between any vector spaces on which there is a natural notion of multiplication by smooth functions[31], e.g. we had $\mathfrak{X}(M)$ in the above example because the product of a smooth vector field with a smooth function is also a smooth vector field, but for similar reasons, one could just as well work with $\Omega^1(M)$, the other spaces of smooth tensor fields $\Gamma(T_\ell^k M)$, or $C^\infty(M)$ itself. From this perspective, the obvious reason why the Lie bracket does not define a tensor field is that it is not C^∞-linear: according to Exercise 6.4, it satisfies
$$[fX, Y] = f[X,Y] - (\mathcal{L}_Y f)X, \qquad [X, fY] = f[X,Y] + (\mathcal{L}_X f)Y,$$

[31]in other words, spaces that are naturally *modules* over $C^\infty(M)$

for $f \in C^\infty(M)$, which is not the desired relation except in the special case where f is constant.

It will be exceedingly useful to observe that C^∞-linearity is not only necessary for a multilinear map on vector fields or 1-forms to define a tensor field—it is also sufficient.

PROPOSITION 8.2. *For a multilinear map*

$$S : \underbrace{\Omega^1(M) \times \ldots \times \Omega^1(M)}_{k} \times \underbrace{\mathfrak{X}(M) \times \ldots \times \mathfrak{X}(M)}_{\ell} \to C^\infty(M)$$

that is C^∞-linear in every argument, there exists a unique tensor field $\hat{S} \in \Gamma(T^k_\ell M)$ such that for every $p \in M$, $X_1, \ldots, X_\ell \in \mathfrak{X}(M)$ and $\lambda^1, \ldots, \lambda^k \in \Omega^1(M)$,

$$\hat{S}(\lambda^1_p, \ldots, \lambda^k_p, X_1(p), \ldots, X_\ell(p)) = S(\lambda^1, \ldots, \lambda^k, X_1, \ldots, X_\ell)(p).$$

Before proving the theorem, let us observe that it can be adapted easily for the slightly different situation in (8.2), where our multilinear map takes values in $\mathfrak{X}(M)$ instead of $C^\infty(M)$:

EXERCISE 8.3. Deduce from Proposition 8.2 that for any multilinear map

$$S : \underbrace{\mathfrak{X}(M) \times \ldots \times \mathfrak{X}(M)}_{\ell} \to \mathfrak{X}(M)$$

that is C^∞-linear in every argument, there exists a unique tensor field $\hat{S} \in \Gamma(T^1_\ell M)$ such that for every $p \in M$, $X_1, \ldots, X_\ell \in \mathfrak{X}(M)$, the multilinear map $\hat{S}_p : T_pM \times \ldots \times T_pM \to T_pM$ satisfies

$$\hat{S}(X_1(p), \ldots, X_\ell(p)) = S(X_1, \ldots, X_\ell)(p).$$

PROOF OF PROPOSITION 8.2. Let us consider only the case $\ell = 1$ and $k = 0$, as there is no substantial difference in the general case beyond requiring more complicated notation. We therefore assume $\Lambda : \mathfrak{X}(M) \to C^\infty(M)$ is a linear map satisfying $\Lambda(fX) = f\Lambda(X)$ for all $f \in C^\infty(M)$ and $X \in \mathfrak{X}(M)$, and we need to find a smooth 1-form $\lambda \in \Omega^1(M)$ such that $\lambda(X(p)) = \Lambda(X)(p)$ for all $p \in M$ and $X \in \mathfrak{X}(M)$. The uniqueness of λ is clear, since every tangent vector at a point $p \in M$ can be the value at that point of a smooth vector field (just write it down in local coordinates, multiply by a smooth cutoff function and extend outside of the coordinate neighborhood as 0).

To prove existence, it suffices to show that for any point $p \in M$, the value of $\Lambda(X)(p)$ is completely determined by $X(p)$ and does not otherwise depend on the choice of vector field X having this particular value at p. This will follow from linearity after proving two claims:

***Claim 1**: If $X \in \mathfrak{X}(M)$ vanishes in a neighborhood of p, then $\Lambda(X)(p) = 0$.*

Indeed, if $\mathcal{U} \subset M$ is an open neighborhood on which X vanishes, choose a smooth function $\beta : M \to [0, 1]$ with compact support in $\mathcal{U}$ satisfying $\beta(p) = 1$. Then $\beta X \equiv 0$, thus by C^∞-linearity,

$$0 = \Lambda(\beta X) = \beta\Lambda(X) \in C^\infty(M),$$

implying in particular that $\Lambda(X)(p) = \beta(p)\Lambda(X)(p) = 0$.

***Claim 2**: If $X \in \mathfrak{X}(M)$ satisfies $X(p) = 0$, then $\Lambda(X)(p) = 0$.*

To see this, choose a chart $(\mathcal{U}, x)$ with $p \in \mathcal{U}$, and write $X = X^i\partial_i$ on $\mathcal{U}$, so the functions $X^i \in C^\infty(\mathcal{U})$ satisfy $X^1(p) = \ldots = X^n(p) = 0$. Using smooth cutoff functions, we can also choose global vector fields $e_1, \ldots, e_n \in \mathfrak{X}(M)$ and functions $f^1, \ldots, f^n \in C^\infty(M)$ such that

$$f^i = X^i \quad \text{and} \quad e_i = \partial_i \qquad \text{near } p \text{, for all } i = 1, \ldots, n,$$

producing another vector field $Y := f^i e_i \in \mathfrak{X}(M)$ which matches X on some small neighborhood of p within $\mathcal{U}$. Claim 1 then implies $\Lambda(Y - X)(p) = \Lambda(Y)(p) - \Lambda(X)(p) = 0$. In light of C^∞-linearity and the condition $f^i(p) = X^i(p) = 0$ for $i = 1, \ldots, n$, we then have

$$\Lambda(X)(p) = \Lambda(Y)(p) = \Lambda(f^i e_i)(p) = f^i(p)\Lambda(e_i)(p) = 0.$$

□

From now on, we will say that a multilinear map on the spaces of vector fields and/or 1-forms **defines a tensor** whenever it is C^∞-linear in every argument, so that Proposition 8.2 or its obvious corollaries such as Exercise 8.3 apply. We can now carry out the "coordinate free" version of Exercise 8.1:

EXERCISE 8.4. Show that for any given 1-form $\lambda \in \Omega^1(M)$, the tensor of type $(0,2)$ that was defined via coordinates in Exercise 8.1 can also be defined via the bilinear map

$$\mathfrak{X}(M) \times \mathfrak{X}(M) \to C^\infty(M) : (X,Y) \mapsto \mathcal{L}_X\left[\lambda(Y)\right] - \mathcal{L}_Y\left[\lambda(X)\right] - \lambda([X,Y]),$$

which is C^∞-linear in both arguments. (In this expression, we associate to each vector field $Z \in \mathfrak{X}(M)$ the smooth real-valued function $\lambda(Z) \in C^\infty(M)$ whose value at $p \in M$ is $\lambda(Z(p))$.)

EXERCISE 8.5. Suppose $J \in \Gamma(T^1_1 M)$ is a smooth almost complex structure, which we will regard as a smooth map $J : TM \to TM$ whose restriction to each tangent space T_pM is a linear map $J_p : T_pM \to T_pM$ with $J_p^2 = -\mathbb{1}$. The **Nijenhuis tensor**[32] is defined from J via the map

$$N : \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M), \qquad N(X,Y) := [JX,JY] - J[JX,Y] - J[X,JY] - [X,Y].$$

(a) Use Exercise 8.3 to prove that this formula defines a tensor field of type $(1,2)$.
(b) Show that in local coordinates, the components of N and J are related by

$$N^i_{\ jk} = J^\ell_{\ j}\,\partial_\ell J^i_{\ k} - J^\ell_{\ k}\,\partial_\ell J^i_{\ j} + J^i_{\ \ell}\left(\partial_k J^\ell_{\ j} - \partial_j J^\ell_{\ k}\right).$$

(c) Show that N vanishes identically if $\dim M = 2$.
Hint: Notice that $N(X,Y)$ is antisymmetric in X and Y. What is $N(X,JX)$?
(d) An almost complex structure J is called *integrable* if near every point $p \in M$ there exists a chart $(\mathcal{U}, x)$ in which the components $J^i_{\ j}$ become the entries of the constant matrix

$$\mathbf{J}_0 := \begin{pmatrix} 0 & -\mathbb{1} \\ \mathbb{1} & 0 \end{pmatrix} \in \mathbb{R}^{2n\times 2n},$$

where each of the four blocks is an n-by-n matrix and $\dim M = 2n$. Show that if J is integrable, then $N \equiv 0$.
Advice: One can use the formula in part (b) for this, but an argument based directly on the definition of N via Lie brackets is also possible.

Remark: The matrix $\mathbf{J}_0$ represents the linear transformation $\mathbb{C}^n \to \mathbb{C}^n : \mathbf{z} \mapsto i\mathbf{z}$ if one identifies $\mathbb{C}^n$ with $\mathbb{R}^{2n}$ via the correspondence $\mathbb{C}^n \ni \mathbf{x} + i\mathbf{y} \leftrightarrow (\mathbf{x},\mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^n = \mathbb{R}^{2n}$, thus an integrable almost complex structure makes M into a "complex manifold". By a deep theorem of Newlander and Nirenberg from 1957, the converse of part (d) is also true: if the Nijenhuis tensor vanishes, then J is integrable.

8.2. Differential forms and the exterior derivative. In Exercises 8.1 and 8.4, we saw that if we "antisymmetrize" the partial derivatives of the components of a 1-form, the result is a well-defined tensor field of type $(0,2)$. We shall now generalize this observation, and in the process, introduce an important special class of tensor fields that will play a major role when we discuss integration on manifolds.

A multilinear map $T : V \times \ldots \times V \to W$ is called **antisymmetric** (*antisymmetrisch*) or **skew-symmetric** (*schiefsymmetrisch*) or **alternating** if the value $T(v_1,\ldots,v_n)$ changes by a sign whenever any two of its arguments are interchanged. One can express this condition equivalently in terms of arbitrary permutations: let S_n denote the **symmetric group** on n elements, which consists of all bijections from the set $\{1,\ldots,n\}$ to itself, also known as **permutations** (*Permutationen*). There are exactly $n!$ elements in S_n, and the group is generated by the so-called *flips*,

[32] Approximate pronounciation: "NIGH-en-house", where "nigh" rhymes with English "sigh".

which satisfy $\sigma(i) = j$ and $\sigma(j) = i$ for two distinct elements $i, j \in \{1, \dots, n\}$ while leaving every other element fixed. Every permutation can therefore be expressed as a composition of flips, and while a given permutation will generally admit many distinct decompositions into varying numbers of flips, one can show that for any fixed $\sigma \in S_n$, the number of flips required is always either even or odd, i.e. a composition of evenly many flips cannot also be expressed as a composition of an odd number of flips, or vice versa. We call each permutation $\sigma \in S_n$ **even** (*gerade*) or **odd** (*ungerade*) accordingly, and define its **parity** by[33]

$$|\sigma| := \begin{cases} 0 & \text{if } \sigma \text{ is even,} \\ 1 & \text{if } \sigma \text{ is odd.} \end{cases}$$

In applications, the parity usually appears in the form $(-1)^{|\sigma|}$, thus one sometimes also refers to odd or even permutations as *negative* or *positive* respectively. With this notion in place, a multilinear map $T : \underbrace{V \times \ldots \times V}_{n} \to W$ is antisymmetric if and only if it satisfies

$$T(v_{\sigma(1)}, \dots, v_{\sigma(n)}) = (-1)^{|\sigma|} T(v_1, \dots, v_n)$$

for all $v_1, \dots, v_n \in V$ and $\sigma \in S_n$. One can turn any multilinear map $T : V \times \ldots \times V \to W$ into one that is antisymmetric by defining

$$(\operatorname{Alt} T)(v_1, \dots, v_n) := \frac{1}{n!} \sum_{\sigma \in S_n} (-1)^{|\sigma|} T(v_{\sigma(1)}, \dots, v_{\sigma(n)}).$$

We observe that $\operatorname{Alt}(T) = T$ if and only if T is antisymmetric, thus Alt defines a linear projection map $\operatorname{Hom}(\bigotimes^n V, W) \to \operatorname{Hom}(\bigotimes^n V, W)$ onto the subspace of antisymmetric maps.

Definition 8.6. For any integer $k \geqslant 0$, an antisymmetric tensor field of type $(0, k)$ on M is called a **differential k-form** (or just **k-form** for short). The vector space of smooth k-forms on M is denoted by

$$\Omega^k(M) := \{\text{smooth } k\text{-forms on } M\}.$$

Note that antisymmetry is a vacuous condition in the cases $k = 0, 1$, which is why $\Omega^1(M) = \Gamma(T_1^0 M)$ and $\Omega^0(M) = \Gamma(T_0^0 M) = C^\infty(M)$. Given a chart $(\mathcal{U}, x)$, a k-form $\omega \in \Omega^k(M)$ can be written in local coordinates as

$$\omega = \omega_{i_1 \dots i_k} \, dx^{i_1} \otimes \ldots \otimes dx^{i_k} \qquad \text{on } \mathcal{U},$$

where antisymmetry means that the component functions $\omega_{i_1 \dots i_k} : \mathcal{U} \to \mathbb{R}$ change by a sign whenever two of the indices are interchanged. In this context, the following notational device is often useful. Suppose $T_{i_1 \dots i_k}$ is a collection of symbols associating to each k-tuple of integers $i_1, \dots, i_k \in \{1, \dots, n\}$ an element of some vector space, e.g. $C^\infty(\mathcal{U})$ in the example above. We can then **antisymmetrize** these symbols to define

$$T_{[i_1 \dots i_k]} := \frac{1}{k!} \sum_{\sigma \in S_k} (-1)^{|\sigma|} T_{i_{\sigma(1)} \dots i_{\sigma(k)}},$$

so the symbols $T_{[i_1 \dots i_k]}$ are antisymmetric with respect to interchanging pairs of indices, and one has $T_{[i_1 \dots i_k]} = T_{i_1 \dots i_k}$ if and only if $T_{i_1 \dots i_k}$ already has this property. Note that in this definition, there is no need to assume that $T_{i_1 \dots i_k}$ are the components of a well-defined tensor, but usefully, it may nonetheless happen that $T_{[i_1 \dots i_k]}$ *does* define a tensor. We saw an example of this already

[33]One easy way to see that the parity is well defined is by associating to each permutation $\sigma \in S_n$ the unique linear map $\mathbf{A}_\sigma : \mathbb{R}^n \to \mathbb{R}^n$ that permutes the standard basis vectors by σ. The matrix of $\mathbf{A}_\sigma$ is obtained from the identity matrix by permuting its columns, and $\det \mathbf{A}_\sigma = (-1)^{|\sigma|}$.

in Exercise 8.1, where the tensor $S \in \Gamma(T_2^0 M)$ defined from any 1-form $\lambda \in \Omega^1(M)$ can now be abbreviated in local coordinates by

$$S_{ij} = 2\,\partial_{[i}\lambda_{j]}$$

PROPOSITION 8.7. *For every smooth differential form $\omega \in \Omega^k(M)$, $k \geqslant 0$, there exists a unique $(k+1)$-form $d\omega \in \Omega^{k+1}(M)$ determined by the formula*

$$\tag{8.4} d\omega(X_0, \dots, X_k) = \sum_{i=0}^{k} (-1)^i \mathcal{L}_{X_i} \left[\omega(X_0, \dots, \hat{X}_i, \dots, X_k) \right] + \sum_{0 \leqslant i < j \leqslant k} (-1)^{i+j} \omega\big([X_i, X_j], X_0, \dots, \hat{X}_i, \dots, \hat{X}_j, \dots, X_k\big)$$

for $X_0, \dots, X_k \in \mathfrak{X}(M)$, where the hats over certain terms in sequences like "$X_0, \dots, \hat{X}_i, \dots, X_k$" mean that those terms do not appear in the sequence but every other term does. For any chart $(\mathcal{U}, x)$, the components of $d\omega$ in local coordinates over $\mathcal{U} \subset M$ are given by

$$(d\omega)_{i_0 \dots i_k} = (k+1)\partial_{[i_0}\omega_{i_1 \dots i_k]}.$$

PROOF. We claim first that both terms on the right hand side of (8.4) are antisymmetric functions of the vector fields $X_0, \dots, X_k$. In fact, the first term satisfies

$$\tag{8.5} \sum_{i=0}^{k} (-1)^i \mathcal{L}_{X_i} \left[\omega(X_0, \dots, \hat{X}_i, \dots, X_k) \right] = \frac{1}{k!} \sum_{\sigma \in S_{k+1}} (-1)^{|\sigma|} \mathcal{L}_{X_{\sigma(0)}} \left[\omega(X_{\sigma(1)}, \dots, X_{\sigma(k)}) \right],$$

where the right hand side is manifestly antisymmetric, and in this setting S_{k+1} means the group of permutations of the elements $\{0, \dots, k\}$. This can be seen by considering separately for each $i = 0, \dots, k$ the permutations σ with $\sigma(0) = i$, and then exploiting the antisymmetry of ω to place $X_{\sigma(1)}, \dots, X_{\sigma(k)}$ in a canonical order. A similar approach shows that the second term is a constant multiple of the antisymmetric expression $\sum_{\sigma \in S_{k+1}} (-1)^{|\sigma|} \omega\big([X_{\sigma(0)}, X_{\sigma(1)}], X_{\sigma(2)}, \dots, X_{\sigma(k)}\big)$.

We claim next that the right hand side of (8.4) is C^∞-linear in X_i for every $i = 0, \dots, k$. By antisymmetry, it suffices to prove this for $i = 0$, and the proof is then a straightforward computation based on Exercise 6.4. We can now conclude from Proposition 8.2 that $d\omega$ is a well-defined $(k+1)$-form. Finally, the coordinate formula for $d\omega$ follows from (8.5) since $[\partial_i, \partial_j] \equiv 0$ for all i, j. $\square$

DEFINITION 8.8. For a smooth k-form on ω, the $(k+1)$-form $d\omega$ defined in Proposition 8.7 is called the **exterior derivative** (*äußere Ableitung*) of ω.

EXAMPLE 8.9. For a 0-form $f \in C^\infty(M) = \Omega^0(M)$, the definition above makes $df \in \Omega^1(M)$ the usual differential of f.

For $k > 0$, the exterior derivative $d\omega$ of $\omega \in \Omega^k(M)$ does not contain *all* information about the first derivative of ω at each point, e.g. in local coordinates, the individual partial derivatives $\partial_j \omega_{i_1 \dots i_k}$ cannot be deduced from $(d\omega)_{i_0 \dots i_k}$, nor can ω be recovered from $d\omega$ up to addition of a constant. We will see more comprehensive (though non-canonical) ways of defining derivatives of ω when we discuss connections. The exterior derivative will be essential, however, due to the role it plays in Stokes' theorem, the n-dimensional generalization of the fundamental theorem of calculus.

8.3. Pullbacks and pushforwards. For a diffeomorphism $\psi : M \to N$, pushforwards and pullbacks of tensor fields can be defined in much the same way as for functions and vector fields in §5.2. Recalling the notation

$$\psi_* := T\psi : TM \to TN, \qquad \psi^* := (T\psi)^{-1} : TN \to TM,$$

we can dualize to define

$$\psi^* : T^*N \to T^*M, \qquad \psi_* : T^*M \to T^*N$$

by

$$(\psi^*\lambda)(X) := \lambda(\psi_* X), \qquad (\psi_*\lambda)(X) := \lambda(\psi^* X).$$

Every $S \in \Gamma(T^k_\ell M)$ with $k > 0$ or $\ell > 0$ then has a **pushforard** $\psi_* S \in \Gamma(T^k_\ell N)$ defined by

$$(\psi_* S)(\lambda^1, \ldots, \lambda^k, X_1, \ldots, X_\ell) := S(\psi^*\lambda^1, \ldots, \psi^*\lambda^k, \psi^* X_1, \ldots, \psi^* X_\ell),$$

and similarly, $S \in \Gamma(T^k_\ell N)$ has a **pullback** $\psi^* S \in \Gamma(T^k_\ell M)$ defined by

$$(\psi^* S)(\lambda^1, \ldots, \lambda^k, X_1, \ldots, X_\ell) := S(\psi_*\lambda^1, \ldots \psi_*\lambda^k, \psi_* X_1, \ldots, \psi_* X_\ell).$$

The reader should take a moment to check that under the canonical identification $\mathfrak{X}(M) = \Gamma(T^1_0 M)$, this definition of the pushforward and pullback for tensor fields of type $(1,0)$ matches what we defined in §5.2 for vector fields. The maps

$$\psi_* : \Gamma(T^k_\ell M) \to \Gamma(T^k_\ell N), \qquad \psi^* : \Gamma(T^k_\ell N) \to \Gamma(T^k_\ell M)$$

are vector space isomorphisms, and are inverse to each other. It is straightforward to show that if $\varphi : N \to Q$ is another diffeomorphism, the composition $\varphi \circ \psi : M \to Q$ satisfies

$$(\varphi \circ \psi)_* = \varphi_* \psi_*, \qquad (\varphi \circ \psi)^* = \psi^* \varphi^*. \tag{8.6}$$

Notice that the pushforward $\psi_* X = T\psi(X) \in TN$ of a tangent vector $X \in TM$ is defined without reference to the inverse ψ^{-1}, and can therefore also be defined when $\psi : M \to N$ is any smooth map, not necessarily a diffeomorphism. The same thus holds for the *pullback* of a fully covariant tensor field $S \in \Gamma(T^0_k N)$: the definition of $\psi^* S \in \Gamma(T^0_k M)$ as

$$\psi^* S(X_1, \ldots, X_\ell) = S(\psi_* X_1, \ldots, \psi_* X_\ell) = S(T\psi(X_1), \ldots, T\psi(X_\ell))$$

makes sense for any smooth map $\psi : M \to N$, though the resulting linear map $\psi^* : \Gamma(T^0_k N) \to \Gamma(T^0_k M)$ need not be invertible if ψ is not a diffeomorphism. This applies in particular for differential forms: they can always be pulled back via smooth maps.

Exercise 8.10. Assume $\psi : M \to N$ is a smooth map and $(\mathcal{U}, x)$ and $(\mathcal{V}, y)$ are charts on M and N respectively such that $\mathcal{U} \cap \psi^{-1}(\mathcal{V}) \neq \emptyset$. Abbreviating $\psi^i := y^i \circ \psi : \psi^{-1}(\mathcal{V}) \to \mathbb{R}$ for the component functions of ψ written in coordinates, show that the components of a k-form $\omega \in \Omega^k(N)$ in the coordinates $y^1, \ldots, y^n$ are related to those of its pullback $\psi^*\omega \in \Omega^k(M)$ in coordinates $x^1, \ldots, x^m$ by

$$(\psi^*\omega)_{i_1 \ldots i_k} = \frac{\partial \psi^{j_1}}{\partial x^{i_1}} \cdots \frac{\partial \psi^{j_k}}{\partial x^{i_k}} (\omega_{j_1 \ldots j_k} \circ \psi) \qquad \text{on } \mathcal{U} \cap \psi^{-1}(\mathcal{V}).$$

8.4. The Lie derivative of a tensor field. As with vector fields in §6.3, there is a natural way to differentiate any tensor field $S \in \Gamma(T^k_\ell M)$ with respect to a vector field $X \in \mathfrak{X}(M)$, giving the most general version of the **Lie derivative**

$$\mathcal{L}_X S := \left.\frac{d}{dt}(\varphi^t_X)^* X\right|_{t=0} \in \Gamma(T^k_\ell M).$$

This is well defined even if none of the flow maps φ^t_X are globally defined on M for $t \neq 0$, since for any point $p \in M$, φ^t_X is at least defined on a neighborhood of p for every t close enough to 0.

As with the Lie derivative of vector fields, one should keep in mind that for each $p \in M$, $(\mathcal{L}_X S)_p$ depends on more than just S and the value of X at p, due to the fact that pulling back via the flow requires differentiating it, and this derivative will also depend on the derivatives of X at p. The only exception is the case $k = \ell = 0$, in which S is just a function $f : M \to \mathbb{R}$ and $\mathcal{L}_X f = df(X)$ as before.

The Lie derivative has important applications to questions of *invariance*, e.g. if $\dim M = n$, we will see that one can use a differential form $\omega \in \Omega^n(M)$ to define a notion of *volume* for regions in M, and the condition $\mathcal{L}_X \omega \equiv 0$ will then characterize vector fields whose flows are volume preserving. We will need to develop the technology somewhat further before we can do nontrivial things with this, as it is typically quite difficult to compute $\mathcal{L}_X S$ directly from the definition, due to the fact that the flow of a vector field is typically not easy to write down. Let us mention however that there is a very user-friendly formula for the Lie derivative of a differential form:

THEOREM 8.11 (Cartan's formula). *For any $\omega \in \Omega^k(M)$ and $X \in \mathfrak{X}(M)$,*

$$\mathcal{L}_X \omega = d(\iota_X \omega) + \iota_X(d\omega),$$

*where the **interior product** $\iota_X \alpha \in \Omega^{q-1}(M)$ of a differential form $\alpha \in \Omega^q(M)$ with a vector field $X \in \mathfrak{X}(M)$ is defined by*

$$(\iota_X \alpha)(Y_1, \dots, Y_{q-1}) := \alpha(X, Y_1, \dots, Y_{q-1}).$$

We will prove this in Lecture 11, after we have discussed the algebra of differential forms in more detail.

9. The algebra of differential forms

Our goal for the next two lectures is to make sense of symbols like $\int_M f$ when M is a manifold. The naive hope would be that one could associate a real number $\int_M f \in \mathbb{R}$ to every (let's say continuous and compactly supported) function $f : M \to \mathbb{R}$, one that weights the values of f in proportion to the amount of volume covered. We will see that this notion does not make sense in general for real-valued *functions*, but if $\dim M = n$, it does make sense when f is replaced by a differential n-form.

9.1. Measure and volume on manifolds. The basic problem with defining $\int_M f$ for a function $f : M \to \mathbb{R}$ is that we have not specified any measure on M with which to define what "volume" means. Certain special classes of manifolds admit canonical measures, e.g. if M is a k-dimensional submanifold of $\mathbb{R}^n$, then one can derive a notion of "k-dimensional volume" on subsets of M from the Euclidean geometry of $\mathbb{R}^n$. But this measure on M will depend on the precise embedding $M \hookrightarrow \mathbb{R}^n$, e.g. the volume of any given region in M will change by a factor of L^k if we modify the embedding by multiplication with a scalar $L > 0$. And in any case, not all manifolds are presented as submanifolds of Euclidean space.

Another idea would be to use local coordinates, meaning that for any chart $(x, \mathcal{U})$ on M, the measure of a subset $\mathcal{O} \subset \mathcal{U}$ could be defined as the Lebesgue measure of $x(\mathcal{O}) \subset \mathbb{R}^n$. This definition, however, clealy depends on the choice of chart: according to the change of variables formula, the Lebesgue measure of $y(\mathcal{O}) \subset \mathbb{R}^n$ for another chart $(\mathcal{V}, y)$ with $\mathcal{O} \subset \mathcal{V}$ will be the Lebesgue integral of $|\det D(y \circ x^{-1})|$ over $x(\mathcal{O})$, and this integral is not typically the same as the measure of $x(\mathcal{O})$.

Let us drop the question of whether M carries a canonical measure (usually it doesn't), and ask instead how one might go about *choosing* a measure on M, i.e. what kinds of properties should a notion of n-dimensional volume on M have? Heuristically, one useful way to approach this question is by thinking of the tangent space T_pM at a point $p \in M$ is an "approximation" of a neighborhood of p in M, so if we can define volumes of regions in that neighborhood, we should also be able to define volumes of regions in the vector space T_pM. How does one define volume in an n-dimensional vector space? For example, given vectors $X_1, \dots, X_n \in T_pM$, consider the so-called **parallelepiped** spanned by $X_1, \dots, X_n$, meaning the set

$$P(X_1, \dots, X_n) := \left\{ t^i X_i \in T_pM \;\middle|\; t^1, \dots, t^n \in [0,1] \right\} \subset T_pM,$$

where as usual there is an implied summation in the expression $t^i X_i$. Suppose $\mu : T_pM \times \dots \times T_pM \to [0, \infty)$ is a function that associates to each n-tuple $(X_1, \dots, X_n)$ the n-dimensional volume of $P(X_1, \dots, X_n)$. What kind of function is μ? Basic geometric considerations dictate the following:

(1) If one of the vectors X_i is multiplied by a nonnegative constant, the volume scales by the same constant, i.e.

$$\mu(X_1, \dots, cX_i, \dots, X_n) = c\mu(X_1, \dots, X_i, \dots, X_n)$$

for $c \geqslant 0$.

(2) The volume is additive[34] with respect to each variable, i.e.

$$\mu(X_1, \dots, X_i + X_i', \dots, X_n) = \mu(X_1, \dots, X_i, \dots, X_n) + \mu(X_1, \dots, X_i', \dots, X_n).$$

An elementary geometric justification of this relation in the case $n = 2$ is shown in Figure 7. Using the letters A through E to denote the areas of the various regions in this picture, one has $\mu(X_1, X_2) = A + B$, $\mu(X_1', X_2) = C + D$, and $\mu(X_1 + X_1', X_2) = A + C + E = A + C + B + D = \mu(X_1, X_2) + \mu(X_1', X_2)$.

(3) If any two of the vectors $X_1, \dots, X_n$ match, then $P(X_1, \dots, X_n)$ is contained in an $(n-1)$-dimensional subspace and thus has zero n-dimensional volume, so

$$\mu(X_1, \dots, X_n) = 0 \qquad \text{whenever} \qquad X_i = X_j \text{ for some } i \neq j.$$

The first two properties suggest multilinearity, though μ itself cannot be multilinear since it only takes nonnegative values, and the scalar multiplication property only involves nonnegative scalars. On the other hand, a good way to find functions μ that satisfy these two properties is by choosing an actual multilinear function $\omega : T_pM \times \dots \times T_pM \to \mathbb{R}$ and setting

$$\mu(X_1, \dots, X_n) := |\omega(X_1, \dots, X_n)|.$$

The third property now imposes a serious restriction on ω:

PROPOSITION 9.1. *If V is a vector space and $\omega : V \times \dots \times V \to \mathbb{R}$ is an n-fold multilinear function that vanishes whenever two of its arguments are identical, then ω is alternating.*

PROOF. In the case $n = 2$, it suffices to choose any $v, w \in V$ and use multilinearity to observe

$$0 = \omega(v + w, v + w) = \omega(v, v) + \omega(w, w) + \omega(v, w) + \omega(w, v) = \omega(v, w) + \omega(w, v).$$

The general case works similarly. ☐

The upshot of this discussion is that a reasonable notion of volume for paralelepipeds in a tangent space T_pM can be defined by choosing an alternating n-fold multilinear form ω on T_pM and taking its absolute value. If the gaps in the discussion leading to this conclusion made you uncomfortable, one could alternatively derive it from a basic result in measure theory: every translation-invariant measure on $\mathbb{R}^n$ is a scalar $c \geqslant 0$ multiplied by the Lebesgue measure (see e.g. [Sal16, Chapter 2]). Moreover, the Lebesgue measure of the parallelepiped spanned by n vectors $\mathbf{v}_1, \dots, \mathbf{v}_n$ in $\mathbb{R}^n$ is given by $\left|\det \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{pmatrix}\right|$. As you learned in linear algebra, the

[34] Strictly speaking, some extra condition on the vectors $X_1, \dots, X_n$ is needed in order for the additivity property to hold, as not all possible configurations (even in the case $n = 2$) can be described by something like Figure 7. Since this is only meant to be a heuristic discussion, let's not worry about this for now.

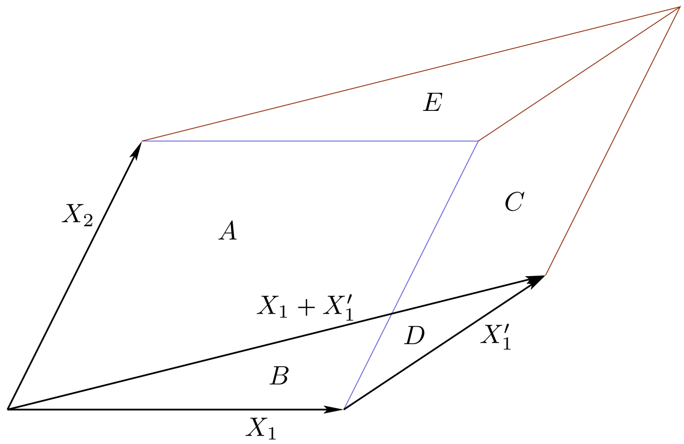

FIGURE 7. A geometric "proof" that volumes of parallelepipeds are determined by *multilinear* functions of their spanning vectors.

determinant of a matrix is an alternating multilinear function of its columns, thus we can now write $\mu = |\omega|$ where $\omega(\mathbf{v}_1, \dots, \mathbf{v}_n) := c \det \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{pmatrix}$ defines an alternating multilinear form.

Since everything in this course is smooth, it will also make sense to assume that for reasonable notions of volume on regions in M, the associated notions of volume on the tangent spaces T_pM depend smoothly on the point p. We can now say precisely what kind of geometric object defines a smoothly varying notion of volume on tangent spaces: it is a smooth n-form $\omega \in \Omega^n(M)$.

9.2. Exterior algebra. The previous section provided some motivation to believe that differential forms are the right objects with which to define integration on manifolds. Before we can fully unpack this idea, we need to develop the algebra of differential forms a bit further.

The tasks of this section are fundamentally algebraic, so there will be no manifolds, only an n-dimensional vector space V with basis $e_1, \dots, e_n \in V$. Let $e_*^1, \dots, e_*^n \in V^*$ denote the corresponding **dual basis**, determined by the condition

$$e_*^i(e_j) = \delta_j^i.$$

Recall from §7.2 that V_ℓ^k denotes the space of multilinear functions $V^* \times \dots \times V^* \times V \times \dots \times V \to \mathbb{R}$ that take k dual vectors in V^* and ℓ vectors in V as arguments; in particular, $V_1^0 = V^*$ and V_0^1 is the "double dual" $(V^*)^*$ of V, which is canonically isomorphic to V itself. The tensor product $\otimes : V_\ell^k \times V_s^r \to V_{\ell+s}^{k+r}$ can be defined in the same way as for tensor fields, and it is associative, so in particular, the tensor product of k dual vectors $\alpha^1, \dots, \alpha^k$ is a k-fold multilinear map $\alpha^1 \otimes \dots \otimes \alpha^k : V \times \dots \times V \to \mathbb{R}$ defined by

$$(\alpha^1 \otimes \dots \otimes \alpha^k)(v_1, \dots, v_k) = \alpha^1(v_1) \cdot \ldots \cdot \alpha^k(v_k).$$

The vector space of real-valued alternating k-fold multilinear maps on V is denoted by

$$\Lambda^k V^* := \left\{ \omega \in V_k^0 \;\middle|\; \omega(\dots, v, \dots, w, \dots) = -\omega(\dots, w, \dots, v, \dots) \text{ for all } v, w \in V \right\},$$

and we often refer to its elements as **alternating k-forms on V**. The antisymmetry condition is vacuous for $k \leqslant 1$, thus $\Lambda^0 V^* = \mathbb{R}$ and $\Lambda^1 V^* = V^*$. Using multilinearity as in Proposition 7.20, any $\omega \in \Lambda^k V^*$ for $k \geqslant 1$ can be written in terms of the basis $e_*^1, \dots, e_*^n \in V^*$ as

$$\omega = \omega_{i_1 \dots i_k} \, e_*^{i_1} \otimes \dots \otimes e_*^{i_k},$$

with unique coefficients

$$\omega_{i_1\ldots i_k} := \omega(e_{i_1}, \ldots, e_{i_k}) \in \mathbb{R}. \tag{9.1}$$

These coefficients are not all independent of each other: the antisymmetry of ω dictates that they satisfy

$$\omega_{i_1\ldots j\ldots \ell\ldots i_k} = -\omega_{i_1\ldots \ell\ldots j\ldots i_k},$$

i.e. there is a sign change whenever two distinct indices are interchanged, and $\omega_{i_1\ldots i_k}$ can only be nontrivial when all of its indices $i_1, \ldots, i_k \in \{1, \ldots, n\}$ have distinct values. It follows that $\omega_{i_1\ldots i_k}$ must always vanish if $k > n$, and otherwise, the number of distinct components that can be specified independently before the rest are determined is $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, hence

$$\dim \Lambda^k V^* = \begin{cases} \binom{n}{k} = \frac{n!}{k!(n-k)!} & \text{for } k \leqslant n, \\ 0 & \text{for } k > n. \end{cases}$$

Observe that while the case $k = 0$ was excluded from the discussion above, the formula $\dim \mathbb{R} = \dim \Lambda^0 V^* = \binom{n}{0} = 1$ is also correct in that case. The most interesting case is $k = n$: the elements of $\Lambda^n V^*$ are sometimes called **top-dimensional** forms, since n is the largest value of k for which $\Lambda^k V^*$ is a nontrivial space. The space is 1-dimensional in this case, due to the fact that all nontrivial components of $\omega \in \Lambda^n V^*$ are obtained by permuting the indices of $\omega_{1\ldots n}$. This elementary observation has nontrivial consequences that will be concretely useful to us, such as:

Proposition 9.2. *For any basis $v_1, \ldots, v_n \in V$ of a vector space V, every $\omega \in \Lambda^n V^*$ is uniquely determined by the number $\omega(v_1, \ldots, v_n) \in \mathbb{R}$; in particular, this number vanishes if and only if $\omega = 0$.* □

Example 9.3. The **determinant** $\det : \mathbb{R}^{n\times n} \to \mathbb{R}$ can be characterized by the property that $\mathbb{R}^n \times \ldots \times \mathbb{R}^n \to \mathbb{R} : (\mathbf{v}_1, \ldots, \mathbf{v}_1) \mapsto \det\begin{pmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_n\end{pmatrix}$ is the unique element of $\Lambda^n(\mathbb{R}^n)^*$ satisfying $\det\begin{pmatrix}\mathbf{e}_1 & \cdots & \mathbf{e}_n\end{pmatrix} = 1$ for the standard basis $\mathbf{e}_1, \ldots, \mathbf{e}_n \in \mathbb{R}^n$. Using the dual basis $\mathbf{e}^1_*, \ldots, \mathbf{e}^n_* \in (\mathbb{R}^n)^*$ to the standard basis, one can write down a concrete element of $\Lambda^n(\mathbb{R}^n)^*$ with this property in the form

$$\sum_{\sigma\in S_n} (-1)^{|\sigma|} \mathbf{e}^{\sigma(1)}_* \otimes \ldots \otimes \mathbf{e}^{\sigma(n)}_* \in \Lambda^n(\mathbb{R}^n)^*.$$

Plugging in the columns of a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ with entries $A^i{}_j$, an explicit formula for the determinant is thus given by

$$\det(\mathbf{A}) = \sum_{\sigma\in S_n} (-1)^{|\sigma|} A^{\sigma(1)}{}_1 \cdot \ldots \cdot A^{\sigma(n)}{}_n. \tag{9.2}$$

Proposition 9.2 now implies that every $\omega \in \Lambda^n(\mathbb{R}^n)^*$ can be written as

$$\omega(\mathbf{v}_1, \ldots, \mathbf{v}_n) = c \cdot \det\begin{pmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_n\end{pmatrix},$$

with a constant given by $c := \omega(\mathbf{e}_1, \ldots, \mathbf{e}_n) \in \mathbb{R}$.

For $k \geqslant 1$, a natural linear projection $\mathrm{Alt} : V^0_k \to V^0_k$ onto the subspace $\Lambda^k V^* \subset V^0_k$ is defined by

$$\mathrm{Alt}(\omega)(v_1, \ldots, v_k) := \frac{1}{k!} \sum_{\sigma\in S_k} (-1)^{|\sigma|} \omega(v_{\sigma(1)}, \ldots, v_{\sigma(k)}).$$

Indeed, one readily checks that $\mathrm{Alt}(\omega)$ is alternating for every $\omega \in V^0_k$, and ω itself is alternating if and only if $\mathrm{Alt}(\omega) = \omega$. If we write $\omega = \omega_{i_1\ldots i_k}\, e^{i_1}_* \otimes \ldots \otimes e^{i_k}_*$ for a general $\omega \in V^0_k$, applying

Alt changes the components via the antisymmetrization operation introduced in §8.2, which can be written succinctly as

$$\mathrm{Alt}(\omega)_{i_1 \dots i_k} = \omega_{[i_1 \dots i_k]}.$$

Note that for $k = 1$, Alt is simply the identity map $V^* \to V^*$. It will be a useful convention to extend this definition to $k = 0$ so that Alt is also the identity map on $V_0^0 = \mathbb{R}$.

We would now like to define a product operation on alternating forms that has geometric meaning. Let us regard each of the chosen basis 1-forms $e_*^i \in \Lambda^1 V^*$ as defining a notion of *length* (also known as "1-dimensional volume") for vectors in the 1-dimensional subspace $V_i := \mathbb{R} e_i \subset V$, so by this definition, the basis vectors $e_i \in V_i$ have unit length. The fact that each e_*^i vanishes on all the other subspaces $V_j \subset V$ for $j \neq i$ can be interpreted moreover as an "orthogonality" condition, so that we regard all the subspaces $V_1, \dots, V_n \subset V$ as orthogonal to each other. Geometrically, the paralelepiped in V spanned by $e_1, \dots, e_n$ should then have volume 1, and we would like to define the product n-form $e_*^1 \wedge \dots \wedge e_*^n \in \Lambda^n V^*$ to reproduce this notion of volume, i.e. it should satisfy

$$(e_*^1 \wedge \dots \wedge e_*^n)(e_1, \dots, e_n) = 1.$$

Since $\dim \Lambda^n V^* = 1$, there is exactly one element of $\Lambda^n V^*$ that satisfies this condition, and it is given by

$$e_*^1 \wedge \dots \wedge e_*^n = n! \, \mathrm{Alt}(e_*^1 \otimes \dots \otimes e_*^n) = \sum_{\sigma \in S_n} (-1)^{|\sigma|} e_*^{\sigma(1)} \otimes \dots \otimes e_*^{\sigma(n)}.$$

We take this observation as motivation for the general definition of the **wedge product**, which is contained in the theorem below. To state it properly, we define the vector space

$$\Lambda^* V^* := \bigoplus_{k=0}^{\infty} \Lambda^k V^*,$$

which is finite dimensional since $\Lambda^k V^* = \{0\}$ for $k > n$, hence $\Lambda^* V^*$ is equivalent to the finite product $\Lambda^0 V^* \times \dots \times \Lambda^n V^*$. We can regard each of the spaces $\Lambda^k V^*$ as subspaces of $\Lambda^* V^*$ in the obvious way. A nontrivial element $\alpha \in \Lambda^* V^*$ is said to be **homogeneous of degree** k if it belongs to the subspace $\Lambda^k V^* \subset \Lambda^* V^*$, in which case we also sometimes write its degree as

$$\deg(\alpha) = |\alpha| := k \quad \text{for} \quad \alpha \in \Lambda^k V^*.$$

One should keep in mind that not all elements of $\Lambda^* V^*$ are homogeneous, but this is of little importance in practice because every nontrivial element is a sum of a unique finite set of homogeneous elements of various degrees.

THEOREM 9.4. *There exists a unique bilinear map $\Lambda^* V^* \times \Lambda^* V^* \to \Lambda^* V^* : (\alpha, \beta) \mapsto \alpha \wedge \beta$ that satisfies*

$$c \wedge \alpha = \alpha \wedge c := c\alpha \qquad \text{for all } \alpha \in \Lambda^* V^* \text{ and } c \in \Lambda^0 V^* = \mathbb{R},$$

the associativity property

$$(\alpha \wedge \beta) \wedge \gamma = \alpha \wedge (\beta \wedge \gamma) \qquad \text{for all } \alpha, \beta, \gamma \in \Lambda^* V^*,$$

and

$$\alpha^1 \wedge \dots \wedge \alpha^k = \sum_{\sigma \in S_k} (-1)^{|\sigma|} \alpha^{\sigma(1)} \otimes \dots \otimes \alpha^{\sigma(k)} \qquad \text{for all } k \in \mathbb{N},\ \alpha^1, \dots, \alpha^k \in \Lambda^1 V^*, \tag{9.3}$$

where the k-fold product on the left hand side is defined by arbitrarily inserting parentheses to produce a sequence of binary operations. Moreover, the following conditions are satisfied:

(1) For any integers $k, \ell \geqslant 0$ and $\alpha \in \Lambda^k V^*$, $\beta \in \Lambda^\ell V^*$,

$$\alpha \wedge \beta = \frac{(k+\ell)!}{k!\ell!} \mathrm{Alt}(\alpha \otimes \beta) \in \Lambda^{k+\ell} V^*. \tag{9.4}$$

(2) *The wedge product is* graded commutative*, i.e. for homogeneous elements* $\alpha, \beta \in \Lambda^* V^*$,

$$\alpha \wedge \beta = (-1)^{|\alpha| \cdot |\beta|} \beta \wedge \alpha.$$

Before proving the theorem, we make the useful observation that if one defines k-fold wedge products of 1-forms via the right hand side of (9.3), then they can be used to turn any basis of V^* into a basis of $\Lambda^k V^*$:

PROPOSITION 9.5. *Given the basis* $e_1, \ldots, e_n \in V$ *and its dual basis* $e_*^1, \ldots, e_*^n \in V^*$*, every* $\omega \in \Lambda^k V^*$ *can be written as*

$$\omega = \sum_{i_1 < \ldots < i_k} \omega_{i_1 \ldots i_k} \, e_*^{i_1} \wedge \ldots \wedge e_*^{i_k} \tag{9.5}$$

for unique coefficients $\omega_{i_1 \ldots i_k} \in \mathbb{R}$*, which are given by*[35]

$$\omega_{i_1 \ldots i_k} = \omega(e_{i_1}, \ldots, e_{i_k}) \in \mathbb{R}.$$

PROOF. One uses the formula (9.3) to show that both sides of (9.5) match when evaluated on any tuple of basis vectors $(e_{i_1}, \ldots, e_{i_k})$ with $i_1 < \ldots < i_k$, and by antisymmetry, it follows that they also match when evaluated on *any* tuple of basis vectors. Multilinearity then implies that they match when evaluated on arbitrary k-tuples of vectors. □

REMARK 9.6. Proposition 9.5 is one of the few places where we are *not* using the Einstein summation convention. The reason is that the summation here does not cover all choices of tuples $i_1, \ldots, i_k \in \{1, \ldots, n\}$, as the summation convention would dictate, but rather only those for which the $i_1, \ldots, i_k$ are in strictly increasing order. Including all permutations of such tuples would produce extra terms that (due to the antisymmetry of both $\omega_{i_1 \ldots i_k}$ and $e_*^{i_1} \wedge \ldots \wedge e_*^{i_k}$) match the terms already present in the sum, i.e. exactly $k!$ copies of each term, plus some trivial terms for tuples in which some of the indices $i_1, \ldots, i_k$ match. This overcounting results in the formula

$$\omega = \frac{1}{k!} \omega_{i_1 \ldots i_k} \, e_*^{i_1} \wedge \ldots \wedge e_*^{i_k},$$

in which the coefficients are defined the same as before but the summation convention *is* in effect.

EXAMPLE 9.7. The following case of (9.3) is worth drawing special attention to: for two 1-forms $\alpha, \beta \in \Lambda^1 V^*$, $\alpha \wedge \beta \in \Lambda^2 V^*$ is given by $\alpha \wedge \beta = \alpha \otimes \beta - \beta \otimes \alpha$, thus

$$(\alpha \wedge \beta)(v, w) = \alpha(v)\beta(w) - \beta(v)\alpha(w).$$

One sees easily from this formula that the wedge product of 1-forms is *anticommutative*, i.e. it satisfies $\alpha \wedge \beta = -\beta \wedge \alpha$, and in particular, $\alpha \wedge \alpha = 0$.

PROOF OF THEOREM 9.4. By Proposition 9.5, every $\alpha \in \Lambda^k V^*$ and $\beta \in \Lambda^\ell V^*$ for $k, \ell \geqslant 1$ can be expressed as sums of wedge products of the basis 1-forms $e_*^1, \ldots, e_*^n \in V^*$ as determined by (9.3), so bilinearity and associativity together with (9.3) then uniquely determine $\alpha \wedge \beta \in \Lambda^{k+\ell} V^*$. The only problem with taking the resulting formula as a general *definition* of $\alpha \wedge \beta$ is that it may a priori depend on the choice of the basis $e_*^1, \ldots, e_*^n$. In order to dismiss this concern, we will show that this definition of $\alpha \wedge \beta$ also satisfies the formula (9.4), and observe that the right hand side of this expression is clearly independent of choices. By bilinearity and Proposition 9.5, it suffices to check that this is true when α and β are themselves products of the form

$$\alpha = e_*^{i_1} \wedge \ldots \wedge e_*^{i_k}, \qquad \beta = e_*^{j_1} \wedge \ldots \wedge e_*^{j_\ell}$$

[35]Notice that the coefficients in Proposition 9.5 are the same ones that appeared in (9.1).

for some choice of $i_1, \dots, i_k, j_1, \dots, j_\ell \in \{1, \dots, n\}$, and to show this, it is enough to evaluate both $\alpha \wedge \beta$ (as defined via (9.3)) and the right hand side of (9.4) on the ordered tuple of basis vectors

$$e_{a_1}, \dots, e_{a_k}, e_{b_1}, \dots, e_{b_\ell} \in V$$

for an arbitrary choice of $a_1, \dots, a_k, b_1, \dots, b_\ell \in \{1, \dots, n\}$. By antisymmetry, both clearly vanish unless the integers $a_1, \dots, a_k, b_1, \dots, b_\ell$ are all distinct, so let us assume this. Both will also vanish if any of those numbers are not contained in the set $\{i_1, \dots, i_k, j_1, \dots, j_\ell\}$, so assume this as well from now on, which implies that the numbers $i_1, \dots, i_k, j_1, \dots, j_\ell$ must also be all distinct, and thus

$$\{a_1, \dots, a_k, b_1, \dots, b_\ell\} = \{i_1, \dots, i_k, j_1, \dots, j_\ell\}.$$

Using antisymmetry, we can now apply a permutation and assume without loss of generality that the two ordered tuples are exactly the same, i.e. $a_m = i_m$ and $b_m = j_m$ for all m, so we need only evaluate both $\alpha \wedge \beta$ and $\frac{(k+\ell)!}{k!\ell!} \operatorname{Alt}(\alpha \otimes \beta)$ on the ordered tuple

$$(v_1, \dots, v_{k+\ell}) := e_{i_1}, \dots, e_{i_k}, e_{j_1}, \dots, e_{j_\ell}.$$

The result for $\alpha \wedge \beta$ is immediate from (9.3): only the trivial permutation produces a nontrivial term, and the answer is 1. Now consider

$$\frac{(k+\ell)!}{k!\ell!} \operatorname{Alt}(\alpha \otimes \beta)(v_1, \dots, v_{k+\ell}) = \frac{1}{k!\ell!} \sum_{\sigma \in S_{k+\ell}} (-1)^{|\sigma|} \alpha(v_{\sigma(1)}, \dots, v_{\sigma(k)}) \cdot \beta(v_{\sigma(k+1)}, \dots, v_{\sigma(k+\ell)}).$$

Since the sets $\{i_1, \dots, i_k\}$ and $\{j_1, \dots, j_\ell\}$ are disjoint, the only permutations that contribute non-trivially to the right hand side of this expression are those which preserve the subsets $\{1, \dots, k\}$ and $\{k+1, \dots, k+\ell\}$, and the sign of such a permutation is the product of the signs of the permutations of these two subsets, so the sum can be rewritten as

$$\frac{1}{k!\ell!} \sum_{(\sigma_1, \sigma_2) \in S_k \times S_\ell} (-1)^{|\sigma_1|} \alpha(e_{i_{\sigma_1(1)}}, \dots, e_{i_{\sigma_1(k)}}) \cdot (-1)^{|\sigma_2|} \beta(e_{j_{\sigma_2(1)}}, \dots, e_{j_{\sigma_2(\ell)}}).$$

Finally, observe that since α and β are both antisymmetric, every term in this last sum is identical, and there are exactly $k!\ell!$ of them, so we can restrict to the trivial permutation and simplify the expression to

$$\alpha(e_{i_1}, \dots, e_{i_k}) \cdot \beta(e_{j_1}, \dots, e_{j_\ell}) = 1,$$

since both terms in the product equal 1 by (9.3). This establishes the existence of the associative product $\wedge : \Lambda^* V^* \times \Lambda^* V^* \to \Lambda^* V^*$ and the formula (9.4). One still has to show that it also satisfies (9.3), i.e. not just for the basis 1-forms e^i_* but for arbitrary tuples of 1-forms $\alpha^1, \dots, \alpha^k \in \Lambda^1 V^*$. This can be derived from (9.4) by induction on k and a bit of combinatorics; we leave the details as an exercise.

To prove graded commutativity, it suffices again to consider the case where α and β are both products of 1-forms, and the relation then follows from the case $k = \ell = 1$ which was observed in Example 9.7. The key observation is that the number of flips required for permuting $i_1, \dots, i_k, j_1, \dots, j_\ell$ to $j_1, \dots, j_\ell, i_1, \dots, i_k$ is $k\ell$. □

The wedge product turns the vector space $\Lambda^* V^*$ into an algebra; it is called the **exterior algebra** (*äußere Algebra*) over V^*.[36]

[36] You may at this point be wondering what the "exterior algebra over V", presumably denoted by $\Lambda^* V$, might be. Since V is finite dimensional, the cheap way to define it is by identifying V with the dual space of V^*, so that homogeneous elements of $\Lambda^* V$ are antisymmetric multilinear maps $V^* \times \ldots \times V^* \to \mathbb{R}$. That is a correct definition, but not the most elegant formulation possible, and it also does not generalize to the case where V is infinite-dimensional since it may then fail to be isomorphic to its double dual. One can define $\Lambda^* V$ in terms of the abstract tensor product of vector spaces, and the details can be found in many standard algebra textbooks, but we will not need them here.

EXERCISE 9.8. Prove that a set of dual vectors $\alpha^1, \dots, \alpha^k \in V^*$ is linearly independent if and only if its wedge product $\alpha^1 \wedge \dots \wedge \alpha^k \in \Lambda^k V^*$ is nonzero.
Hint: Consider products of the form $\left(\sum_{i=1}^k c_i \alpha^i\right) \wedge \alpha^2 \wedge \dots \wedge \alpha^k$.

EXERCISE 9.9. Show that if $\alpha \in \Lambda^k V^*$ and $\beta \in \Lambda^\ell V^*$ are written in terms of the basis $e^1_*, \dots, e^n_* \in V^*$ as $\alpha = \alpha_{i_1 \dots i_k} e^{i_1}_* \otimes \dots \otimes e^{i_k}_*$ and $\beta = \beta_{i_1 \dots i_\ell} e^{i_1}_* \otimes \dots \otimes e^{i_\ell}_*$, then $\alpha \wedge \beta = (\alpha \wedge \beta)_{i_1 \dots i_{k+\ell}} e^{i_1}_* \otimes \dots \otimes e^{i_{k+\ell}}_*$ where

$$(\alpha \wedge \beta)_{i_1 \dots i_{k+\ell}} = \frac{(k+\ell)!}{k!\ell!} \alpha_{[i_1 \dots i_k} \beta_{i_{k+1} \dots i_{k+\ell}]}.$$

The following formula for top-dimensional forms will have many useful applications:

PROPOSITION 9.10. *Given a basis* $e_1, \dots, e_n \in V$ *with dual basis* $e^1_*, \dots, e^n_* \in V^*$, *we have*

$$\lambda^1 \wedge \dots \wedge \lambda^n = \det \begin{pmatrix} \lambda^1(e_1) & \cdots & \lambda^1(e_n) \\ \vdots & \ddots & \vdots \\ \lambda^n(e_1) & \cdots & \lambda^n(e_n) \end{pmatrix} e^1_* \wedge \dots \wedge e^n_*$$

for any $\lambda^1, \dots, \lambda^n \in \Lambda^1 V^*$.

PROOF. Use (9.3) to evaluate $(\lambda^1 \wedge \dots \wedge \lambda^n)(e_1, \dots, e_n)$, then plug in the formula (9.2) for the determinant. □

EXERCISE 9.11. Find a second proof of Proposition 9.10 using the following idea. Associate to each $\mathbf{v} = (v_1, \dots, v_n) \in \mathbb{R}^n$ the 1-form $\mathbf{v}_\flat := v_i e^i_* \in \Lambda^1 V^*$. What can you say about the multilinear function $\omega : \mathbb{R}^n \times \dots \times \mathbb{R}^n \to \mathbb{R}$ defined by $\omega(\mathbf{v}^1, \dots, \mathbf{v}^n) := (\mathbf{v}^1_\flat \wedge \dots \wedge \mathbf{v}^n_\flat)(e_1, \dots, e_n)$?

REMARK 9.12. The formula (9.4) for the product of $\alpha \in \Lambda^k V^*$ and $\beta \in \Lambda^\ell V^*$ can be written in more verbose form as

$$(\alpha \wedge \beta)(v_1, \dots, v_{k+\ell}) = \frac{1}{k!\ell!} \sum_{\sigma \in S_{k+\ell}} (-1)^{|\sigma|} \alpha(v_{\sigma(1)}, \dots, v_{\sigma(k)}) \cdot \beta(v_{\sigma(k+1)}, \dots, v_{\sigma(k+\ell)}). \tag{9.6}$$

The factor in front makes this formula a bit hard to memorize, but there is a combinatorial trick that makes it easier. Let

$$S_{k,\ell} \subset S_{k+\ell}$$

denote the subset consisting of permutations σ that satisfy

$$\sigma(1) < \dots < \sigma(k) \qquad \text{and} \qquad \sigma(k+1) < \dots < \sigma(k+\ell);$$

such permutations are sometimes called **shuffles**. They do not form a subgroup, but every permutation in $S_{k+\ell}$ is obtained from a unique shuffle by composing it with something in the subgroup $S_k \times S_\ell \subset S_{k+\ell}$ consisting of permutations that preserve the subsets $\{1, \dots, k\}$ and $\{k+1, \dots, k+\ell\}$. The key observation is that there are exactly $k!\ell!$ elements in this subgroup, and applying them has the effect of permuting the sets of vectors that are plugged into each of α and β in (9.6), while simultaneously changing the sign $(-1)^{|\sigma|}$ in a way that *cancels* the resulting change in the product of α and β. The result is that (9.6) contains $k!\ell!$ times as many terms as it actually needs: it is equivalent to the simpler formula

$$(\alpha \wedge \beta)(v_1, \dots, v_{k+\ell}) = \sum_{\sigma \in S_{k,\ell}} (-1)^{|\sigma|} \alpha(v_{\sigma(1)}, \dots, v_{\sigma(k)}) \cdot \beta(v_{\sigma(k+1)}, \dots, v_{\sigma(k+\ell)}), \tag{9.7}$$

in which no combinatorial factor is needed because the sum ranges only over shuffles.

9.3. The differential graded algebra of forms. Everything stated in the previous section implies a statement about differential forms on a manifold M, simply by replacing the vector space V with a tangent space T_pM and then letting $p \in M$ vary. In particular, a k-form $\omega \in \Omega^k(M)$ can now be understood as a function that associates to each $p \in M$ an element

$$\omega_p \in \Lambda^k T_p^* M := \Lambda^k(T_pM)^*.$$

It follows that if $\dim M = n$, then k-forms for $k > n$ are identically 0, hence the direct sum

$$\Omega^*(M) := \bigoplus_{k=0}^{\infty} \Omega^k(M)$$

has only finitely many nontrivial summands. (It is an infinite-dimensional space nonetheless, since each $\Omega^k(M)$ for $k = 0, \dots, n$ is infinite dimensional.) The wedge product of differential forms is now defined pointwise, i.e. given $\alpha \in \Omega^k(M)$ and $\beta \in \Omega^\ell(M)$, we define $\alpha \wedge \beta \in \Omega^{k+\ell}(M)$ by

$$(\alpha \wedge \beta)_p = \alpha_p \wedge \beta_p \in \Lambda^{k+\ell} T_p^* M.$$

The smoothness of $\alpha \wedge \beta$ by this definition will become clear momentarily when we write it down in local coordinates. Given a chart $(\mathcal{U}, x)$, the natural basis of T_pM to use at points $p \in \mathcal{U}$ is given by the coordinate vector fields $\partial_1, \dots, \partial_n$, and its dual basis consists of the coordinate differentials $dx^1, \dots, dx^n$. Any smooth k-form $\omega \in \Omega^k(M)$ can thus be written over $\mathcal{U}$ as

$$\begin{aligned} \omega &= \omega_{i_1 \dots i_k}\, dx^{i_1} \otimes \dots \otimes dx^{i_k} = \frac{1}{k!} \omega_{i_1 \dots i_k}\, dx^{i_1} \wedge \dots \wedge dx^{i_k} \\ &= \sum_{i_1 < \dots < i_k} \omega_{i_1 \dots i_k}\, dx^{i_1} \wedge \dots \wedge dx^{i_k}, \end{aligned} \tag{9.8}$$

where the Einstein summation convention is in effect for the first line but (in order to eliminate redundancy caused by antisymmetry) not for the second, and the smooth component functions are given by

$$\omega_{i_1 \dots i_k} = \omega(\partial_{i_1}, \dots, \partial_{i_k}) \in C^\infty(\mathcal{U}).$$

A coordinate formula for the wedge product can then be extracted from Exercise 9.9, namely

$$(\alpha \wedge \beta)_{i_1 \dots i_{k+\ell}} = \frac{(k+\ell)!}{k!\ell!} \alpha_{[i_1 \dots i_k} \beta_{i_{k+1} \dots i_{k+\ell}]},$$

so assuming that α and β have smooth components, the same is clearly true for $\alpha \wedge \beta$. Theorem 9.4 now carries over to the statement that $\wedge$ defines a bilinear map

$$\Omega^*(M) \times \Omega^*(M) \to \Omega^*(M) : (\alpha, \beta) \mapsto \alpha \wedge \beta$$

that is associative and graded commutative, where the latter again means that for homogeneous elements $\alpha \in \Omega^k(M)$ and $\beta \in \Omega^\ell(M)$, $\alpha \wedge \beta = \pm \beta \wedge \alpha$, with the minus sign appearing if and only if k and ℓ are both odd.

Example 9.13. Using Cartesian coordinates (x, y, z) on $\mathbb{R}^3$, the second line of (9.8) says that every $\omega \in \Omega^2(\mathbb{R}^3)$ has a unique presentation in the form

$$\omega = \omega_{xy}\, dx \wedge dy + \omega_{xz}\, dx \wedge dz + \omega_{yz}\, dy \wedge dz,$$

determined by three smooth functions $\omega_{xy}, \omega_{xz}, \omega_{yz} : \mathbb{R}^3 \to \mathbb{R}$.

Example 9.14. For $k = n$, the summation in the second line of (9.8) contains only one term. It follows that on an n-manifold M with smooth chart $(\mathcal{U}, x)$, every $\omega \in \Omega^n(M)$ can be written in local coordinates as

$$\omega = f\, dx^1 \wedge \dots \wedge dx^n \qquad \text{on } \mathcal{U},$$

where the real-valued function $f \in C^\infty(\mathcal{U})$ is given by $f = \omega(\partial_1, \dots, \partial_n)$.

Exercise 9.15. Beginners sometimes fixate on the antisymmetry of the wedge product for 1-forms and thus expect $\omega \wedge \omega = 0$ to hold always, but graded commutativity only implies this when ω has odd degree. Find a concrete example of a 2-form ω on $\mathbb{R}^4$ such that $\omega \wedge \omega \neq 0$.

We can now give a more practically useful characterization of the exterior derivative $d : \Omega^k(M) \to \Omega^{k+1}(M)$, which was defined in §8.2 via C^∞-linearity. A quick word about signs: you've already noticed that in the wedge product, a minus sign gets introduced whenever the order of two elements with odd degree is changed. One can use this same rule to remember the sign in the Leibniz rule below if one thinks of the operator d itself as an object with odd degree; it makes sense in fact to define its degree as 1, since that is the amount by which it raises the degree of any homogeneous element of $\Omega^*(M)$ fed into it.

Proposition 9.16. *The exterior derivative $d : \Omega^*(M) \to \Omega^*(M)$ is the unique linear map that satisfies the following conditions:*

(1) d is local, *meaning that for every form $\omega \in \Omega^*(M)$ and every $p \in M$, $(d\omega)_p \in \Lambda^* T_p^* M$ depends only on the restriction of ω to a neighborhood of p.*
(2) For each $f \in \Omega^0(M) = C^\infty(M)$, $df \in \Omega^1(M)$ is the differential of f.
(3) For any homogeneous elements $\alpha, \beta \in \Omega^*(M)$, d satisfies the "graded Leibniz rule"
$$d(\alpha \wedge \beta) = d\alpha \wedge \beta + (-1)^{|\alpha|} \alpha \wedge d\beta.$$
(4) $d \circ d = 0$.

Corollary 9.17. *For any chart $(\mathcal{U}, x)$ and any smooth function $f : \mathcal{U} \to \mathbb{R}$,*

$$d\left(f\, dx^{i_1} \wedge \ldots \wedge dx^{i_k}\right) = df \wedge dx^{i_1} \wedge \ldots \wedge dx^{i_k} = \partial_j f\, dx^j \wedge dx^{i_1} \wedge \ldots \wedge dx^{i_k} \qquad \text{on } \mathcal{U}. \tag{9.9}$$

□

Proof of Proposition 9.16. Let us start by ignoring the definition of $d : \Omega^*(M) \to \Omega^*(M)$ given in §8.2 and showing that a map satisfying the four properties stated above exists and is unique. The uniqueness follows from the observation that for any chart $(\mathcal{U}, x)$, every k-form on $\mathcal{U}$ is a sum of terms of the form $f\, dx^{i_1} \wedge \ldots \wedge dx^{i_k}$, and if d satisfies properties (2)–(4) then its action on this particular product is given by (9.9). To prove existence, suppose first that $M = \mathcal{U}$ is the domain of a global chart x, in which case the only possible definition of d satisfying the required properties is again via (9.9). It is immediate that d by this definition satisfies properties (1) and (2); let us verify that it also satisfies (3) and (4). To prove the graded Leibniz rule, we observe first that it is true for a pair of 0-forms $f, g \in \Omega^0(M) = C^\infty(M)$, as the product rule from first-year analysis implies
$$d(fg) = df \cdot g + f \cdot dg.$$
For the general case, bilinearity allows us to restrict attention to a pair $\alpha, \beta \in \Omega^*(\mathcal{U})$ of the form $\alpha = f\, dx^{i_1} \wedge \ldots \wedge dx^{i_k}$ and $\beta = g\, dx^{j_1} \wedge \ldots \wedge dx^{j_\ell}$. To make the notation more manageable, let us abbreviate $dx^I := dx^{i_1} \wedge \ldots \wedge dx^{i_k}$ and $dx^J := dx^{j_1} \wedge \ldots \wedge dx^{j_\ell}$; then
$$\begin{aligned} d(\alpha \wedge \beta) &= d\left(fg\, dx^I \wedge dx^J\right) = d(fg) \wedge dx^I \wedge dx^J = (df \cdot g + f \cdot dg) \wedge dx^I \wedge dx^J \\ &= \left(df \wedge dx^I\right) \wedge \left(g\, dx^J\right) + (-1)^k \left(f\, dx^I\right) \wedge \left(dg \wedge dx^J\right) = d\alpha \wedge \beta + (-1)^k \alpha \wedge d\beta, \end{aligned}$$
where the sign $(-1)^k$ arose when we changed the order of $dg \in \Omega^1(\mathcal{U})$ and $dx^I \in \Omega^k(\mathcal{U})$. To prove $d \circ d = 0$, we can similarly consider $\alpha = f\, dx^I$ and compute
$$d(d\alpha) = d(df \wedge dx^I) = d(\partial_j f\, dx^j \wedge dx^I) = d(\partial_j f) \wedge dx^j \wedge dx^I = \partial_k \partial_j f\, dx^k \wedge dx^j \wedge dx^I.$$
This last expression contains implied summations over both k and j, and we observe that while exchanging the roles of k and j leaves $\partial_k \partial_j f$ unchanged, it switches the sign of $dx^k \wedge dx^j$, so that every term in this sum is balanced by a cancelling term, and the sum if therefore 0.

Observe next that while our definition of $d : \Omega^*(\mathcal{U}) \to \Omega^*(\mathcal{U})$ above was expressed in terms of the specific coordinates $x^1, \dots, x^n$, the fact that it satisfies properties (1)–(4) implies that any other choice of coordinates would have given the same result, as it would also have given a definition satisfying properties (1)–(4). On a general manifold M, one can now define $d : \Omega^*(M) \to \Omega^*(M)$ on small neighborhoods using local coordinates and appeal to the fact that the definition is independent of coordinates, producing a global definition.

It remains only to prove that our definition of d via properties (1)–(4) matches the definition in §8.2. We will prove this by showing that (9.9) implies the same local coordinate formula that was derived in Proposition 8.7. Recall that in a local chart $(\mathcal{U}, x)$, an arbitrary k-form with components $\omega_{i_1 \dots i_k} = \omega(\partial_{i_1}, \dots, \partial_{i_k})$ can be written as

$$\omega = \sum_{i_1 < \dots < i_k} \omega_{i_1 \dots i_k} \, dx^{i_1} \wedge \dots \wedge dx^{i_k} = \frac{1}{k!} \omega_{i_1 \dots i_k} \, dx^{i_1} \wedge \dots \wedge dx^{i_k},$$

where the summation convention is in effect only in the second expression, in which the combinatorial factor accounts for the fact that each term in the implied summation appears in $k!$ identical copies arising from permutations of the indices $i_1, \dots, i_k$. The formula (9.9) then implies

$$d\omega = \frac{1}{k!} d\omega_{i_1 \dots i_k} \wedge dx^{i_1} \wedge \dots \wedge dx^{i_k} = \frac{1}{k!} \partial_{i_0} \omega_{i_1 \dots i_k} \, dx^{i_0} \wedge \dots \wedge dx^{i_k}.$$

In this last sum, nonzero contributions come only from terms in which the numbers $i_0, \dots, i_k \in \{1, \dots, n\}$ are all distinct, and if we write S_{k+1} for the group of bijections on $\{0, \dots, k\}$, each of these terms can be permuted by some $\sigma \in S_{k+1}$ to produce a product $dx^{i_0} \wedge \dots \wedge dx^{i_k}$ with $i_0 < \dots < i_k$, at the cost of applying the inverse permutation to the indices of $\partial_{i_0} \omega_{i_1 \dots i_k}$ and multiplying by the sign $(-1)^{|\sigma|}$. The expression therefore becomes

$$\begin{aligned} \frac{1}{k!} & \sum_{i_0 < \dots < i_k} \sum_{\sigma \in S_{k+1}} (-1)^{|\sigma|} \partial_{i_{\sigma(0)}} \omega_{i_{\sigma(1)} \dots i_{\sigma(k)}} \, dx^{i_0} \wedge \dots \wedge dx^{i_k} \\ & = \frac{(k+1)!}{k!} \sum_{i_0 < \dots < i_k} \partial_{[i_0} \omega_{i_1 \dots i_k]} \, dx^{i_0} \wedge \dots \wedge dx^{i_k} = (k+1) \sum_{i_0 < \dots < i_k} \partial_{[i_0} \omega_{i_1 \dots i_k]} \, dx^{i_0} \wedge \dots \wedge dx^{i_k}, \end{aligned}$$

which matches Proposition 8.7. □

The wedge product and exterior derivative make $\Omega^*(M)$ into an example of a (commutative) **differential graded algebra** (*graduierte Differentialalgebra*), or "DGA" for short. The inclusion of the word "graded" refers in the first place to the direct sum decomposition $\Omega^*(M) = \bigoplus_{k \geqslant 0} \Omega^k(M)$, but more importantly it refers to the sign appearing in the Leibniz rule of Proposition 9.16. A similar sign prevents $\Omega^*(M)$ from satisfying the commutativity relation $\alpha \wedge \beta = \beta \wedge \alpha$ in general, but the convention is nonetheless to call it a "commutative DGA" if it satisfies the graded commutativity relation $\alpha \wedge \beta = (-1)^{|\alpha||\beta|} \beta \wedge \alpha$.

Recall from §8.3 that pullbacks of differential forms can be defined for arbitrary smooth maps $\varphi : M \to N$, not just diffeomorphisms.

Proposition 9.18. *For any smooth map $\varphi : M \to N$:*

(1) $\varphi^*(\alpha \wedge \beta) = \varphi^*\alpha \wedge \varphi^*\beta$ for all $\alpha, \beta \in \Omega^*(N)$;
(2) $\varphi^*(d\omega) = d(\varphi^*\omega)$ for all $\omega \in \Omega^*(N)$.

Proof. The first statement follows directly from the definitions. For the second, we start with the case $\omega = f \in C^\infty(N) = \Omega^0(N)$ and use the chain rule: $\varphi^*(df) := df \circ T\varphi = d(f \circ \varphi) =: d(\varphi^* f)$. Since every differential form is locally a finite sum of wedge products of functions and differentials, the graded Leibniz rule then extends this result to all $\omega \in \Omega^k(N)$. □

10. Oriented manifolds and the integral

10.1. Change of variables. One of the messages of the previous lecture was that on an n-manifold M, one can use differential n-forms to define sensible notions of "n-dimensional volume" and thus measures, from which a notion of integration should emerge. Let's consider first how this might work when M is an open subset $\mathcal{U} \subset \mathbb{R}^n$ in Euclidean space.

There is a canonical choice of coordinates $x^1, \ldots, x^n$ on $\mathcal{U} \subset \mathbb{R}^n$, leading us naturally to consider the n-form $dx^1 \wedge \ldots \wedge dx^n \in \Omega^n(\mathcal{U})$. It has the desirable property that at every point $p \in \mathcal{U}$, if one feeds into it the standard basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$ of $\mathbb{R}^n = T_p\mathcal{U}$, the result (by (9.3)) is 1, which happens also to be the Lebesgue measure of the paralelepiped spanned by these vectors, i.e. the n-dimensional unit cube. It follows that if one interprets $dx^1 \wedge \ldots \wedge dx^n$ as a way of computing volumes on tangent spaces $T_p\mathcal{U} = \mathbb{R}^n$, the volume it computes is the *standard* notion of volume, i.e. the Lebesgue measure.

This observation motivates the following definition, which (in light of Example 9.14) tells us how to integrate an arbitrary compactly supported n-form on $\mathcal{U} \subset \mathbb{R}^n$.

Definition 10.1. For any integer $n \geqslant 1$, any compactly supported smooth function $f : \mathcal{U} \to \mathbb{R}$ on an open subset $\mathcal{U} \subset \mathbb{R}^n$ and any Lebesgue-measurable subset $A \subset \mathcal{U}$, the **integral** of the n-form $\omega := f\, dx^1 \wedge \ldots \wedge dx^n$ over A is defined to be the Lebesgue integral of f on A with respect to the standard Lebesgue measure m on $\mathbb{R}^n$, i.e.

$$\int_A \omega = \int_A f\, dx^1 \wedge \ldots \wedge dx^n := \int_A f\, dm \in \mathbb{R}.$$

Remark 10.2. If you prefer to think in terms of Riemann integrals rather than Lebesgue integrals, you are free to do so in Definition 10.1 at the cost of being slightly more restrictive about the subset $A \subset \mathcal{U}$, e.g. for almost all[37] applications it suffices to imagine that A is an open or closed subset. Nothing in our discussion of integration will depend in any serious way on the distinction between the Riemann and Lebesgue integrals. We will continue to use the language of Lebesgue integration because it seems the most natural.

Analysis conventions sometimes denote the Lebesgue measure on $\mathbb{R}^n$ more suggestively as "$dx^1 \ldots dx^n$", so that Definition 10.1 becomes the easy-to-remember formula

$$\int_A f\, dx^1 \wedge \ldots \wedge dx^n := \int_A f(x^1, \ldots, x^n)\, dx^1 \ldots dx^n.$$

Let's get a bit more ambitious now: suppose M is a more general n-manifold and $\omega \in \Omega^n(M)$ is a compactly supported top-dimensional differential form that happens to have its support contained in the domain $\mathcal{U} \subset M$ of some chart $(\mathcal{U}, x)$. In the corresponding local coordinates, ω can therefore also be written within $\mathcal{U}$ as $f\, dx^1 \wedge \ldots \wedge dx^n$ for a smooth compactly supported function $f : \mathcal{U} \to \mathbb{R}$. Expressing f as a function of the coordinates $x^1, \ldots, x^n$ on $\mathcal{U}$, it now seems natural to define

$$\int_A \omega := \int_{x(A)} f(x^1, \ldots, x^n)\, dx^1 \ldots dx^n \tag{10.1}$$

for any subset $A \subset \mathcal{U}$ such that $x(A) \subset x(\mathcal{U}) \subset \mathbb{R}^n$ is measurable, i.e. the function whose Lebesgue integral we are actually computing is $f \circ x^{-1} : x(\mathcal{U}) \to \mathbb{R}$. To see why this might be a sensible definition, write the standard Cartesian coordinates on $\mathbb{R}^n$ as $t^1, \ldots, t^n$ so as to distinguish them from the coordinates $x^1, \ldots, x^n$ on $\mathcal{U}$; regarding both sets of coordinates as functions on their respective domains, they are related by

$$t^i \circ x = x^i \quad \text{on } \mathcal{U}, \qquad i = 1, \ldots, n. \tag{10.2}$$

[37] no pun intended

Definition 10.1 now identifies the Lebesgue integral we just described with the integral of the n-form $(f \circ x^{-1})\, dt^1 \wedge \ldots \wedge dt^n$ over $x(A) \subset x(\mathcal{U}) \subset \mathbb{R}^n$. According to Proposition 9.18 and (10.2), the diffeomorphism $M \supset \mathcal{U} \xrightarrow{x} x(\mathcal{U}) \subset \mathbb{R}^n$ pulls this n-form back to $\mathcal{U}$ as

$$x^* \left((f \circ x^{-1})\, dt^1 \wedge \ldots \wedge dt^n \right) = f \cdot \left(x^* dt^1 \wedge \ldots \wedge x^* dt^n \right) = f \cdot \left(d(x^* t^1) \wedge \ldots \wedge d(x^* t^n) \right)$$
$$= f\, dx^1 \wedge \ldots \wedge dx^n = \omega,$$

so (10.1) follows from Definition 10.1 if we stipulate that the integral should satisfy

$$\int_A x^* \alpha = \int_{x(A)} \alpha$$

for all compactly supported n-forms α on $x(\mathcal{U}) \subset \mathbb{R}^n$. This identity is consistent with our intuition about pullbacks via diffeomorphisms: x^* gives a bijection allowing geometric data on $x(\mathcal{U}) \subset \mathbb{R}^n$ to be identified with geometric data on $\mathcal{U} \subset M$, and it would make sense for our definition of the integral to respect such identifications.

But there is still a crucial question to be answered: does our definition of $\int_A \omega$ as described above depend on the choice of chart $x : \mathcal{U} \to \mathbb{R}^n$?

Suppose $y : \mathcal{U} \to \mathbb{R}^n$ is a second chart defined on the same domain, so ω can also be written as $\omega = g\, dy^1 \wedge \ldots \wedge dy^n$ for some function $g : \mathcal{U} \to \mathbb{R}$, and $\int_A \omega$ according to this chart should be $\int_{y(A)} g \circ y^{-1}\, dm$, so we need to know whether this is the same as $\int_{x(A)} f \circ x^{-1}\, dm$. To clarify this, let us abbreviate $\psi := y \circ x^{-1} : x(\mathcal{U}) \to y(\mathcal{U})$ for the transition map relating x and y, and use Proposition 9.10 to write

$$dy^1 \wedge \ldots \wedge dy^n = \det \left(\frac{\partial y}{\partial x} \right) dx^1 \wedge \ldots \wedge dx^n \qquad \text{on } \mathcal{U},$$

where we abbreviate the matrix-valued function

$$\frac{\partial y}{\partial x} := \begin{pmatrix} \frac{\partial y^1}{\partial x^1} & \cdots & \frac{\partial y^1}{\partial x^n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y^n}{\partial x^1} & \cdots & \frac{\partial y^n}{\partial x^n} \end{pmatrix} : \mathcal{U} \to \mathbb{R}^{n \times n}.$$

The identity $f\, dx^1 \wedge \ldots \wedge dx^n = \omega = g\, dy^1 \wedge \ldots \wedge dy^n$ thus implies $f = g \cdot \det \left(\frac{\partial y}{\partial x} \right)$. At any point $p \in \mathcal{U}$, $\frac{\partial y}{\partial x}(p)$ is just the Jacobian matrix of the transition map ψ at $x(p)$, and this last identity thus implies

$$f \circ x^{-1} = \left(g \circ x^{-1} \right) \cdot \det D\psi.$$

If we now write $G := g \circ y^{-1}$, then $f \circ x^{-1}$ becomes $(G \circ \psi) \cdot \det D\psi$, and the identity we were hoping for becomes

$$\int_{y(A)} g \circ y^{-1}\, dm = \boxed{\int_{\psi(x(A))} G\, dm \stackrel{?}{=} \int_{x(A)} (G \circ \psi) \cdot \det D\psi\, dm} = \int_{x(A)} f \circ x^{-1}\, dm. \tag{10.3}$$

This should look familiar, as it is *almost* the classical change-of-variables formula, except for one detail: in the classical formula, the Jacobian determinant $\det(D\psi)$ is replaced by its absolute value. That is fine if $\det(D\psi)$ happens to be positive—we do of course know that it can never be 0, since ψ is a diffeomorphism and $D\psi(q) : \mathbb{R}^n \to \mathbb{R}^n$ is therefore an isomorphism for all $q \in x(\mathcal{U})$. But nothing in our discussion so far has ruled out the possibility that $\det(D\psi)$ may sometimes be *negative*, and there certainly do exist diffeomorphisms between regions in $\mathbb{R}^n$ that have negative Jacobian determinant, e.g. the reflection $(x, y) \mapsto (x, -y)$ on $\mathbb{R}^2$. The answer to the crucial question about (10.1) is therefore a resounding *sometimes*:

PROPOSITION 10.3. *In the setting of* (10.1)*, two charts defined on* $\mathcal{U}$ *give matching definitions of* $\int_A \omega$ *if the Jacobian determinant of their transition map is everywhere positive.* $\square$

10.2. Orientations. The upshot of our change-of-variables discussion is that integrating an n-form $\omega \in \Omega^n(M)$ by writing it in local coordinates as $\omega = f\, dx^1 \wedge \ldots \wedge dx^n$ and then integrating the function f in coordinates does not give a fully coordinate-invariant result, but it will *become* coordinate-invariant if for some reason we never have to worry about transition maps whose Jacobian determinant is negative. This is our first encounter in this course with the notion of *orientation*.

DEFINITION 10.4. Given open subsets $\mathcal{U}, \mathcal{V} \subset \mathbb{R}^n$ for $n \geqslant 1$, a diffeomorphism $\psi : \mathcal{U} \to \mathcal{V}$ is called **orientation preserving** (*orientierungserhaltend*) if the Jacobian matrix $D\psi(p) \in \mathrm{GL}(n, \mathbb{R})$ at every point $p \in \mathcal{U}$ has positive determinant. It is called **orientation reversing** (*orientierungsumkehrend*) if $\det D\psi(p) < 0$ for all p.

We will say more about the intuitive meaning of this definition in a moment, but for now, you may want to keep the following linear examples in mind:

(1) Every rotation $\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$ defines an orientation-preserving diffeomorphism $\mathbb{R}^2 \to \mathbb{R}^2$. More generally, every element of the special orthogonal group $\mathrm{SO}(n)$ (cf. Exercise 4.25) defines an orientation-preserving diffeomorphism $\mathbb{R}^n \to \mathbb{R}^n$.

(2) The reflection $(x, y) \mapsto (x, -y)$ is an orientation-reversing diffeomorphism $\mathbb{R}^2 \to \mathbb{R}^2$, and more generally, every element of $\mathrm{O}(n) \setminus \mathrm{SO}(n)$ defines an orientation-reserving diffeomorphism $\mathbb{R}^n \to \mathbb{R}^n$. In particular, this includes every linear transformation on $\mathbb{R}^n$ that is defined by reflecting across an $(n-1)$-dimensional subspace.

DEFINITION 10.5. A smooth atlas $\mathcal{A} = \{(\mathcal{U}_\alpha, x_\alpha)\}_{\alpha \in I}$ on a manifold M of dimension $n \geqslant 1$ is called **oriented** (*orientiert*) if all of its transition maps $x_\alpha \circ x_\beta^{-1}$ are orientation preserving. An **orientation** (*Orientierung*) of a manifold M with maximal smooth atlas $\mathcal{A}$ is a subset $\mathcal{A}^+ \subset \mathcal{A}$ that forms a maximal oriented atlas for M. A smooth manifold that has been equipped with an orientation $\mathcal{A}^+$ is called an **oriented manifold** (*orientierte Mannigfaltigkeit*), and the smooth charts in $\mathcal{A}^+$ are then called the **oriented charts**. A manifold is called **orientable** (*orientierbar*) if it admits an orientation.

One can argue as in Lemma 2.5 that given a smooth structure $\mathcal{A}$, every oriented atlas $\mathcal{A}^+ \subset \mathcal{A}$ has a unique extension to a maximal one and thus determines an orientation. In practice, we will see that there are usually more convenient ways to specify an orientation than by explicitly finding an oriented atlas, but here are a few examples where the latter can easily be done:

EXERCISE 10.6. Show that the atlas we defined on S^1 in Lecture 1 is oriented.

EXERCISE 10.7. Use the atlas from Exercise 1.7 to show that S^2 is orientable. (Depending on how you constructed the charts in that exercise, you might now have to modify them slightly for the sake of orientations.)

EXAMPLE 10.8. The manifold $\mathbb{R}^n$ carries a canonical global chart defined by the identity map, so this chart forms an oriented atlas and thus endows $\mathbb{R}^n$ with a canonical orientation.

EXAMPLE 10.9. If M has an oriented atlas $\mathcal{A}^+$ and $\mathcal{O} \subset M$ is an open subset, then the atlas $\mathcal{A}^+_{\mathcal{O}}$ on $\mathcal{O}$ constructed as in §2.4.2 is automatically also oriented, thus open subsets of oriented manifolds inherit natural orientations. In light of the previous example, this applies in particular to open subsets of $\mathbb{R}^n$.

EXERCISE 10.10. Show that if M and N are both orientable, then so is $M \times N$.

Exercise 10.11. Convince yourself that the atlases on the projective plane and Klein bottle described in §2.4.7 are not oriented. (This does not yet prove that these manifolds are not orientable, since one might imagine that there are other ways to construct an oriented atlas. But we will see below that this is impossible.)

Definition 10.12. For two oriented smooth manifolds M and N, a diffeomorphism $f : M \to N$ is called **orientation preserving** or **orientation reversing** if the map $y \circ f \circ x^{-1}$ is orientation preserving / reversing respectively for every choice of oriented smooth charts $(\mathcal{U}, x)$ on M and $(\mathcal{V}, y)$ on N.

Exercise 10.13. Show that for the orientations of S^1 and S^2 defined in Exercises 10.6 and 10.7, the antipodal map $S^n \to S^n : p \mapsto -p$ is orientation preserving for $n = 1$ but orientation reversing for $n = 2$.

Remark 10.14. In light of Definition 10.12 and the canonical orientations of $\mathbb{R}^n$ and open subsets specified by Examples 10.8 and 10.9, a smooth chart $(\mathcal{U}, x)$ on an oriented manifold M is an oriented chart if and only if the diffeomorphism $M \supset \mathcal{U} \xrightarrow{x} x(\mathcal{U}) \subset \mathbb{R}^n$ is orientation preserving.

Let's discuss next some useful alternative perspectives on the notion of orientation. We recall first the basic notion from topology of *connected components*. In topology one distinguishes between two slightly different notions of connectedness, but we will not need to worry about this distinction since for manifolds, they are equivalent.

Definition 10.15. A manifold M is **connected** (*zusammenhängend*) if for every pair of points $p, q \in M$, there exists a continuous path $\gamma : [0, 1] \to M$ with $\gamma(0) = p$ and $\gamma(1) = q$. The **connected components** (*Zusammenhangskomponenten*) of M are the maximal connected subsets.

It should be easy to convince yourself that each connected component of a manifold is both closed and open as a subset, hence it is also a manifold. In fact, if M has connected components $\{M_\alpha\}_{\alpha \in I}$, then there is a natural diffeomorphism $\coprod_{\alpha \in I} M_\alpha \cong M$.

Returning to the subject of orientations, consider a 2-dimensional subspace $P \subset \mathbb{R}^3$, i.e. a plane. One common way of characterizing what it should mean intuitively for P to be "oriented" in one way or the other is to decide which side of P is the "top" and which is the "bottom"; in other words, we draw a distinction between the two connected components of $\mathbb{R}^3 \backslash P$, labelling one component as "above" the plane and the other as "below" it. An equivalent way to say this is that one makes a choice of a unit vector $\mathbf{n} \in \mathbb{R}^3$ orthogonal to P, so that one can then decide to call the direction indicated by $\mathbf{n}$ "above" and the opposite direction "below". There are obviously two possible choices of the vector $\mathbf{n}$, and for an arbitrary plane $P \subset \mathbb{R}^3$, neither choice can be considered canonical.

Now, the case of a plane $P \subset \mathbb{R}^3$ is rather special since it is a submanifold of $\mathbb{R}^3$, and we do not want to have to assume all manifolds we consider are presented to us as submanifolds of Euclidean space. But actually, there is another way to characterize the choice of normal vector $\mathbf{n}$ in terms of vectors that are tangent to P. You may have learned it as the "right hand rule" when you first encountered vectors and the cross product in school: imagine positioning your right hand along the plane $P \subset \mathbb{R}^3$ so that your thumb points orthogonal to it in the direction of $\mathbf{n}$, but your other four fingers are tangent to P. Those four fingers will want to curl in a particular manner, defining a direction of rotation on the plane that one might choose to label "counterclockwise". (This is exactly what one does—at least in the northern hemisphere—when one visualizes the Earth "from above" and says that it rotates counterclockwise. In that situation, "from above" means that one chooses to view the Earth from a vantage point that is centered on the north pole; if one centered the picture on the south pole instead, the rotation would look clockwise! For the same reason, it is important to consistently use the *right* hand rather than the left hand when implementing the right hand rule, as switching hands would indicate a rotation in the other direction.)

The upshot of this heuristic discussion is this: our intuitive notion of what it means to orient a plane $P \subset \mathbb{R}^3$ is equivalent to making a choice of which direction of rotation on P should be labelled as counterclockwise instead of clockwise. This notion can be defined on *any* surface Σ by talking about rotations in the tangent spaces $T_p\Sigma$, and there is no longer any need to discuss normal vectors or assume an embedding $\Sigma \hookrightarrow \mathbb{R}^3$ is given. Moreover, we will see presently that instead of specifying a preferred direction of rotation in $T_p\Sigma$, it is equivalent to specify a preferred class of ordered bases.

Definition 10.16. For a vector space V of dimension $n \geqslant 1$, let

$$\mathcal{B}(V) \subset V^{\times n} := \underbrace{V \times \ldots \times V}_{n}$$

denote the set of all ordered n-tuples $(v_1, \ldots, v_n)$ that form bases of V.

Observe that $\mathcal{B}(V)$ is an open subset of $V^{\times n}$ since linear independence cannot be destroyed by small perturbations. In fact, after choosing any isomorphism $V \to \mathbb{R}^n$, the vectors in any tuple $(v_1, \ldots, v_n) \in \mathcal{B}(V)$ can be put together as columns of an n-by-n matrix, thus identifying $\mathcal{B}(V)$ with the general linear group $\mathrm{GL}(n, \mathbb{R})$, which is indeed an open subset of the space of matrices $\mathbb{R}^{n \times n}$.

Now consider the case $V = \mathbb{R}^2$. Given any $(v_1, v_2) \in \mathcal{B}(\mathbb{R}^2)$, moving from the direction of v_1 to that of v_2 requires a rotation of less than 180 degrees that is either counterclockwise or clockwise; for example, a counterclockwise rotation is required in order to move from the first standard basis vector $e_1 = (1, 0)$ to the second one $e_2 = (0, 1)$, but if we exchange their roles and order the standard basis as $(e_2, e_1) \in \mathcal{B}(\mathbb{R}^2)$, then getting from e_2 to e_1 requires a clockwise rotation. For a tangent space $T_p\Sigma$ to a surface Σ, the implication is that if one has chosen which rotations to call counterclockwise as opposed to clockwise, then one has also chosen a preferred class of ordered bases $(X_1, X_2) \in \mathcal{B}(T_p\Sigma)$, i.e. we call (X_1, X_2) a *positively oriented* basis of the rotation moving from X_1 to X_2 is counterclockwise, and *negatively oriented* if that rotation is clockwise. The following facts should now be apparent:

(1) If $(X_1, X_2) \in \mathcal{B}(T_p\Sigma)$ is positively oriented, then every $(X_1', X_2') \in \mathcal{B}(T_p\Sigma)$ that can be connected to (X_1, X_2) by a continuous path in $\mathcal{B}(T_p\Sigma)$ is also positively oriented. Conversely, any two choices of positively oriented basis are related to each other by a continuous deformation of ordered bases, meaning they are connected by a continuous path in $\mathcal{B}(T_p\Sigma)$. Both statements also apply of course to negatively oriented bases.
(2) Any choice of basis $(X_1, X_2) \in \mathcal{B}(T_p\Sigma)$ can be used to *define* the distinction between clockwise and counterclockwise rotation in $T_p\Sigma$: one simply chooses it so that (X_1, X_2) is a positively oriented basis.
(3) An ordered basis (X_1, X_2) is positively oriented if and only if (X_2, X_1) is negatively oriented.

There is a basic fact about $\mathrm{GL}(2, \mathbb{R})$ in the background of the first observation above: it has exactly two connected components, characterized by the conditions $\det(\mathbf{A}) > 0$ and $\det(\mathbf{A}) < 0$. This turns out to be true in every dimension:

Proposition 10.17. *For every $n \in \mathbb{N}$, the sets of* $\mathrm{GL}_+(n, \mathbb{R}) := \left\{\mathbf{A} \in \mathrm{GL}(n, \mathbb{R}) \;\middle|\; \det(\mathbf{A}) > 0\right\}$ *and* $\mathrm{GL}_-(n, \mathbb{R}) := \left\{\mathbf{A} \in \mathrm{GL}(n, \mathbb{R}) \;\middle|\; \det(\mathbf{A}) < 0\right\}$ *are both connected.*

Proof. Since $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$, it suffices to prove that $\mathrm{GL}_+(n, \mathbb{R})$ is connected. To start with, we use polar decomposition to reduce this to a statement about the special orthogonal group $\mathrm{SO}(n)$. Given $\mathbf{A} \in \mathrm{GL}_+(n, \mathbb{R})$, the matrix $\mathbf{A}^T\mathbf{A}$ is symmetric and positive definite, thus it is diagonalizable with only positive eigenvalues, and therefore admits a "square root"

$$\mathbf{P} := \sqrt{\mathbf{A}^T\mathbf{A}},$$

defined in the same orthogonal basis by taking the square roots of the eigenvalues. Clearly $\mathbf{P}$ is also symmetric and positive definite, and it is now straightforward to check that $\mathbf{R} := \mathbf{A}\mathbf{P}^{-1}$ satisfies $\mathbf{R}^T\mathbf{R} = \mathbb{1}$, i.e. it is orthogonal; moreover, $\mathbf{R} \in \mathrm{SO}(n)$ since $\mathbf{A}$ and $\mathbf{P}^{-1}$ each have positive determinant. Now choose a continuous path of symmetric positive-definite matrices $\{\mathbf{P}_t\}_{t\in[0,1]}$ such that $\mathbf{P}_1 = \mathbf{P}$ and $\mathbf{P}_0 = \mathbb{1}$; such a path can be found by fixing the orthonormal eigenbasis of $\mathbf{P}$ while deforming all its (positive!) eigenvalues to 1. The path $\mathbf{A}_t := \mathbf{R}\mathbf{P}_t$ then connects $\mathbf{A}_1 = \mathbf{A}$ to $\mathbf{A}_0 = \mathbf{R} \in \mathrm{SO}(n)$, so we will be done if we can show that $\mathrm{SO}(n)$ is connected.

We argue the latter by induction: the case $n = 1$ is already clear since $\mathrm{SO}(1) = \{1\}$. Assuming $\mathrm{SO}(n-1)$ is already known to be connected, suppose $\mathbf{A} \in \mathrm{SO}(n)$ is given. We claim that there exists a continuous path $\{\mathbf{A}_t \in \mathrm{SO}(n)\}_{t\in[0,1]}$ such that $\mathbf{A}_1 = \mathbf{A}$ and $\mathbf{A}_0$ is a matrix of the form

$$\mathbf{A}_0 = \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{B} \end{pmatrix}, \qquad \text{for some } \mathbf{B} \in \mathrm{SO}(n-1).$$

Observe that this claim implies the inductive step, as $\mathrm{SO}(n-1)$ is already known to be connected. To prove the claim, first choose any continuous path of unit vectors $v_1(t) \in \mathbb{R}^n$ such that $v_1(1)$ is the first column of $\mathbf{A}$ and $v_1(0)$ is the first standard basis vector $e_1 = (1, 0, \ldots, 0)$; this is possible since the unit sphere S^{n-1} is connected. For any $t_0 \in [0,1]$, one can complete $v_1(t_0)$ to an orthonormal basis $v_1(t_0), \ldots, v_n(t_0) \in \mathbb{R}^n$, and then find a connected neighborhood $J \subset [0,1]$ of t_0 such that the set of vectors $v_1(t), v_2(t_0), \ldots, v_n(t_0)$ remains linearly independent for every $t \in J$. Now define a continuous family of orthonormal bases $v_1(t), v_2(t), \ldots, v_n(t)$ for $t \in J$ by applying the Gram-Schmidt algorithm to $v_1(t), v_2(t_0), \ldots, v_n(t_0)$; regarding these as columns of a matrix, we have in this way constructed a continuous family of orthogonal matrices $\{\widehat{\mathbf{A}}_t \in \mathrm{O}(n)\}_{t\in J}$ whose first columns are $v_1(t)$. Their determinants depend continuously on t and are thus either $+1$ or -1 for all $t \in J$; in the latter case, we can replace $v_n(t)$ by $-v_n(t)$ in order to assume $\widehat{\mathbf{A}}_t \in \mathrm{SO}(n)$ without loss of generality. Since $[0,1]$ is compact, we can cover it with finitely many neighborhoods J as described above, and in this way construct a family of matrices $\{\widehat{\mathbf{A}}_t \in \mathrm{SO}(n)\}_{t\in[0,1]}$ that satisfy $\mathbf{A}_1 = \mathbf{A}$ and $\mathbf{A}_0 = \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{B} \end{pmatrix}$, and such that the first column of $\mathbf{A}_t$ depends continuously on t, while the other columns are continuous except at finitely many points $0 < t_1 < \ldots t_N < 1$, where there are jump discontinuities. At any of these points t_j, the two matrices

$$\widehat{\mathbf{A}}_{t_j}^- := \lim_{t\to t_j^-} \widehat{\mathbf{A}}_t, \qquad \widehat{\mathbf{A}}_{t_j}^+ := \lim_{t\to t_j^+} \widehat{\mathbf{A}}_t$$

may differ, but they have the same first column, namely $v_1(t_j)$. But expressing these matrices in any orthonormal basis that starts with $v_1(t_j)$ puts both of them in the form $\begin{pmatrix} 1 & 0 \\ 0 & \mathbf{B}_\pm \end{pmatrix}$ for some $\mathbf{B}_\pm \in \mathrm{SO}(n-1)$, and by the inductive hypothesis, there exists a continuous path in $\mathrm{SO}(n-1)$ from $\mathbf{B}_-$ to $\mathbf{B}_+$. In this way, we can insert extra intervals at each of the points t_j and fill in the discontinuities, then reparametrize the interval to construct the continuous family $\mathbf{A}_t$ in the claim. □

COROLLARY 10.18. *For any vector space V of dimension $n \geqslant 1$, the set of ordered bases $\mathcal{B}(V)$ has exactly two connected components.* □

REMARK 10.19. It is very important in this entire discussion that we are talking about *real* vector spaces, not complex. In particular, the analogous set of ordered complex bases on a complex vector space is *connected*, due to the fact that $\mathrm{GL}(n, \mathbb{C})$ is connected. A hint of this is provided by the fact that the determinant on $\mathrm{GL}(n, \mathbb{C})$ takes values in $\mathbb{C}\backslash\{0\}$, which is connected, unlike $\mathbb{R}\backslash\{0\}$. As a consequence, there is no meaningful notion of orientations for *complex* manifolds; actually, every complex manifold can also be regarded as a real manifold and is orientable as a real manifold, but the orientation is *canonically* determined by its complex structure. The reason for the latter is that if we identify $\mathbb{C}^n$ with $\mathbb{R}^{2n}$ via the correspondence $\mathbb{C}^n \ni \mathbf{x}+i\mathbf{y} \leftrightarrow (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^n = \mathbb{R}^{2n}$, then every complex-linear isomorphism $\mathbf{A} \in \mathrm{GL}(n, \mathbb{C})$ becomes an element of $\mathrm{GL}(2n, \mathbb{R})$ with *positive* determinant.

Exercise 10.20 (just for fun). Adapt the proof of Proposition 10.17 to prove that $\mathrm{GL}(n, \mathbb{C})$ is connected for every $n \in \mathbb{N}$.
Hint: $\mathrm{O}(1)$ is not connected, but $\mathrm{U}(1)$ is.

We can now give a general definition of orientations of vector spaces and relate it to the previously defined notion of oriented manifolds.

Definition 10.21. An **orientation** $\mathfrak{o}_V$ of an n-dimensional vector space V for $n \geqslant 1$ is a labelling of the two connected components of $\mathcal{B}(V)$ as $\mathcal{B}^+(V)$ and $\mathcal{B}^-(V)$, which are then said to consist of the **positively oriented** and **negatively oriented** bases respectively. An **oriented vector space** is a vector space that has been equipped with an orientation. A linear isomorphism $A : V \to W$ between two oriented vector spaces is called **orientation preserving** if for every positively-oriented basis $(v_1, \ldots, v_n)$ of V, $(Av_1, \ldots, Av_n)$ is a positively-oriented basis of W, and A is otherwise called **orientation reversing**.

Notice that unlike manifolds, vector spaces always admit orientations, and there are always exactly two possible choices of orientation.

Example 10.22. As a vector space, $\mathbb{R}^n$ carries a canonical orientation for which the standard basis is regarded as positively oriented.

Exercise 10.23. Show that for the vector space $\mathbb{R}^n$ with its canonical orientation, an invertible linear map $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^n$ is orientation preserving if and only if $\det(\mathbf{A}) > 0$.
Hint: The identity map $\mathbb{R}^n \to \mathbb{R}^n$ is clearly orientation preserving.

In light of Exercise 10.23, a diffeomorphism $\psi : \mathcal{U} \to \mathcal{V}$ between two open subsets $\mathcal{U}, \mathcal{V} \subset \mathbb{R}^n$ is orientation preserving as in Definition 10.4 if and only if its derivative at every point is an orientation-preserving isomorphism $\mathbb{R}^n \to \mathbb{R}^n$ in the sense of Definition 10.21. We only need one more notion before we can set up a precise correspondence between orientations of manifolds and of their tangent spaces:

Definition 10.24. Suppose M is an n-manifold with $n \geqslant 1$, P is a topological space, $\phi : P \to M$ is a continuous map, and we consider the family of tangent spaces $\{T_{\phi(s)}M\}_{s \in P}$ at points parametrized by the map ϕ. A **continuous family of orientations along** $\phi : P \to M$ is a family $\{\mathfrak{o}_s\}_{s \in P}$, where $\mathfrak{o}_s$ is an orientation of $T_{\phi(s)}M$ for each $s \in P$, such that for every $s_0 \in P$, there exists a neighborhood $\mathcal{O} \subset P$ of s_0 and a collection of continuous maps $X_1, \ldots, X_n : \mathcal{O} \to TM$ for which $(X_1(s), \ldots, X_n(s))$ is a positively-oriented basis of $T_{\phi(s)}M$ with respect to $\mathfrak{o}_s$ for each $s \in \mathcal{O}$. In the case $P = M$ with ϕ chosen to be the identity map, we will simply refer to this as a **continuous family of orientations of the tangent spaces** of M.

Proposition 10.25. *On smooth manifolds M of dimension $n \geqslant 1$, there is a natural bijective correspondence between orientations of M and continuous families of orientations of the tangent spaces of M, and it is uniquely determined by the condition that for any diffeomorphism $f : M \to N$ between two smooth oriented manifolds, f is orientation preserving if and only if the isomorphism $T_pf : T_pM \to T_{f(p)}N$ is orientation preserving for every $p \in M$. Equivalently, a chart $(\mathcal{U}, x)$ is* *oriented if and only if the corresponding basis of coordinate vector fields $(\partial_1, \dots, \partial_n)$ is positively oriented for every $p \in \mathcal{U}$.*

PROOF. If M is oriented, one defines the orientation of T_pM for any $p \in M$ such that for any oriented chart $(\mathcal{U}, x)$ with $p \in \mathcal{U}$, the isomorphism $d_px : T_pM \to \mathbb{R}^n$ is orientation preserving (for the canonical orientation of $\mathbb{R}^n$). This is equivalent to the condition stated above involving coordinate vector fields, and the definition is independent of the choice of oriented chart since if $(\mathcal{V}, y)$ is a different choice, then d_py is the composition of d_px with an isomorphism $\mathbb{R}^n \to \mathbb{R}^n$ (defined by differentiating a transition map) that is orientation preserving. Conversely, given a continuous family of orientations of the tangent spaces T_pM, one defines the corresponding orientation of M such that a chart $(\mathcal{U}, x)$ is oriented if and only if $d_px : T_px \to \mathbb{R}^n$ is orientation preserving for every $p \in \mathcal{U}$. We leave it as an exercise to check that these definitions satisfy all of the stated properties. □

The fact that the orientations of the tangent spaces T_pM vary *continuously* with p is crucial, and it provides the easiest means of proving statements about orientations in many concrete examples.

EXERCISE 10.26. For a smooth n-manifold M with $n \geqslant 1$, prove:

(1) If M is connected and orientable, then it admits exactly two choices of orientation.
(2) M is orientable if and only if for every continuous path $\gamma : [0, 1] \to M$ with $\gamma(0) = \gamma(1)$ and every continuous family of orientations $\{\mathfrak{o}_t\}_{t\in[0,1]}$ along γ, $\mathfrak{o}_0 = \mathfrak{o}_1$.

EXERCISE 10.27. Show that S^n is orientable for every $n \in \mathbb{N}$.
Hint: For every $p \in S^n$ and any basis $X_1, \dots, X_n$ of T_pS^n, $(X_1, \dots, X_n, p)$ forms a basis of $\mathbb{R}^{n+1}$. Use the fact that $\mathbb{R}^{n+1}$ is orientable.

EXERCISE 10.28. Use Exercise 10.26 to show that the projective plane $\mathbb{RP}^2$ and the Klein bottle are not orientable.

EXAMPLE 10.29. The physical universe is a 3-manifold, as you can plainly see by looking around you; from your local perspective it looks like $\mathbb{R}^3$, but since you cannot see the whole thing, it could in theory be diffeomorphic to any 3-manifold, even one that is not orientable. If indeed it is not orientable, then it is possible in theory for an astronaut to return from a long journey through space and find that what she used to call her right hand is now on the left side, and vice versa. She would not see it that way since her right and left eyes would also have been interchanged, but she would think that all writing now appears backwards, and the Earth (when viewed from the north pole) is now rotating clockwise. I am not aware of any law of physics that would rule out this scenario.

10.3. Definition of the integral. We are now in a position to define the integral of a compactly supported n-form on an oriented n-manifold for each $n \geqslant 1$. Denote the **support** (*Träger*) of a k-form $\omega \in \Omega^k(M)$ by

$$\operatorname{supp}(\omega) := \overline{\{p \in M \mid \omega_p \neq 0\}} \subset M,$$

and define the vector space

$$\Omega^k_c(M) := \{\omega \in \Omega^k(M) \mid \operatorname{supp}(\omega) \subset M \text{ is compact}\} \subset \Omega^k(M).$$

In the most interesting examples for our purposes, M will often be a compact manifold, in which case $\Omega^k_c(M) = \Omega^k(M)$. We will call a subset $A \subset M$ **measurable** if for every smooth chart $(\mathcal{U}, x)$ on M, the set $x(\mathcal{U} \cap A) \subset \mathbb{R}^n$ is Lebesgue measurable. The following theorem serves simultaneously as a definition.

THEOREM 10.30. *For $n \in \mathbb{N}$, one can uniquely associate to every smooth oriented n-manifold M and measurable subset $A \subset M$ a linear map*

$$\Omega_c^n(M) \to \mathbb{R} : \omega \mapsto \int_A \omega$$

such that the following conditions are satisfied:

(1) *If $\mathcal{U} \subset M$ is an open subset containing $\operatorname{supp}(\omega) \cap A$, then then $\int_{\mathcal{U} \cap A} \omega = \int_A \omega$.*
(2) *For $M = \mathcal{U} \subset \mathbb{R}^n$ an open subset of Euclidean space with its canonical orientation and the standard Cartesian coordinates $x^1, \ldots, x^n$,*

$$\int_A f\, dx^1 \wedge \ldots \wedge dx^n = \int_A f\, dm$$

for all smooth compactly supported functions $f : \mathcal{U} \to \mathbb{R}$, where the right hand side is the standard Lebesgue integral of f.
(3) *For any orientation-preserving diffeomorphism $\psi : M \to N$ between a pair of oriented n-manifolds,*

$$\int_A \psi^* \omega = \int_{\psi(A)} \omega$$

holds for all $\omega \in \Omega_c^n(N)$ and measurable subsets $A \subset M$.

To summarize, the integral on arbitrary oriented manifolds is uniquely determined by its definition on open subsets of $\mathbb{R}^n$ and the change-of-variables formula, which now appears as the condition that integrals are invariant under pullbacks via orientation-preserving diffeomorphisms. We will prove this in the next lecture, but it is already easy to explain the idea. For forms $\omega \in \Omega_c^n(M)$ with $\operatorname{supp}(\omega)$ contained in the domain of a single oriented chart $(\mathcal{U}, x)$, one can write

$$\omega = f\, dx^1 \wedge \ldots \wedge dx^n = x^* \left((f \circ x^{-1})\, dt^1 \wedge \ldots \wedge dt^n \right) \qquad \text{on } \mathcal{U}$$

in terms of the standard Cartesian coordinates $t^1, \ldots, t^n$ on $x(\mathcal{U}) \subset \mathbb{R}^n$ and a uniquely determined function $f : \mathcal{U} \to \mathbb{R}$. The three properties in the statement above then reproduce the definition of $\int_A \omega$ that we saw in §10.1, namely

$$\begin{aligned}\int_A \omega = \int_{\mathcal{U} \cap A} \omega = \int_{\mathcal{U} \cap A} x^*((f \circ x^{-1})\, dt^1 \wedge \ldots \wedge dt^n) = \int_{x(\mathcal{U} \cap A)} (f \circ x^{-1})\, dt^1 \wedge \ldots \wedge dt^n \\ = \int_{x(\mathcal{U} \cap A)} f \circ x^{-1}\, dm.\end{aligned}$$

The restriction to oriented charts guarantees moreover in light of Proposition 10.3 that this result does not depend on the choice of the chart $(\mathcal{U}, x)$, though it does depend on the orientation. Linearity will then determine $\int_A \omega$ uniquely for every $\omega \in \Omega_c^n(M)$ if we can be assured that every such form is a finite sum of forms that each have compact support in the domain of some oriented chart. This is true, but not completely obvious—it will require a brief digression on the topic of *partitions of unity*, which will have many further uses as we move forward.

11. Integration and volume

11.1. Existence of the integral. I owe you a proof of Theorem 10.30 on the existence and properties of the linear map $\Omega_c^n \to \mathbb{R} : \omega \mapsto \int_A \omega$ for all oriented n-manifolds M and measurable subsets $A \subset M$. The following will serve as a useful tool for "localizing" such constructions:

LEMMA 11.1. *Given a smooth manifold M, a compact subset $K \subset M$ and a finite collection of open sets $\{\mathcal{U}_\alpha\}_{\alpha \in I}$ that cover K, there exists a collection of smooth functions $\{\varphi_\alpha : M \to [0,1]\}_{\alpha \in I}$ satisfying the following two conditions:*

(1) *For each* $\alpha \in I$, φ_α *has compact support contained in* $\mathcal{U}_\alpha$*;*
(2) $\sum_{\alpha \in I} \varphi_\alpha \equiv 1$ *on* K.

Proof. For each $p \in K$, choose any $\alpha_p \in I$ such that $p \in \mathcal{U}_{\alpha_p}$, and choose also a smooth function $\psi_p : M \to [0,1]$ with compact support in $\mathcal{U}_{\alpha_p}$ such that $\psi_p > 0$ on some open neighborhood $\mathcal{V}_p \subset \mathcal{U}_{\alpha_p}$ of p. The sets $\{\mathcal{V}_p\}_{p \in K}$ then form an open cover of the compact set K and therefore admit a finite subcover, i.e. there is a finite subset $K_0 \subset K$ such that $K \subset \bigcup_{p \in K_0} \mathcal{V}_p$. Now for each $\alpha \in I$, define a smooth function $\psi_\alpha : M \to [0, \infty)$ by

$$\psi_\alpha := \sum_{\{p \in K_0 \mid \alpha_p = \alpha\}} \psi_p.$$

By construction, ψ_α has compact support in $\mathcal{U}_\alpha$, and for each $q \in K$, there exists $p \in K_0$ such that $q \in \mathcal{V}_p$ and thus $\psi_p(q) > 0$, implying $\psi_{\alpha_p}(q) > 0$. It follows that $\sum_{\alpha \in I} \psi_\alpha > 0$ everywhere on K, and therefore also on some neighborhood $\mathcal{V} \subset M$ of K. On the neighborhood $\mathcal{V}$, we define

$$\varphi_\alpha := \frac{\psi_\alpha}{\sum_{\beta \in I} \psi_\beta}, \qquad \text{for each } \alpha \in I,$$

so that each φ_α takes values in $[0,1]$ and $\sum_{\alpha \in I} \varphi_\alpha \equiv 1$ by construction. Now choose any smooth function $f : M \to [0,1]$ that equals 1 on K and has compact support in $\mathcal{V}$, modify each φ_α by multiplying it by f, and extend the modified function to the rest of M so that it vanishes outside of $\mathcal{V}$. $\square$

The collection of functions $\{\varphi_\alpha\}_{\alpha \in I}$ in this lemma is a special case of a general construction called a **partition of unity** subordinate to the cover $\{\mathcal{U}_\alpha\}_{\alpha \in I}$ (*eine der Überdeckung untergeordnete Zerlegung der Eins*). We will extend this notion later, when we discuss more general existence theorems for geometric structures such as Riemannian metrics.

Proof of Theorem 10.30. Given an oriented n-manifold M with measurable subset $A \subset M$ and $\omega \in \Omega^n_c(M)$, choose an open subset $M_0 \subset M$ that contains $\operatorname{supp}(\omega) \cap A$ but has compact closure $\overline{M}_0 \subset M$. By compactness, we can cover $\overline{M}_0$ with a finite collection of open sets $\{\mathcal{U}_\alpha \subset M\}_{\alpha \in I}$ that are domains of oriented charts $(\mathcal{U}_\alpha, x_\alpha)$, and Lemma 11.1 provides a partition of unity $\{\varphi_\alpha : M \to [0,1]\}_{\alpha \in I}$ such that

(i) φ_α has compact support contained in $\mathcal{U}_\alpha$ for each $\alpha \in I$;
(ii) $\sum_{\alpha \in I} \varphi_\alpha \equiv 1$ on M_0.

We can now write

$$\omega = \sum_{\alpha \in I} \varphi_\alpha \omega \qquad \text{on } M_0,$$

and observe that $\varphi_\alpha \omega \in \Omega^n_c(\mathcal{U}_\alpha)$, so if the integral satisfies the properties stated in the theorem, then

$$\int_A \omega = \int_{M_0 \cap A} \omega = \sum_{\alpha \in I} \int_{M_0 \cap A} \varphi_\alpha \omega = \sum_{\alpha \in I} \int_{\mathcal{U}_\alpha \cap A} \varphi_\alpha \omega = \sum_{\alpha \in I} \int_{x_\alpha(\mathcal{U}_\alpha \cap A)} f_\alpha \, dm, \tag{11.1}$$

where $f_\alpha : x_\alpha(\mathcal{U}_\alpha) \to \mathbb{R}$ is the unique function such that $\varphi_\alpha \omega = x_\alpha^*(f_\alpha \, dx^1 \wedge \ldots \wedge dx^n)$ on $\mathcal{U}_\alpha$. This specifies the integral uniquely.

We claim next that if $\int_A \omega \in \mathbb{R}$ is defined via (11.1), then the result is independent of all choices, namely the open subset $M_0 \subset M$ containing $\operatorname{supp}(\omega) \cap A$, the finite collection of oriented charts $\{(\mathcal{U}_\alpha, x_\alpha)\}_{\alpha \in I}$ and the functions $\{\varphi_\alpha\}_{\alpha \in I}$ satisfying (i) and (ii) above. Independence of the choice of charts follows from the discussion in §10.1, in particular Proposition 10.3. This is the step at which it is crucial that M comes with an orientation, so the transition maps that we feed into Proposition 10.3 are all orientation preserving. With this out of the way, suppose

$\{(\mathcal{V}_\beta, y_\beta)\}_{\beta\in J}$ is another finite collection of oriented charts and $\{\psi_\beta : M \to [0,1]\}_{\beta\in J}$ a collection of smooth functions that each have compact support in the corresponding subsets $\mathcal{V}_\beta$ and satisfy $\sum_{\beta\in J}\psi_\beta \equiv 1$ on some open set $M_1 \subset M$ containing $\operatorname{supp}(\omega)\cap A$. The open set $M_0 \cap M_1 \subset M$ then also contains $\operatorname{supp}(\omega)\cap A$, and is covered by the finite collection of open sets

$$\{\mathcal{U}_\alpha \cap \mathcal{V}_\beta\}_{(\alpha,\beta)\in I\times J},$$

with the functions $\{\varphi_\alpha\psi_\beta : M \to [0,1]\}_{(\alpha,\beta)\in I\times J}$ having compact support in $\mathcal{U}_\alpha\cap\mathcal{V}_\beta$ and satisfying $\sum_{(\alpha,\beta)\in I\times J}\varphi_\alpha\psi_\beta \equiv 1$ on $M_0\cap M_1$. Any oriented chart x_α defined on $\mathcal{U}_\alpha$ is also defined on $\mathcal{U}_\alpha\cap\mathcal{V}_\beta$ for each $\beta\in J$, so we can use it to compute $\int_{\mathcal{U}_\alpha\cap\mathcal{V}_\beta\cap A}\varphi_\alpha\psi_\beta\omega$ as a Lebesgue integral over $x_\alpha(\mathcal{U}_\alpha\cap A)\subset\mathbb{R}^n$ of a function with compact support in the region $x_\alpha(\mathcal{U}_\alpha\cap\mathcal{V}_\beta)$, and the additivity of the Legesgue integral then implies

$$\int_{\mathcal{U}_\alpha\cap A}\varphi_\alpha\omega = \sum_{\beta\in J}\int_{\mathcal{U}_\alpha\cap\mathcal{V}_\beta\cap A}\varphi_\alpha\psi_\beta\omega,$$

and therefore also

$$\sum_{\alpha\in I}\int_{\mathcal{U}_\alpha\cap A}\varphi_\alpha\omega = \sum_{(\alpha,\beta)\in I\times J}\int_{\mathcal{U}_\alpha\cap\mathcal{V}_\beta\cap A}\varphi_\alpha\psi_\beta\omega.$$

But if we carry out the same argument instead with the charts $(\mathcal{V}_\beta, y_\beta)$ and write $\psi_\beta\omega = \sum_{\alpha\in I}\varphi_\alpha\psi_\beta\omega$, we find that the right hand side is also equal to $\sum_{\beta\in J}\int_{\mathcal{V}_\beta\cap A}\psi_\beta\omega$, proving that the two definitions of $\int_A\omega$ obtained from these different partitions of unity match.

It remains to check that our general definition of $\int_A\omega$ satisfies the three properties stated in the theorem, but this is easy, so we will leave it as an exercise with the following hint: the freedom to choose any convenient collection of oriented charts makes the formula $\int_A\psi^*\omega = \int_{\psi(A)}\omega$ for orientation-preserving diffeomorphisms $\psi : M \to N$ virtually a tautology. □

11.2. Computational tools. The notion of integration defined in Theorem 10.30 has several useful properties that were not mentioned yet, some of which can be applied to make actual calculations considerably easier, e.g. it is rarely actually necessary in practice to choose a partition of unity. We begin with two properties whose proofs are easy exercises.

Exercise 11.2. Prove that for an oriented n-manifold M and $\omega\in\Omega^n_c(M)$, the following properties hold:

(1) If $A, B\subset M$ are two disjoint measurable subsets, then $\int_{A\cup B}\omega = \int_A\omega + \int_B\omega$.
(2) If $A\subset M$ has the property that $x(\mathcal{U}\cap A)\subset\mathbb{R}^n$ has Lebesgue measure zero[38] for all smooth charts $(\mathcal{U}, x)$, then $\int_A\omega = 0$.

One frequently occurring situation in simple examples is that the domain $A\subset M$ where we want to integrate lies almost entirely inside the domain of a single chart, where the word "almost" in this case carries its usual measure-theoretic meaning, i.e. "outside of a set of measure zero". In combination with the exercise above, the next result will then allow us to dispense entirely with partitions of unity and compute the integral in a single chart:

Proposition 11.3. *Suppose M is an oriented n-manifold and $(\mathcal{U}, x)$ is an oriented chart on M. Then for any measurable subset $A\subset\mathcal{U}$ and $\omega\in\Omega^n_c(M)$ taking the form $f\,dx^1\wedge\ldots\wedge dx^n$ in $\mathcal{U}$, the function $f\circ x^{-1}$ is Lebesgue integrable on $x(A)\subset\mathbb{R}^n$ and*

$$\int_A\omega = \int_{x(A)} f\circ x^{-1}\,dm.$$

[38] We say in this case that $A\subset M$ has **measure zero**. Note that it is not actually necessary to define a measure on M in order to define this notion.

PROOF. Let $K \subset M$ denote the closure of $\mathrm{supp}(\omega) \cap A \subset M$, and observe that this set is compact since it is a closed subset of $\mathrm{supp}(\omega)$, and it is also contained in the closure of $\mathcal{U}$ since $A \subset \mathcal{U}$. In particular, the set

$$\partial K := K \cap (M \backslash \mathcal{U})$$

is contained in the boundary of the closure of $\mathcal{U}$, and by assumption it is disjoint from A. Next choose a finite collection of oriented charts $\{(\mathcal{O}_\alpha, x_\alpha)\}_{\alpha \in I}$ such that

$$K \subset \mathcal{U} \cup \bigcup_{\alpha \in I} \mathcal{O}_\alpha,$$

and for each $N \in \mathbb{N}$ and $\alpha \in I$, let

$$\mathcal{O}_\alpha^N := \left\{ p \in \mathcal{O} \mid |x_\alpha(p) - x_\alpha(q)| < 1/N \text{ for some } q \in \partial K \cap \mathcal{O}_\alpha \right\}.$$

We observe the following:

(1) $K \subset \mathcal{U} \cup \bigcup_{\alpha \in I} \mathcal{O}_\alpha^N$ for every $N \in \mathbb{N}$.
(2) For each $\alpha \in I$, $\mathcal{O}_\alpha^1 \supset \mathcal{O}_\alpha^2 \supset \mathcal{O}_\alpha^3 \supset \ldots$, and, since $A \cap \partial K = \varnothing$,

$$A \cap \bigcap_{N \in \mathbb{N}} \mathcal{O}_\alpha^N = \varnothing. \tag{11.2}$$

For each $N \in \mathbb{N}$, we can choose a partition of unity consisting of functions $\varphi^N, \varphi_\alpha^N : M \to [0,1]$ for each $\alpha \in I$ with compact supports $\mathrm{supp}(\varphi^N) \subset \mathcal{U}$ and $\mathrm{supp}(\varphi_\alpha^N) \subset \mathcal{O}_\alpha^N$ such that $\varphi^N + \sum_{\alpha \in I} \varphi_\alpha^N \equiv 1$ on K. Since K contains $A \cap \mathrm{supp}(\omega)$, we then have

$$\int_A \omega = \int_A \varphi^N \omega + \sum_{\alpha \in I} \varphi_\alpha^N \omega$$

for every $N \in \mathbb{N}$. But for each $\alpha \in I$, (11.2) implies that the Lebesgue measure of $x_\alpha(\mathcal{O}_\alpha^N \cap A)$ converges to 0 as $N \to \infty$, thus

$$\lim_{N \to \infty} \int_A \varphi_\alpha^N \omega = 0,$$

from which follows

$$\int_A \varphi^N \omega \to \int_A \omega \qquad \text{as } N \to \infty.$$

Writing $\omega = x^*(f\, dx^1 \wedge \ldots \wedge dx^n)$ on $\mathcal{U}$ for a suitable function $f : x(\mathcal{U}) \to \mathbb{R}$, $\int_A \varphi^N \omega$ becomes the Lebesgue integral

$$\int_{x(A)} (\varphi^N \circ x^{-1}) f\, dm,$$

in which the integrand converges pointwise to f since each point in A is outside the support of all the φ_α^N for N sufficiently large. If you already believe that f is Lebesgue integrable on $x(A)$, then since $|(\varphi^N \circ x^{-1}) f| \leqslant |f|$, the dominated convergence theorem now implies that this integral converges to $\int_{x(A)} f\, dm$ as $N \to \infty$, and the latter is therefore $\int_A \omega$.

Here is a quick sketch of the proof that f really is Lebesgue integrable on $x(A)$: suppose ω is replaced by a *continuous* n-form $|\omega|$ on M that equals $-\omega$ at any point where ω evaluates negatively on some positive basis, but is otherwise identical to ω. In general $|\omega|$ will not be smooth—just as $|f|$ need not be smooth when f is a smooth function—but continuity is good enough for defining the integral $\int_A |\omega|$ as in Theorem 10.30. Changing ω to $|\omega|$ has the effect of replacing f with $|f|$ in the calculation above, and similarly in all other oriented charts. The same argument as above then proves

$$\int_{x(A)} (\varphi^N \circ x^{-1}) |f|\, dm \to \int_A |\omega| \qquad \text{as } N \to \infty.$$

Since φ^N equals 1 on subsets that exhaust all of A as $N \to \infty$, this implies a uniform upper bound for the integral of $|f|$ over arbitrary compact subsets of $x(A)$, and thus $\int_{x(A)} |f| \, dm < \infty$. □

EXERCISE 11.4. For every oriented n-manifold M with $n \geqslant 1$, there exists another oriented manifold $-M$ that is defined as the same manifold with the "reversed" orientation, meaning that one changes the orientation of every tangent space T_pM. Show that for every $\omega \in \Omega^n_c(M)$,

$$\int_{-M} \omega = -\int_M \omega.$$

Hint: If you fix the reflection map $r(t^1, t^2, \ldots, t^n) := (-t^1, t^2, \ldots, t^n)$ on $\mathbb{R}^n$ and take any oriented chart $(\mathcal{U}, x)$ on M, then $(\mathcal{U}, r \circ x)$ will be an oriented chart on $-M$.

REMARK 11.5. At long last, we can now clarify a notational issue that often bothers newcomers to integral calculus: what does $\int_b^a f(x) \, dx$ actually mean when $a < b$? It is traditional to define this as a synonym for $-\int_a^b f(x) \, dx := -\int_{[a,b]} f \, dm$ and regard it as a meaningless but useful convention, but now we can assign a deeper meaning to it: for the 1-manifold $M := (a, b) \subset \mathbb{R}$ with its canonical orientation and the 1-form $f \, dx \in \Omega^1_c(M)$ defined via the canonical coordinate x and a compactly supported[39] function $f : (a, b) \to \mathbb{R}$, the correct definition is

$$\int_b^a f(x) \, dx := \int_{-(a,b)} f \, dx,$$

where $-(a, b)$, denotes the manifold (a, b) with the opposite of its canonical orientation. This is consistent with the way that substitution is typically applied in calculations of 1-dimensional integrals: orientation-reversing diffeomorphisms are sometimes used for substitution, but they produce integrals over intervals with reversed orientation.

11.3. Volume forms. We now consider the first true geometric application of integration: how does one compute volumes of subsets in a manifold?

In an ordinary measure space X with measure μ, the volume of $A \subset X$ is simply $\int_A d\mu$. We have seen that in n-dimensional oriented manifolds, the role of measures is played by differential n-forms; however, not all of these define geometrically appropriate notions of volume. Indeed, a form $\omega \in \Omega^n(M)$ gives a way to define volumes of paralelepipeds in each tangent space T_pM, but it can happen that $\omega_p = 0$ at some point $p \in M$, implying that all regions in T_pM have volume zero, which is not very reasonable geometrically. The objects that we will refer to as "volume forms" specifically exclude this possibility:

DEFINITION 11.6. A **volume form** (*Volumenform*) on an n-manifold M is an n-form $\omega \in \Omega^n(M)$ such that $\omega_p \neq 0$ for all $p \in M$.

NOTATION. In these notes, we will usually denote volume forms by

$$d\mathrm{vol} \in \Omega^n(M),$$

or sometimes $d\mathrm{vol}_M$ if there are various manifolds in the picture and we want to specify which one $d\mathrm{vol}$ is defined on. The notation is slightly misleading since in many cases, our volume form will not actually be the exterior derivative of anything; nonetheless, the presence of the symbol "d" is consistent with the way that measures are often written in integrals, and that is the role that we intend for $d\mathrm{vol}$ to play.

[39]We are assuming compact support in (a, b) here because we have not yet defined manifolds *with boundary*, and thus cannot define an integral over the *closed* interval $[a, b]$. This will come in the next lecture, however.

Observe that since $\dim \Lambda^n T_p^* M = 1$ for every $p \in M$, $d\mathrm{vol} := \omega \in \Omega^n(M)$ is a volume form if and only if ω_p is a basis of $\Lambda^n T_p^* M$ for every p, and it follows in this case that any other n-form $\alpha \in \Omega^n(M)$ can be written as

$$\alpha = f\, d\mathrm{vol}$$

for a unique function $f \in C^\infty(M)$. In this situation, α is also a volume form if and only if the function f is nowhere zero.

Proposition 11.7. *Any volume form $d\mathrm{vol} \in \Omega^n(M)$ on a manifold M determines a unique orientation of M such that for each $p \in M$, an ordered basis $(X_1, \ldots, X_n) \in T_p M$ is positively oriented if and only if $d\mathrm{vol}(X_1, \ldots, X_n) > 0$.*

Proof. Assuming $d\mathrm{vol}_p \neq 0$, Proposition 9.2 implies that $d\mathrm{vol}(X_1, \ldots, X_n) \neq 0$ for every basis $X_1, \ldots, X_n$ of $T_p M$. It follows that $d\mathrm{vol}$ determines a continuous map $\mathcal{B}(T_p M) \to \mathbb{R} : (X_1, \ldots, X_n) \mapsto d\mathrm{vol}(X_1, \ldots, X_n)$ that is never zero, and since it clearly can take values of both signs, it must take positive values on one connected component of $\mathcal{B}(T_p M)$ and negative values on the other. Since its values also vary continuously with the point p, this distinction between the signs of $d\mathrm{vol}(X_1, \ldots, X_n)$ determines a continuous family of orientations of the tangent spaces $T_p M$. $\square$

If M is equipped with the orientation determined by a volume form $d\mathrm{vol}$ via Proposition 11.7, then it is common to write this condition as

$$d\mathrm{vol} > 0,$$

meaning literally that $d\mathrm{vol}(X_1, \ldots, X_n) > 0$ for every $p \in M$ and every *positively-oriented* basis $(X_1, \ldots, X_n)$ of $T_p M$, and $d\mathrm{vol}$ is in this case called a **positive volume form** on the oriented manifold M. Another n-form $\alpha = f\, d\mathrm{vol}$ is then also a positive volume form if and only if $f > 0$ everywhere. In particular, for any oriented chart $(\mathcal{U}, x)$, $dx^1 \wedge \ldots \wedge dx^n$ is a positive volume form on $\mathcal{U}$ since $(dx^1 \wedge \ldots \wedge dx^n)(\partial_1, \ldots, \partial_n) = 1$, thus a positive volume form $d\mathrm{vol} \in \Omega^n(M)$ always locally takes the form

$$d\mathrm{vol} = f\, dx^1 \wedge \ldots \wedge dx^n, \qquad f : \mathcal{U} \to (0, \infty). \tag{11.3}$$

If $(M, d\mathrm{vol})$ is an oriented manifold equipped with a positive volume form, the volume of a measurable subset $A \subset M$ is now defined simply as

$$\mathrm{Vol}(A) := \int_A d\mathrm{vol},$$

which is always nonnegative due to (11.3).

The definition of volume in M clearly depends on a choice of volume form, and for arbitrary manifolds there is generally no canonical choice—this reflects the fact that volumes of regions can appear very different when viewed in different coordinate systems. However, there are situations in which extra data determines a natural choice of volume form.

Suppose for instance that $M \subset \mathbb{R}^n$ is a k-dimensional submanifold of Euclidean space. Each tangent space $T_p M$ is then a k-dimensional linear subspace of $T_p \mathbb{R}^n = \mathbb{R}^n$, and can thus be assigned the standard Euclidean inner product $\langle\ ,\ \rangle$, which we can then use to define lengths of vectors in $T_p M$ and angles between them. In particular, this defines the notion of an *orthonormal* basis of $T_p M$. The paralelepiped spanned by an orthonormal basis of a k-dimensional subspace in $\mathbb{R}^n$ has the same dimensions as the k-dimensional unit cube, so its k-dimensional volume is 1, and it would therefore be natural to choose a volume form $d\mathrm{vol} \in \Omega^k(M)$ that evaluates to 1 on some orthonormal basis.

To bring this discussion into its most natural setting, recall that a **Riemannian metric** (*Riemannsche Metrik*) on a manifold M is a smooth type $(0, 2)$ tensor field $g \in \Gamma(T_2^0 M)$ such that $g_p : T_p M \times T_p M \to \mathbb{R}$ defines an inner product on $T_p M$ for every $p \in M$. The pair (M, g)

is in this case called a **Riemannian manifold** (*Riemannsche Mannigfaltigkeit*). The data of a Riemannian metric makes it possible to define norms of tangent vectors and angles between them, so in particular, every tangent space T_pM acquires a well-defined notion of orthonormality.

DEFINITION 11.8. On a Riemannian manifold (M, g), a volume form $d\mathrm{vol} \in \Omega^n(M)$ is said to be **compatible** with the metric g if for every $p \in M$ and every orthonormal basis $X_1, \dots, X_n \in T_pM$, $|d\mathrm{vol}(X_1, \dots, X_n)| = 1$.

Since $\dim \Lambda^n T_p^*M = 1$ for an n-manifold M, there are clearly at most two volume forms compatible with a given metric g at any given point $p \in M$. The following algebraic lemma guarantees that there are, in fact, exactly two, corresponding to the two possible orientations of T_pM.

LEMMA 11.9. *Suppose V is an n-dimensional oriented vector space equipped with an inner product $\langle\ ,\ \rangle$, $v_1, \dots, v_n \in V$ is a positively-oriented orthonormal basis and $v_*^1, \dots, v_*^n \in V^*$ denotes its dual basis. Then the top-dimensional form*

$$\omega := v_*^1 \wedge \dots \wedge v_*^n \in \Lambda^n V^*$$

satisfies $\omega(w_1, \dots, w_n) = 1$ for every positively-oriented orthonormal basis $w_1, \dots, w_n \in V$.

PROOF. By (9.3), it will suffice to establish that if $w_*^1, \dots, w_*^n \in V^*$ is the dual basis of another positively-oriented orthonormal basis $w_1, \dots, w_n \in V$, then

$$v_*^1 \wedge \dots \wedge v_*^n = w_*^1 \wedge \dots \wedge w_*^n.$$

By Proposition 9.10, the scaling factor relating these two n-forms is the determinant of the matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ with entries $A^i{}_j := w_*^i(v_j)$. Writing v_k as a linear combination of the w_i gives $v_k = w_*^i(v_k)w_i$, and orthonormality then implies

$$\begin{aligned}\delta_{k\ell} = \langle v_k, v_\ell\rangle = \langle w_*^i(v_k)w_i, w_*^j(v_\ell)w_j\rangle = w_*^i(v_k)w_*^j(v_\ell)\langle w_i, w_j\rangle = w_*^i(v_k)w_*^j(v_\ell)\delta_{ij} \\ = \sum_{i=1}^n w_*^i(v_k)w_*^i(v_\ell) = \sum_{i=1}^n A^i{}_k A^i{}_\ell,\end{aligned}$$

where in the second line we can no longer use the summation convention since the index to be summed does not appear in an upper-lower pair. This calculation implies that the rows of $\mathbf{A}$ form an orthonormal set, meaning $\mathbf{A} \in \mathrm{O}(n)$ and thus $\det(\mathbf{A}) = \pm 1$. Since both bases are also positively oriented, there exists a continuous family of orthonormal bases connecting one to the other, implying that there is also a continuous family of orthogonal matrices connecting $\mathbf{A}$ to $\mathbb{1}$, thus $\det(\mathbf{A}) = 1$. $\square$

COROLLARY 11.10. *Every oriented Riemannian n-manifold (M, g) admits a unique so-called **Riemannian volume form** $d\mathrm{vol} \in \Omega^n(M)$ that is positive and compatible with g.*

PROOF. The existence and uniqueness of $d\mathrm{vol}_p \in \Lambda^n T_p^*M$ for each $p \in M$ follows from Lemma 11.9, so it remains only to check that the n-form defined in this way is smooth. To see this, note that for any $p \in M$, one can find a neighborhood $\mathcal{U} \subset M$ of p and smooth vector fields $X_1, \dots, X_n \in \mathfrak{X}(\mathcal{U})$ that form a positively-oriented orthonormal basis at every point in $\mathcal{U}$; simply start e.g. with a basis of coordinate vector fields near p and then use the Gram-Schmidt process to make them orthonormal at each point. Now if $\lambda^1, \dots, \lambda^n \in \Omega^1(\mathcal{U})$ are defined so that $\lambda_q^1, \dots, \lambda_q^n \in T_q^*M$ is the dual basis to $X_1(q), \dots, X_n(q) \in T_qM$ for every $q \in \mathcal{U}$, then $\lambda^1 \wedge \dots \wedge \lambda^n$ is a smooth n-form on $\mathcal{U}$ that matches $d\mathrm{vol}$ according to Lemma 11.9. $\square$

EXAMPLE 11.11. On $\mathbb{R}^n$, there is a standard choice of Riemannian metric defined by assigning to each $T_p\mathbb{R}^n = \mathbb{R}^n$ the Euclidean inner product. This makes the standard coordinate vector fields $\partial_1, \ldots, \partial_n$ into a positively-oriented orthonormal basis at every point, and the unique positive volume form compatible with the standard metric is thus the so-called **standard volume form** $dx^1 \wedge \ldots \wedge dx^n$. The notion of volume defined by integrating it is of course just the Lebesgue measure.

EXERCISE 11.12. In local coordinates with respect to an oriented n-dimensional chart $(\mathcal{U}, x)$, a Riemannian metric $g \in \Gamma(T_2^0 M)$ is described in terms of its components $g_{ij} := g(\partial_i, \partial_j)$, so that vectors $X, Y \in T_pM$ at points $p \in \mathcal{U}$ satisfy $g(X,Y) = g_{ij}X^iY^j$. The goal of this exercise is to prove that the Riemannian volume form is then given by

$$\text{dvol} = \sqrt{\det \mathbf{g}}\, dx^1 \wedge \ldots \wedge dx^n \qquad \text{on } \mathcal{U}, \tag{11.4}$$

where $\mathbf{g} : \mathcal{U} \to \mathbb{R}^{n\times n}$ denotes the matrix-valued function whose ith row and jth column is g_{ij}. Note that this matrix necessarily has positive determinant since g is positive definite. Fix a point $p \in \mathcal{U}$ and a positively-oriented orthonormal basis $(X_1, \ldots, X_n)$ of T_pM, whose dual basis we will denote by $\lambda^1, \ldots, \lambda^n \in T_p^*M$. According to Lemma 11.9, $\text{dvol}_p = \lambda^1 \wedge \ldots \wedge \lambda^n$. Define matrices $\mathbf{X}, \boldsymbol{\lambda} \in \mathbb{R}^{n\times n}$ whose ith row and jth column are $dx^i(X_j)$ and $\lambda^i(\partial_j)$ respectively. By Proposition 9.10, $(\lambda^1 \wedge \ldots \wedge \lambda^n)(\partial_1, \ldots, \partial_n) = \det \boldsymbol{\lambda}$.

(1) Prove $\boldsymbol{\lambda} = \mathbf{X}^{-1}$.
(2) Prove $\mathbf{X}^T\mathbf{g}\mathbf{X} = \mathbb{1}$.
(3) Deduce (11.4).

Most people's favorite manifolds are submanifolds of Euclidean space—especially surfaces in $\mathbb{R}^3$. Generalizing this notion slightly, an $(n-1)$-dimensional submanifold M of an n-manifold N is called a **hypersurface** (*Hyperfläche*) in N. Any Riemannian metric g on N induces a Riemannian metric on any submanifold $M \subset N$, defined simply by restricting each of the inner products g_p on tangent spaces T_pN to the subspaces $T_pM \subset T_pN$. To put this another way, one can denote the inclusion map of M into N by $i : M \hookrightarrow N$ and observe that for every $p \in M$, $i_* : T_pM \hookrightarrow T_pN$ is the corresponding inclusion map of vector spaces, so the Riemannian metric induced by $g \in \Gamma(T_2^0 N)$ on M is the pullback $i^*g \in \Gamma(T_2^0 M)$. With this understood, we will show next that there is an easy way to derive from the compatible volume form on an oriented Riemannian manifold the compatible volume form on any oriented hypersurface.

DEFINITION 11.13. For an n-dimensional vector space V and an integer $k = 1, \ldots, n$, the **interior product** is the bilinear map

$$V \times \Lambda^k V^* \to \Lambda^{k-1}V^* : (v, \alpha) \mapsto \iota_v\alpha$$

defined by $\iota_v\alpha(w_1, \ldots, w_{k-1}) := \alpha(v, w_1, \ldots, w_{k-1})$. On a manifold M, the map

$$\mathfrak{X}(M) \times \Omega^k(M) \to \Omega^{k-1}(M) : (X, \omega) \mapsto \iota_X\omega$$

is defined similarly by $(\iota_X\omega)_p := \iota_{X(p)}\omega_p$ for all $p \in M$.

PROPOSITION 11.14. *Assume (N, g) is a Riemannian manifold, $M \subset N$ is a hypersurface with inclusion map $i : M \hookrightarrow N$, and $\nu : M \to TN$ is a continuous map*[40] *such that for every $p \in M$, $\nu(p) \in T_pN$ is a unit vector orthogonal to T_pM. (In this situation we call ν a* ***unit normal vector field*** *for M.) Then if $\text{dvol}_N \in \Omega^n(N)$ is a volume form on N compatible with g,*

$$\text{dvol}_M := (\iota_\nu \text{dvol}_N)|_{TM} \in \Omega^{n-1}(M)$$

[40] In fact it will follow from these assumptions that ν is also smooth, but one does not need to know that in advance.

is a volume form on M compatible with the induced metric i^*g.

PROOF. For any $p \in M$ and an orthonormal basis $X_1, \dots, X_{n-1}$ of T_pM, the n-tuple $\nu(p), X_1, \dots, X_{n-1}$ forms an orthonormal basis of T_pN, thus

$$|\iota_\nu d\mathrm{vol}_N(X_1, \dots, X_{n-1})| = |d\mathrm{vol}(\nu(p), X_1, \dots, X_{n-1})| = 1.$$

□

EXERCISE 11.15. Using Cartesian coordinates (x, y, z) on $\mathbb{R}^3$, let $\omega := x\, dy \wedge dz + y\, dz \wedge dx + z\, dx \wedge dy \in \Omega^2(\mathbb{R}^3)$, and let $i : S^2 \hookrightarrow \mathbb{R}^3$ denote the inclusion of the unit sphere.

(a) Show that $d\mathrm{vol}_{S^2} := i^*\omega \in \Omega^2(S^2)$ is a volume form compatible with the Riemannian metric on S^2 induced by the Euclidean inner product.
Hint: Pick a good vector field $X \in \mathfrak{X}(\mathbb{R}^3)$ with which to write ω as $\iota_X(dx \wedge dy \wedge dz)$.
(b) Show that in the spherical coordinates (θ, ϕ) of Exercise 1.7, $d\mathrm{vol}_{S^2} = \cos\phi\, d\theta \wedge d\phi$.
(c) On the open upper hemisphere $\mathcal{U}_+ := \{z > 0\} \subset S^2 \subset \mathbb{R}^3$, one can define a chart $(x, y) : \mathcal{U}_+ \to \mathbb{R}^2$ by restricting to $\mathcal{U}_+$ the usual Cartesian coordinates x and y, which are then related to the z-coordinate on this set by $z = \sqrt{1 - x^2 - y^2}$. Show that $d\mathrm{vol}_{S^2} = \frac{1}{z}\, dx \wedge dy$ on $\mathcal{U}_+$.
(d) Compute the surface area of $S^2 \subset \mathbb{R}^3$ in two ways: once using the formula for $d\mathrm{vol}_{S^2}$ in part (b), and once using part (c) instead. In both cases, the results of §11.2 will allow you to express the answer in terms of a single Lebesgue integral over a region in $\mathbb{R}^2$, and there will be no need for any partition of unity.

11.4. Densities. [41]

You may have wondered: what if M is non-orientable, but I still want to compute its volume? There are two problems in this situation: one is that according to Proposition 11.7, M cannot admit a volume form if it does not also admit an orientation, but there is also the more fundamental issue that the integral of an n-form over an n-manifold is not defined unless M comes with an orientation. Recall from §10.1: the trouble was that if $\omega = f\, dx^1 \wedge \dots \wedge dx^n = g\, dy^1 \wedge \dots \wedge dy^n$ for two different local coordinate systems $x, y : \mathcal{U} \to \mathbb{R}^n$ on the same region, then the Legesgue integrals $\int_{x(\mathcal{U}\cap A)} f \circ x^{-1}\, dm$ and $\int_{y(\mathcal{U}\cap A)} g \circ y^{-1}\, dm$ cannot generally be assumed to match unless the transition map $\psi := y \circ x^{-1} : x(\mathcal{U}) \to y(\mathcal{U})$ is orientation preserving. This problem is summarized by Equation (10.3), which resembles the classical change-of-variables formula, but does not match it exactly unless $\det(D\psi)$ is everywhere positive.

One way to circumvent this problem is to give up on intergrating the real-valued functions f and g and instead integrate their absolute values, so that (10.3) gives rise to the completely true statement

$$\int_{y(A)} \left|g \circ y^{-1}\right| dm = \int_{\psi(x(A))} |G|\, dm = \int_{x(A)} |(G \circ \psi)| \cdot |\det D\psi|\, dm = \int_{x(A)} \left|f \circ x^{-1}\right| dm,$$

in which we are again writing $G := g \circ y^{-1}$. The message of this calculation is that if we are willing to ignore the sign of an n-form and pay attention only to its magnitude, then we will no longer need to restrict ourselves to orientation-preserving transition maps.

DEFINITION 11.16. A (nonnegative) **density** on a smooth n-manifold M is a map

$$\mu : (TM)^{\oplus n} \to [0, \infty)$$

[41] The contents of §11.4 were not covered in the lecture and will not be referred to again in this course, at least not in any serious way. This section of the notes is provided only for your information.

whose restriction to $T_pM \times \ldots \times T_pM$ for each $p \in M$ takes the form $\mu_p(X_1, \ldots, X_n) = |\omega_p(X_1, \ldots, X_n)|$ for some $\omega_p \in \Lambda^n T_p^* M$. In a smooth chart $(\mathcal{U}, x)$, every density can thus be written in terms of the standard volume form $dx^1 \wedge \ldots \wedge dx^n \in \Omega^n(\mathcal{U})$ as

$$\mu = f \cdot \left|dx^1 \wedge \ldots \wedge dx^n\right|$$

for a unique function $f : \mathcal{U} \to [0, \infty)$. We call μ a **smooth density** if the function f defined in this way is smooth for all choices of smooth chart on M.

Remark 11.17. It is also possible to define densities with negative values (see e.g. [**Lee13a**]), but we will not need this. Our refusal to define negative densities means that the space

$$\mathscr{D}(M) := \{\text{smooth densities on } M\}$$

is not a vector space, but it does admit natural notions of addition and multiplication by nonnegative scalars.

The **support** of a density $\mu \in \mathscr{D}(M)$ is of course the closure of the set $\{p \in M \mid \mu_p \neq 0\} \subset M$, and we will denote

$$\mathscr{D}_c(M) := \left\{\mu \in \mathscr{D}(M) \;\middle|\; \mu \text{ has compact support}\right\}.$$

For smooth maps $\varphi : M \to N$, there is a natural **pullback** operation $\varphi^* : \mathscr{D}(N) \to \mathscr{D}(M)$ defined by

$$(\varphi^*\mu)(X_1, \ldots, X_n) := \mu(\varphi_* X_1, \ldots, \varphi_* X_n).$$

If we revise the discussion of §10.1 to work with densities instead of n-forms, then the key fact is that for any two charts x and y defined on the same domain $\mathcal{U}$, we have

$$\left|dy^1 \wedge \ldots \wedge dy^n\right| = \left|\det\left(\frac{\partial y}{\partial x}\right)\right| \cdot \left|dx^1 \wedge \ldots \wedge dx^n\right| \qquad \text{on } \mathcal{U},$$

thus if $\mu = f\,|dx^1 \wedge \ldots \wedge dx^n| = g\,|dy^1 \wedge \ldots \wedge dy^n|$ on this region, the nonnegative functions f and g are related by $f = g \cdot \left|\det\left(\frac{\partial y}{\partial x}\right)\right|$. The presence of the absolute value in this expression repairs our previous problem with orientations, and it now follows that the integrals $\int_{x(A)} f \circ x^{-1}\, dm$ and $\int_{y(A)} g \circ y^{-1}\, dm$ will *always* match, even if $y \circ x^{-1}$ is orientation reversing. The proof of Theorem 10.30 can now easily be adapted to establish the following:

Theorem 11.18. *For $n \in \mathbb{N}$, one can uniquely associate to every smooth n-manifold M and measurable subset $A \subset M$ a map*

$$\mathscr{D}_c(M) \to [0, \infty) : \mu \mapsto \int_A \mu$$

such that the following conditions are satisfied:

(1) $\int_A(\mu_1 + \mu_2) = \int_A \mu_1 + \int_A \mu_2$ for any $\mu_1, \mu_2 \in \mathscr{D}_c(M)$.
(2) If $\mathcal{U} \subset M$ is an open subset containing $\operatorname{supp}(\mu) \cap A$, then then $\int_{\mathcal{U} \cap A} \mu = \int_A \mu$.
(3) For $M = \mathcal{U} \subset \mathbb{R}^n$ an open subset of Euclidean space and the standard Cartesian coordinates $x^1, \ldots, x^n$,

$$\int_A f\,\left|dx^1 \wedge \ldots \wedge dx^n\right| = \int_A f\, dm$$

for all smooth compactly supported functions $f : \mathcal{U} \to [0, \infty)$, where the right hand side is the standard Lebesgue integral of f.

(4) For any diffeomorphism $\psi : M \to N$ between a pair of n-manifolds,

$$\int_A \psi^* \mu = \int_{\psi(A)} \mu$$

holds for all $\mu \in \mathscr{D}_c(N)$ and measurable subsets $A \subset M$.

□

The freedom in this theorem to allow non-orientable manifolds and diffeomorphisms that are not orientation preserving is paid for by the fact that integrals of nonnegative densities are always nonnegative, and thus tend to deliver less information than the *real*-valued integrals of differential forms. As mentioned in Remark 11.17 above, one can also allow densities with negative values and thus obtain negative integrals, but this does not add very much generality: it is tantamount to defining a measure μ via integrals of a positive density and then computing integrals $\int_A f\, d\mu$ of functions f that are also allowed to have negative values. Integration of densities is a somewhat less elegant and less useful construction on the whole than integration of forms; in particular, there are many more beautiful theorems involving the latter. Nonetheless, there are of course geometric situations in which an integral that is guaranteed to be nonnegative is exactly what one wants:

Definition 11.19. A **volume element** on a smooth n-manifold M is a density $d\mathrm{vol}$ such that $d\mathrm{vol}_p \neq 0$ for every $p \in M$. If M is equipped with a volume element $d\mathrm{vol}$, one defines the **volume** of measurable sets $A \subset M$ by

$$\mathrm{Vol}(A) := \int_A d\mathrm{vol} \geqslant 0.$$

We can now state a version of Corollary 11.10 that does not depend on orientability; its proof is an easy adaptation of arguments in the previous section.

Proposition 11.20. *Every Riemannian manifold (M, g) admits a unique volume element $d\mathrm{vol}$ such that for all $p \in M$ and every orthonormal basis $X_1, \ldots, X_n$ of T_pM, $d\mathrm{vol}(X_1, \ldots, X_n) = 1$.* □

We will not have any more occasions to talk about densities and volume elements in this course, but it is good to be aware that a theory of integration exists for non-orientable manifolds, even if it is less versatile and less powerful than the orientable case.

12. Stokes' theorem

It is finally time to tell you the true reason why the exterior derivative is important: it is "dual" in some sense to the operation of replacing a manifold by its boundary. First we will have to discuss what is meant by the *boundary* of a manifold, and we will have to be fairly careful with orientations if we want to get all the signs right.

12.1. A word about dimension zero. You may or may not have noticed that manifolds of dimension zero have been explicitly excluded from all discussion of orientations and integration so far. You probably didn't miss it, because in truth, integrals of 0-forms on 0-manifolds are not very interesting. But we have to define them now, because as soon as we start talking about manifolds with boundary, 0-manifolds will inevitably arise, namely as boundaries of 1-manifolds.

A 0-manifold M, you may recall, is simply a discrete set, and it can have at most countably many elements; it is compact if and only if it is finite. A 0-form on M is then an arbitrary function $f : M \to \mathbb{R}$. There is no need to worry about continuity or smoothness since M is discrete, and the support of f is just the set of all points p where $f(p) \neq 0$, so $f : M \to \mathbb{R}$ has compact support if and only if it is zero outside of a finite set.

Since there is no such thing as a "basis" of a 0-dimensional vector space and no meaningful sense in which one can say that a (the) map $\mathbb{R}^0 \to \mathbb{R}^0$ preserves or reverses orientation, the entire discussion of orientations in §10.2 is useless for $n = 0$. What we will use instead looks terribly naive at first glance, but we will see that it works:

DEFINITION 12.1. An **orientation** of a 0-manifold M is a function $\varepsilon : M \to \{1, -1\}$, i.e. it a assigns to each point of M a label as either "positive" or "negative". A bijection $\varphi : M \to N$ between two oriented 0-manifolds is **orientation preserving** if it maps all positive points to positive points and all negative points to negative points, and it is **orientation reversing** if it exchanges the sets of positive and negative points.

DEFINITION 12.2. For M a 0-manifold with orientation $\varepsilon : M \to \{1, -1\}$ and $f \in \Omega^0_c(M)$, the integral of f on a subset $A \subset M$ is defined by

$$\int_A f := \sum_{p \in A} \varepsilon(p) f(p),$$

where the sum is necessarily finite since f has compact support.

The only other thing worth saying for now about this definition is that it trivially satisfies the usual change-of-variables formula

$$\int_A \varphi^* f = \int_{\varphi(A)} f, \qquad f \in \Omega^0_c(N)$$

whenever $\varphi : M \to N$ is an orientation-preserving bijection of oriented 0-manifolds.

12.2. Manifolds with boundary. The definitions from Lectures 1 and 2 need to be generalized if we want to accommodate examples like the unit n**-disk**

$$\mathbb{D}^n := \left\{ x \in \mathbb{R}^n \;\middle|\; |x| \leqslant 1 \right\},$$

whose interior is accurately described as a smooth n-manifold, but there are no n-dimensional charts (by our current definition) describing neighborhoods in $\mathbb{D}^n$ of points on the boundary

$$\partial\mathbb{D}^n := S^{n-1} \subset \mathbb{D}^n.$$

An even simpler example is the **half-plane**

$$\mathbb{H}^n := (-\infty, 0] \times \mathbb{R}^{n-1} \subset \mathbb{R}^n,$$

whose boundary is the linear subspace

$$\partial\mathbb{H}^n := \{0\} \times \mathbb{R}^{n-1} \subset \mathbb{R}^n.$$

Just as subspaces of this form serve as local models of submanifolds as seen through slice charts, the half-plane will serve as our local model for a manifold with boundary.

DEFINITION 12.3. An n-dimensional **boundary chart** $(\mathcal{U}, x)$ on a set M consists of a subset $\mathcal{U} \subset M$ and an injective map $x : \mathcal{U} \hookrightarrow \mathbb{H}^n$ whose image $x(\mathcal{U}) \subset \mathbb{H}^n$ is an open set.[42]

The only difference between this and Definition 1.4 is the replacement of $\mathbb{R}^n$ by the half-space $\mathbb{H}^n$. A boundary chart $(\mathcal{U}, x)$ will sometimes also be a chart according to our original definition, because an open subset $x(\mathcal{U}) \subset \mathbb{H}^n$ might also be an open subset of $\mathbb{R}^n$; indeed, it will be so if $x(\mathcal{U}) \cap \partial\mathbb{H}^n = \varnothing$. For this reason, any set that is covered by charts can equally well be covered by boundary charts: one need only modify each chart $(\mathcal{U}, x)$ by a translation so that its

[42]One finds a few variations on this definition in the literature, in which the half-space $\mathbb{H}^n = (-\infty, 0] \times \mathbb{R}^{n-1}$ gets replaced by different half-spaces such as $[0, \infty) \times \mathbb{R}^{n-1}$ or $\mathbb{R}^{n-1} \times [0, \infty)$. This detail makes no meaningful difference for the definition of a smooth manifold with boundary, but it starts to matter as soon as one has to think about orientations. The definition in the form we've given here leads to the simplest possible definition of boundary orientations, and a relatively straightforward proof of Stokes' theorem.

image lies in the interior of the half-plane, or if this is impossible because $x(\mathcal{U})$ is unbounded in the x^1-direction, first break it up into countably many open subsets so that this can be done. However, if $x(\mathcal{U})$ does contain points in the boundary $\partial\mathbb{H}^n$, then it is not open in $\mathbb{R}^n$. A typical example is the "open" half-disk

$$\mathring{\mathbb{D}}^n_- := \left\{(x^1,\ldots,x^n) \in \mathbb{R}^n \;\middle|\; (x^1)^2 + \ldots + (x^n)^2 < 1 \text{ and } x^1 \leqslant 0\right\},$$

which is open in $\mathbb{H}^n$ but not open in $\mathbb{R}^n$ since it does not contain any ball around points in $\mathring{\mathbb{D}}^n_- \cap \partial\mathbb{H}^n$. In this sense, Definition 12.3 is strictly more general than our original definition of a chart.

The notion of **transition maps** between two charts $(\mathcal{U}, x)$ and $(\mathcal{V}, y)$ generalizes in an obvious way to boundary charts,

$$\begin{aligned} &\mathbb{H}^n \supset x(\mathcal{U}\cap\mathcal{V}) \xrightarrow{y\circ x^{-1}} y(\mathcal{U}\cap\mathcal{V}) \subset \mathbb{H}^n, \\ &\mathbb{H}^n \supset y(\mathcal{U}\cap\mathcal{V}) \xrightarrow{x\circ y^{-1}} x(\mathcal{U}\cap\mathcal{V}) \subset \mathbb{H}^n, \end{aligned} \tag{12.1}$$

though since $x(\mathcal{U}\cap\mathcal{V})$ and $y(\mathcal{U}\cap\mathcal{V})$ may be open in $\mathbb{H}^n$ but not in $\mathbb{R}^n$, the notion of smooth compatibility requires a bit of clarification. The quickest approach is to say that a map $f : \mathcal{O} \to \mathbb{R}^m$ defined on some (not necessarily open) subset $\mathcal{O} \subset \mathbb{R}^n$ is of class C^k if and only if it admits an extension of class C^k to some open neighborhood of $\mathcal{O}$ in $\mathbb{R}^n$. With this understood, we will call $(\mathcal{U}, x)$ and $(\mathcal{V}, y)$ **smoothly compatible** if both of the transition maps in (12.1) admit smooth extensions over open (in $\mathbb{R}^n$) neighborhoods of their domains.

Remark 12.4. For open subsets $\mathcal{O} \subset \mathbb{H}^n$ in half-space, the notion of a C^k-map $f : \mathcal{O} \to \mathbb{R}^m$ admits various alternative characterizations that do not require extending f over a larger neighborhood in $\mathbb{R}^n$. Denote $\partial\mathcal{O} := \mathcal{O} \cap \partial\mathbb{H}^n$ and $\mathring{\mathcal{O}} := \mathcal{O}\backslash\partial\mathcal{O}$. Then $f : \mathcal{O} \to \mathbb{R}^m$ is of class C^k if and only if its restriction $f|_{\mathring{\mathcal{O}}} : \mathring{\mathcal{O}} \to \mathbb{R}^m$ is of class C^k and either of the following equivalent conditions are satisfied:

- All partial derivatives of $f|_{\mathring{\mathcal{O}}} : \mathring{\mathcal{O}} \to \mathbb{R}^m$ up to order k admit continuous extensions over $\mathcal{O}$;
- All partial derivatives of $f|_{\mathring{\mathcal{O}}} : \mathring{\mathcal{O}} \to \mathbb{R}^m$ up to order k are uniformly continuous on bounded subsets of $\mathring{\mathcal{O}}$.

It is an easy analysis exercise to show that these two conditions are equivalent, and they clearly also follow from the assumption that $f : \mathcal{O} \to \mathbb{R}^m$ admits a C^k-extension to a neighborhood, but the converse takes more effort to prove. We will not do so here since we will never need to use this fact, but the details can be found e.g. in [**AF03**, §5.19–§5.21].

A **smooth n-manifold with boundary** can now be defined by generalizing our previous definition of a smooth n-manifold so that all charts in its maximal smooth atlas are allowed to be boundary charts. Implicit in this definition is the fact that an atlas of boundary charts on M determines a natural topology on M such that the domains of boundary charts are also open sets in M and the charts themselves are homeomorphisms onto their images. This definition is *strictly* more general than what we have been working with so far: a manifold with boundary can sometimes also be a manifold in our previous sense, because its atlas might consist only of regular charts whose images are open subsets of $\mathbb{R}^n$. But if M is a manifold with boundary, it contains a distinuished subset

$$\partial M := \left\{p \in M \;\middle|\; x(p) \in \partial\mathbb{H}^n \text{ for some smooth boundary chart } (\mathcal{U}, x)\right\},$$

called its **boundary** (*Rand*). It should be easy to convince yourself that if $x(p) \in \partial\mathbb{H}^n$ for some particular boundary chart $(\mathcal{U}, x)$, then this also holds for every other boundary chart $(\mathcal{V}, y)$ with $p \in \mathcal{V}$; this is because by the inverse function theorem, the transition maps in (12.1) necessarily preserve the interior of $\mathbb{H}^n$, and therefore also preserve its boundary $\partial\mathbb{H}^n$. Moreover, every boundary chart whose domain intersects ∂M can be viewed as a slice chart for ∂M, so that it is appropriate to call ∂M a smooth $(n-1)$-dimensional submanifold of M. In particular, ∂M inherits from M a natural smooth structure and becomes a smooth $(n-1)$-manifold. We observe that M itself is a manifold in our previous sense if and only if $\partial M = \varnothing$; one sometimes says in this case that M is a *manifold without boundary.* Since $x(\mathcal{U}) \cap \partial\mathbb{H}^n$ is *always* an open subset of $\partial\mathbb{H}^n = \{0\} \times \mathbb{R}^{n-1}$ for a boundary chart $(\mathcal{U}, x)$, the manifold ∂M never has boundary, i.e.

$$\partial(\partial M) = \varnothing.$$

Remark 12.5. One can define even more general notions such as a "manifold with boundary and corners," in which images of charts are allowed to be open subsets of quadrants like $(-\infty, 0] \times (-\infty, 0] \times \mathbb{R}^{n-2}$, in which case ∂M may also be a manifold with nonempty boundary (and possibly corners). The literature on these objects seems however to be not entirely unanimous on what the correct definitions are. In this course, we will occasionally mention corners in heuristic discussions, but we will not study them in any serious way.

Remark 12.6. From now on, you must pay careful attention whenever you see the word "manifold" without further modifiers, as its default meaning may be either "manifold without boundary" or "manifold with boundary" depending on the context. Keep in mind also that these categories are not mutually exclusive: a "manifold with boundary" may have $\partial M = \varnothing$. I generally make a point of saying "manifold with nonempty boundary" if I want to explicitly assume $\partial M \neq \varnothing$. I also will often refer to boundary charts simply as "charts" when working in the context of manifolds with boundary.

Example 12.7. Suppose N is an n-manifold without boundary and $M \subset N$ is an open subset such that $\overline{M}\backslash M \subset N$ is a smooth $(n-1)$-dimensional submanifold, i.e. a hypersurface. Then the closure $\overline{M} \subset N$ is naturally a smooth n-manifold with boundary and

$$\partial\overline{M} = \overline{M}\backslash M,$$

because every slice chart for $\overline{M}\backslash M$ can be modified in straightforward ways so as to be interpreted as a boundary chart for $\overline{M}$. Most interesting examples of manifolds with boundary arise in this way, and it can be shown that *all* manifolds with boundary are diffeomorphic to examples of this type, though the ambient manifold N might not always be a natural part of the picture. As an important special case, if $f : N \to \mathbb{R}$ is a smooth function with $c \in \mathbb{R}$ as a regular value, then $f^{-1}((-\infty, c])$ and $f^{-1}([c, \infty))$ are naturally manifolds with boundary, the boundary in each case being the regular level set $f^{-1}(c) \subset N$. Examples of this type include the n-disk $\mathbb{D}^n \subset \mathbb{R}^n$ mentioned at the beginning of this section.

Almost all of the notions we have discussed in this course so far—tangent vectors and tangent maps, vector fields, tensors, forms, orientations—can be generalized in straightforward ways for manifolds with boundary so long as one remembers what smoothness means on open sets in half-space. The tangent spaces T_pM are defined exactly as before for $p \in M\backslash\partial M$, though it takes a bit more thought to arrive at the right definition for $p \in \partial M$. Here it is useful to keep Example 12.7 in mind and imagine M as a closed subset of a larger manifold N without boundary such that $\partial M \subset N$ is a smooth hypersurface: the correct definition for $p \in \partial M$ is then $T_pM := T_pN$, so that T_pM is still a vector space of the same dimension as M. If there is no ambient manifold N in the picture, then one can instead modify the original definition of T_pM in terms of paths through p by allowing paths of the form $\gamma : (-\epsilon, 0] \to M$ or $\gamma : [0, \epsilon) \to M$ that run "out of" or "into" M through its boundary. The crucial thing to remember is that for any chart $(\mathcal{U}, x)$ with $p \in \mathcal{U}$, $d_px : T_pM \to \mathbb{R}^n$ is still a linear isomorphism, even if $p \in \partial M$. Since $\partial M \subset M$ is an $(n-1)$-dimensional submanifold, $T_p(\partial M) \subset T_pM$ is an $(n-1)$-dimensional subspace. The complement

$T_pM \backslash T_p(\partial M)$ has two connected components: one consists of all vectors that point **outward**, meaning they are derivatives of "departing" paths $\gamma : (-\epsilon, 0] \to M$, and the other contains vectors that point **inward**, which are derivatives of "entering" paths $\gamma : [0, \epsilon) \to M$. It should go without saying that flows of vector fields $X \in \mathfrak{X}(M)$ require extra care when $\partial M \neq \emptyset$, because e.g. if $p \in \partial M$ and $X(p)$ points outward/inward, then there is no forward/backward flow line starting at p for any nonzero time. There is no problem however if $X|_{\partial M}$ is everywhere tangent to the boundary, since it then also defines a flow on ∂M, and Theorem 5.1 in this case goes through without changes.

The notion of a submanifold also requires slight modification when boundaries are involved: the appropriate definition is to call $M \subset N$ a **submanifold** (with boundary) whenever it is the image of an embedding of some manifold with boundary. This allows a few possibilities that were not covered by our original definition in terms of slice charts: one of them was already mentioned above, namely the natural embedding of the boundary $\partial M \hookrightarrow M$. Another is Example 12.7: if N is an n-manifold and $M \subset N$ is an open subset such that $\partial \overline{M} := \overline{M} \backslash M$ is a smooth hypersurface in M, then $\overline{M}$ is a smooth n-dimensional submanifold with boundary in N. This opens the previously excluded possibility that a manifold and submanifold may have the same dimension without one being an open subset of the other.

Proposition 12.8. *If M is an oriented manifold of dimension $n \geqslant 2$ with boundary, then the $(n-1)$-manifold ∂M inherits a natural orientation such that for every oriented boundary chart $(\mathcal{U}, x)$ on M, $(\mathcal{U} \cap \partial M, x|_{\mathcal{U} \cap \partial M})$ is an oriented chart on ∂M. This orientation can also be characterized as follows: for every point $p \in \partial M$ and any tangent vector $\nu \in T_pM \backslash T_p(\partial M)$ that points outward, a basis $(X_1, \ldots, X_{n-1})$ of $T_p(\partial M)$ is positively oriented if and only if the basis $(\nu, X_1, \ldots, X_{n-1})$ of T_pM is positively oriented.*

The orientation defined on ∂M from an orientation of M via this proposition is called the **boundary orientation**. We will always assume unless otherwise specified that when M is oriented, ∂M is endowed with the boundary orientation.

Proof of Proposition 12.8. The main point is that any orientation-preserving transition map $\psi := y \circ x^{-1} : x(\mathcal{U} \cap \mathcal{V}) \to y(\mathcal{U} \cap \mathcal{V})$ not only preserves the subset $\partial \mathbb{H}$ but is also orientation preserving on this subset. To see this, observe that the derivative $D\psi(q) : \mathbb{R}^n \to \mathbb{R}^n$ at any point q must be an isomorphism that preserves each of the subsets $\mathbb{H}^n$ and $\partial \mathbb{H}^n$, thus it is represented by a matrix of the form

$$D\psi(q) = \begin{pmatrix} a & 0 \\ \mathbf{v} & \mathbf{B} \end{pmatrix}, \qquad a > 0, \ \mathbf{v} \in \mathbb{R}^{n-1}, \ \mathbf{B} \in \mathbb{R}^{(n-1)\times(n-1)},$$

where $\mathbf{B}$ is the derivative at q of the restricted transition map on $\partial \mathbb{H}$. Clearly $\det D\psi(q) > 0$ if and only if $\det \mathbf{B} > 0$. This shows that the restriction of an oriented atlas of M to ∂M is an oriented atlas of ∂M.

To characterize the boundary orientation in terms of bases, choose any oriented chart $(\mathcal{U}, x)$ near a point $p \in \partial M$, so the coordinate vector fields $\partial_1, \ldots, \partial_n$ define a positively-oriented basis of T_pM. The restriction of $(\mathcal{U}, x)$ to ∂M now defines an oriented chart for ∂M near p, and the coordinate vector fields for this restricted chart are $(\partial_2, \ldots, \partial_n)$, which therefore form a positively-oriented basis of $T_p(\partial M)$, and this can then be deformed continuously through bases to any other positively-oriented basis $(X_1, \ldots, X_{n-1})$ of $T_p(\partial M)$. Since ∂_1 points outward at p, it follows that for any other vector $\nu \in T_pM \backslash T_p(\partial M)$ pointing outward, the basis $(\nu, X_1, \ldots, X_{n-1})$ of T_pM can be deformed continuously through bases to $(\partial_1, \ldots, \partial_n)$, simply by deforming $(X_1, \ldots, X_{n-1})$ through bases of $T_p(\partial M)$ to $(\partial_2, \ldots, \partial_n)$ and simultaneously deforming ν through outward-pointing vectors to ∂_1. This proves that $(\nu, X_1, \ldots, X_{n-1})$ is a positively-oriented basis of T_pM, and conversely, if

$(X_1, \dots, X_{n-1})$ had been negatively oriented, we could apply the same argument to the positively-oriented basis $(-X_1, X_2, \dots, X_{n-1})$ and thus conclude that $(\nu, X_1, \dots, X_{n-1})$ is also negatively oriented. $\square$

We had to exclude the case $\dim M = 1$ from Proposition 12.8 because orientations of 0-manifolds cannot be described in terms of charts or bases.

DEFINITION 12.9. If M is an oriented 1-manifold with boundary, the **boundary orientation** of the 0-manifold ∂M is defined by calling a point $p \in \partial M$ positive if the basis of $T_p M$ formed by an outward-pointing vector $\nu \in T_p M$ is positively oriented, and negative otherwise.

EXAMPLE 12.10. Any nontrivial compact interval $[a, b] \subset \mathbb{R}$ is a 1-manifold with boundary, and if we assign it the canonical orientation of $\mathbb{R}$ then the boundary orientation of $\partial[a, b] = \{a, b\}$ makes b a positive point and a a negative point. Informally, we write

$$\partial[a, b] = -\{a\} \sqcup \{b\}.$$

A slightly different example is

$$\partial(-\infty, 0] = \{0\},$$

in which the point 0 is assigned a positive orientation; this will be relevant in the proof of Stokes' theorem below.

12.3. The boundary operator is a graded derivation. I want to point out something about boundary orientations that is not an essential part of this discussion, but it may help you to understand more intuitively why graded Leibniz rules keep showing up.

In the previous section we defined an operator "∂" that takes an oriented n-manifold M (with boundary) and returns an oriented $(n-1)$-manifold ∂M. It satisfies $\partial(\partial M) = \varnothing$ for all M, which seems formally similar to the relation $d \circ d = 0$ satisfied by the exterior derivative. We will see in the next section that the operators ∂ and d are in fact dual to each other in a sense that can be made precise, thus it should not be surprising that they have formally similar properties. We claim in particular that ∂ also satisfies a graded Leibniz rule.

To understand what this means, suppose M and N are two oriented manifolds with boundary, with $\dim M = m$ and $\dim N = n$. This discussion will be heuristic, so we will choose not to worry about the fact that $M \times N$ might not actually be a smooth manifold with boundary: in particular, the neighborhood of a point $(p, q) \in \partial M \times \partial N \subset M \times N$ cannot be described smoothly via our usual notion of a boundary chart, and a completely correct description would require the notion of manifolds with boundary *and corners* (cf. Remark 12.5). Nonetheless, it seems sensible to write

$$\partial(M \times N) = (\partial M \times N) \cup (M \times \partial N)\,, \tag{12.2}$$

and outside of the exceptional subset $\partial M \times \partial N$, it is literally true that $M \times N$ is a smooth manifold whose boundary is the union of these two pieces. Formally, $M \times N$ is a smooth manifold with boundary and corners, and its boundary consists of two smooth *faces* $\partial M \times N$ and $M \times \partial N$, each of which are smooth manifolds with boundary, and they are attached to each other at their common boundary $\partial M \times \partial N$.

Now, let's say all that again but pay attention to orientations. The product of two oriented manifolds M and N carries a natural **product orientation** such that for any $(p, q) \in M \times N$ and any pair of positively oriented bases $(X_1, \dots, X_m)$ of $T_p M$ and $(Y_1, \dots, Y_n)$ of $T_q N$, $(X_1, \dots, X_m, Y_1, \dots, Y_n)$ is a positively-oriented basis of $T_{(p,q)}(M \times N) = T_p M \times T_q N$; here we identify each $X_i \in T_p M$ with $(X_i, 0) \in T_p M \times T_q N = T_{(p,q)}(M \times N)$ and similarly identify $Y_j \in T_q N$ with $(0, Y_j) \in T_p M \times T_q N = T_{(p,q)}(M \times N)$. Now, if ∂M and ∂N are each endowed with their natural boundary orientations, then the two faces $\partial M \times N$ and $M \times \partial N$ of the boundary of $M \times N$ inherit product orientations, but these may or may not match the boundary orientation of

$\partial(M \times N)$. Indeed, at a point $(p,q) \in \partial M \times N$, if we choose a positively-oriented basis $(X_2, \ldots, X_m)$ of $T_p(\partial M)$ and an outward-pointing vector $\nu \in T_pM \backslash T_p(\partial M)$, then $(\nu, 0) \in T_{(p,q)}(M \times N)$ also points outward through $\partial M \times N$ and $(\nu, X_2, \ldots, X_m, Y_1, \ldots, Y_n)$ forms a positively-oriented basis of $T_{(p,q)}(M \times N)$, implying that the boundary orientation of $\partial(M \times N)$ does match the product orientation of $\partial M \times N$. But things are different at a point $(p,q) \in M \times \partial N$. Choosing a positively-oriented basis $(Y_2, \ldots, Y_n)$ of $T_q(\partial N)$ and an outward-pointing vector $\nu \in T_qY \backslash T_q(\partial Y)$, a positively-oriented basis of $M \times N$ is given by $(X_1, \ldots, X_m, \nu, Y_2, \ldots, Y_n)$, but m flips are required in order to permute this basis to $(\nu, X_1, \ldots, X_m, Y_2, \ldots, Y_n)$, in which ν serves as an outward-pointing vector in $T_{(p,q)}(M \times N) \backslash T_{(p,q)}(\partial(M \times N))$ and $(X_1, \ldots, X_m, Y_2, \ldots, Y_n)$ as a positively-oriented basis for the product orientation on $M \times \partial N$. This means that the product orientation of $M \times \partial N$ matches the boundary orientation of $\partial(M \times N)$ if and only if $(-1)^m = 1$, i.e. if m is even. The oriented version of (12.2) can thus be written as

$$\partial(M \times N) = (\partial M \times N) \cup ((-1)^m (M \times \partial N)), \tag{12.3}$$

where we define $-(M \times \partial N)$ to mean the oriented manifold obtained from $M \times \partial N$ by assigning it the *opposite* of the product orientation. The formal resemblance of this formula to a graded Leibniz rule is difficult to ignore, though we cannot make this notion precise in the present context since we have not defined any algebraic structure on the "set" of manifolds with boundary and corners. The easiest way to make such notions precise is probably by defining homology theory, which is a topic for a topology course and not for this one, but I wanted in any case to provide (12.3) as further evidence of a formal similarity between the operators ∂ and d.

12.4. The main result. We can now define precisely what is meant by the informal statement that the operators d and ∂ are "dual" to each other. To understand the following statement, note that a k-form $\omega \in \Omega^k(M)$ induces a k-form $\Omega^k(L)$ on every submanifold $L \subset M$ by restriction, and this applies in particular to the boundary $\partial M \subset M$. Strictly speaking, the induced k-form on ∂M in this situation is $i^*\omega \in \Omega^k(\partial M)$ for the inclusion map $i : \partial M \hookrightarrow M$, but in the following we will also denote it by $\omega \in \Omega^k(\partial M)$ instead of $i^*\omega$.

THEOREM 12.11 (Stokes). *Assume M is an oriented n-manifold with boundary, where $n \geqslant 1$, and ∂M is equipped with its natural boundary orientation. Then for every $\omega \in \Omega_c^{n-1}(M)$,*

$$\int_M d\omega = \int_{\partial M} \omega.$$

PROOF. As in the proof of Theorem 10.30, we can choose an open subset $M_0 \subset M$ with compact closure $\overline{M}_0$ such that $\operatorname{supp}(\omega) \subset M_0$, and then choose a finite covering of $\overline{M}_0$ by oriented charts $\{(\mathcal{U}_\alpha, x_\alpha)\}_{\alpha \in I}$ and a partition of unity $\{\varphi_\alpha : M \to [0,1]\}$ such that each φ_α has compact support in $\mathcal{U}_\alpha$ and $\sum_{\alpha \in I} \varphi_\alpha \equiv 1$ on M_0. Then each $\omega_\alpha := \varphi_\alpha \omega$ belongs to $\Omega_c^{n-1}(\mathcal{U}_\alpha)$, and we have $\omega = \sum_{\alpha \in I} \omega_\alpha$ and $d\omega = \sum_{\alpha \in I} d\omega_\alpha$ on M_0. If we can then prove $\int_{\mathcal{U}_\alpha} d\omega_\alpha = \int_{\partial \mathcal{U}_\alpha} \omega_\alpha$ for each α, we will have

$$\int_M d\omega = \int_{M_0} d\omega = \sum_{\alpha \in I} \int_{M_0} d\omega_\alpha = \sum_{\alpha \in I} \int_{\mathcal{U}_\alpha} d\omega_\alpha = \sum_{\alpha \in I} \int_{\partial \mathcal{U}_\alpha} \omega_\alpha = \sum_{\alpha \in I} \int_{\partial M} \omega_\alpha = \int_{\partial M} \omega.$$

In this way, the problem has been reduced to the special case in which M is covered by a single chart.

Next, observe that if the theorem has been proven to hold on another oriented manifold N and there is an orientation-preserving diffeomorphism $\psi : M \to N$, then we can write $\omega = \psi^*\alpha$ for $\alpha := \psi_*\omega \in \Omega_c^{n-1}(N)$ and use Proposition 9.18 along with the invariance of the integral under pullbacks to conclude

$$\int_M d\omega = \int_M d(\psi^*\alpha) = \int_M \psi^*(d\alpha) = \int_N d\alpha = \int_{\partial N} \alpha = \int_{\partial M} \psi^*\alpha = \int_{\partial M} \omega,$$

where we have also used the fact that a diffeomorphism $M \to N$ necessarily maps ∂M to ∂N. The latter is true since diffeomorphisms between regions in $\mathbb{R}^n$ map open sets to open sets, and neighborhoods of boundary points in $\mathbb{H}^n$ are not open in $\mathbb{R}^n$.

The combined result of the previous two paragraphs is that it will suffice to prove Stokes' theorem in the case where M is an open subset $\mathcal{U} \subset \mathbb{H}^n$ in half-space; in fact, since we are going to assume $\omega \in \Omega_c^{n-1}(\mathcal{U})$ has compact support, we may as well also assume M is the whole half-space $\mathbb{H}^n$. The proof now becomes a simple computation based on Fubini's theorem and the fundamental theorem of calculus. We can write ω in terms of n compactly supported smooth functions $f_1, \ldots, f_n : \mathbb{H}^n \to \mathbb{R}$ as

$$\omega = f_i\,\alpha^i, \qquad \text{where} \qquad \alpha^i := dx^1 \wedge \ldots \wedge \widehat{dx^i} \wedge \ldots \wedge dx^n \in \Omega^{n-1}(\mathbb{H}^n),$$

and the hat indicates again that the corresponding term does *not* appear. Then $d\alpha^i = 0$ for each i, and $dx^j \wedge \alpha^i = 0$ for every $j \neq i$, thus

$$d\omega = df_i \wedge \alpha^i = \sum_{i=1}^n \partial_i f_i\, dx^i \wedge \alpha^i = \sum_{i=1}^n (-1)^{i-1} \partial_i f_i\, dx^1 \wedge \ldots \wedge dx^n,$$

where we have refrained from using the summation convention in the last two expressions in order to avert confusion. Of the n terms in this sum, we claim that $n-1$ of them vanish when integrated over $\mathbb{H}^n$. Let us check this specifically for $i = n$: choosing $N > 0$ large enough for the supports of the functions $f_1, \ldots, f_n$ to be contained in $[-N/2, 0] \times [-N/2, N/2]^{n-1}$, we use Fubini and the fundamental theorem of calculus to compute

$$\int_{\mathbb{H}^n} \partial_n f_n(x^1, \ldots, x^n)\, dx^1 \ldots dx^n = \int_{(-\infty,0]\times\mathbb{R}^{n-2}} \left(\int_{\mathbb{R}} \partial_n f_n(x^1, \ldots, x^n)\, dx^n \right) dx^1 \ldots dx^{n-1} = 0$$

since the assumption on the support of f_n implies

$$\begin{aligned}\int_{\mathbb{R}} \partial_n f_n(x^1, \ldots, x_n)\, dx^n &= \int_{-N}^{N} \partial_n f_n(x^1, \ldots, x_n)\, dx^n \\ &= f_n(x^1, \ldots, x^{n-1}, N) - f_n(x^1, \ldots, x^{n-1}, -N) = 0.\end{aligned}$$

This calculation works out the same way for each $i = 2, \ldots, n$, thus we find

$$\begin{aligned}\int_{\mathbb{H}^n} \omega &= \int_{\mathbb{H}^n} \partial_1 f_1(x^1, \ldots, x^n)\, dx^1 \ldots dx^n = \int_{\mathbb{R}^{n-1}} \left(\int_{(-\infty,0]} \partial_1 f_1(x^1, \ldots, x^n)\, dx^1 \right) dx^2 \ldots dx^n \\ &= \int_{\mathbb{R}^{n-1}} \left(\int_{-N}^{0} \partial_1 f_1(x^1, \ldots, x^n)\, dx^1 \right) dx^2 \ldots dx^n \\ &= \int_{\mathbb{R}^{n-1}} \left(f_1(0, x^2, \ldots, x^n) - f_1(-N, x^2, \ldots, x^n) \right) dx^2 \ldots dx^n \\ &= \int_{\mathbb{R}^{n-1}} f_1(0, x^2, \ldots, x^n)\, dx^2 \ldots dx^n = \int_{\partial\mathbb{H}^n} f_1\ dx^2 \wedge \ldots \wedge dx^n.\end{aligned}$$

This last expression is $\int_{\partial\mathbb{H}^n} \omega$, as all other terms in ω contain dx^1, which vanishes when restricted to $\partial\mathbb{H}^n$. $\square$

EXAMPLE 12.12. For a smooth function $f : [a, b] \to \mathbb{R}$ on a nontrivial compact interval, we can denote the standard coordinate on $\mathbb{R}$ by x and write $df = f'\, dx$. The fundamental theorem of calculus then amounts to the following special case of Stokes' theorem,

$$\int_a^b f'(x)\, dx = \int_{[a,b]} df = \int_{-\{a\} \sqcup \{b\}} f = f(b) - f(a).$$

With this example in mind, Stokes' theorem is considered to be the natural n-dimensional generalization of the fundamental theorem of calculus.

EXERCISE 12.13. Prove the following version of *integration by parts*: if M is a compact oriented n-manifold with boundary, $\alpha \in \Omega^k(M)$ and $\beta \in \Omega^\ell(M)$ with $k + \ell = n - 1$, then

$$\int_M d\alpha \wedge \beta = \int_{\partial M} \alpha \wedge \beta - (-1)^k \int_M \alpha \wedge d\beta.$$

EXAMPLE 12.14. Heuristically, the discussion of §12.3 suggests that if M and N are compact manifolds with boundary having dimensions m and n respectively, then for any $\omega \in \Omega^{m+n-1}(M \times N)$, one should have

$$\int_{M\times N} d\omega = \int_{\partial M\times N} \omega + (-1)^m \int_{M\times\partial N} \omega. \tag{12.4}$$

Here the right hand side is obtained from the integral of ω over $\partial(M \times N)$ by splitting the latter into the two *almost* disjoint subsets $\partial M \times N$ and $M \times \partial N$ (whose intersection $\partial M \times \partial N$ is a set of measure zero in either one), and then including a sign (cf. Exercise 11.4) to account for the fact that the product orientation of $M \times \partial N$ only matches the boundary orientation of $\partial(M \times N)$ when m is odd. As it stands, the left hand side of (12.4) does not immediately make sense unless either ∂M or ∂N is empty (in which case (12.4) follows from Stokes' theorem), because $M \times N$ is otherwise not a smooth manifold with boundary. There are at least two ways that one could nonetheless make sense of (12.4):

(1) Define the notion of an oriented manifold *with boundary and corners* by allowing open subsets of $(-\infty, 0]^2 \times \mathbb{R}^{n-2}$ as local coordinate models, generalize the definition of the integral to this wider class of manifolds and prove that Stokes' theorem still holds if $\partial(M \times N)$ is understood in the sense of §12.3. This requires a bit of extra bookkeeping, but is not fundamentally more difficult than what we have already done.

(2) Choose a nested sequence of closed subsets $A_1 \subset A_2 \subset \ldots \bigcup_{j\in\mathbb{N}} A_j = M \times N$ such that each A_j is a smooth manifold with boundary (obtained by "smoothing the corner" of $M \times N$ in progressively small neighborhoods of $\partial M \times \partial N$), then define $\int_{M\times N} d\omega$ to mean $\lim_{j\to\infty} \int_{A_j} d\omega$ and deduce (12.4) from $\int_{A_j} d\omega = \int_{\partial A_j} \omega$.

REMARK 12.15. Much time and effort has been wasted by well-intentioned mathematicians trying to determine whether the correct orthography should be "Stokes' theorem" or "Stokes's theorem". After a years-long struggle I came to the conclusion that it is, essentially, a matter of personal taste. What I can say with absolute certainty is that it is not "Stoke's theorem".

12.5. The classical integration theorems. Various results that are considered central in classical vector calculus are easy consequences of Stokes' theorem.

12.5.1. *Divergence.* The **divergence** (*Divergenz*) of a vector field $X \in \mathfrak{X}(M)$ with respect to a volume form $d\mathrm{vol} \in \Omega^n(M)$ is defined as the unique real-valued function $\mathrm{div}(X) : M \to \mathbb{R}$ such that

$$d(\iota_X d\mathrm{vol}) = \mathrm{div}(X) \cdot d\mathrm{vol}. \tag{12.5}$$

The definition makes sense because $\iota_X d\mathrm{vol}$ is an $(n-1)$-form and thus $d(\iota_X d\mathrm{vol})$ is an n-form, and every n-form is at each point a scalar multiple of the given volume form. It may not seem obvious at this stage why $\mathrm{div}(X)$ is a natural thing to define—we will address this question more thoroughly next week—but the following exercise should at least make it look familiar.

EXERCISE 12.16. Assume M is an n-manifold with a fixed volume form $d\mathrm{vol} \in \Omega^n(M)$, $(\mathcal{U}, x)$ is a chart on M and $f : \mathcal{U} \to \mathbb{R}$ is the unique function such that $d\mathrm{vol} = f\, dx^1 \wedge \ldots \wedge dx^n$ on $\mathcal{U}$. Show that for any $X \in \mathfrak{X}(M)$,

$$\operatorname{div}(X) = \frac{1}{f}\partial_i(fX^i) \qquad \text{on } \mathcal{U}.$$

In particular for the standard volume form $d\mathrm{vol} = dx^1 \wedge \ldots \wedge dx^n$ on $\mathbb{R}^n$, this reduces to the standard definition of divergence in vector calculus.

If M is a compact oriented n-manifold with boundary carrying a positive volume form $d\mathrm{vol}_M \in \Omega^n(M)$ and $X \in \mathfrak{X}(M)$ is a vector field, Stokes' theorem now implies

$$\int_M \operatorname{div}(X)\, d\mathrm{vol}_M = \int_M d(\iota_X d\mathrm{vol}_M) = \int_{\partial M} \iota_X d\mathrm{vol}_M. \tag{12.6}$$

The geometric meaning of this last integral is best understood in the special case where $d\mathrm{vol}_M$ is the Riemannian volume form compatible with a Riemannian metric g on M, which we shall write in the following using the usual notation for inner products,

$$\langle X, Y\rangle := g(X, Y) \qquad \text{for } X, Y \in T_pM,\ p \in M.$$

By Proposition 11.14, the Riemannian volume form $d\mathrm{vol}_{\partial M}$ on ∂M is then

$$d\mathrm{vol}_{\partial M} := \iota_\nu d\mathrm{vol}_M|_{T(\partial M)} \in \Omega^{n-1}(\partial M),$$

where ν is the unique outward-pointing normal vector field to ∂M. (You should take a moment to convince yourself that we are getting the orientations right, i.e. $d\mathrm{vol}_{\partial M}$ really is a *positive* volume form with respect to the boundary orientation of ∂M.) To relate this to $\iota_X d\mathrm{vol}_M$, observe that along ∂M, $X = \langle X, \nu\rangle\nu + Y$ for a unique vector field $Y \in \mathfrak{X}(\partial M)$, but $\iota_Y d\mathrm{vol}_M$ vanishes when restricted to the boundary because feeding it any $(n-1)$-tuple of vectors $Y_1, \ldots, Y_{n-1}$ tangent to ∂M means evaluating $d\mathrm{vol}_M$ on $(Y, Y_1, \ldots, Y_{n-1})$, and those are *all* tangent to the $(n-1)$-dimensional boundary and thus cannot be linearly independent. We conclude

$$\iota_X d\mathrm{vol}_M|_{T(\partial M)} = \langle X, \nu\rangle\ \iota_\nu d\mathrm{vol}_M|_{T(\partial M)} = \langle X, \nu\rangle\, d\mathrm{vol}_{\partial M},$$

and the implication of (12.6) is thus

$$\boxed{\int_M \operatorname{div}(X)\, d\mathrm{vol}_M = \int_{\partial M} \langle X, \nu\rangle\, d\mathrm{vol}_{\partial M}}. \tag{12.7}$$

This is a mild generalization of the classical result known as *Gauss's divergence theorem.*[43] Physics textbooks like to write their favorite special case of this result in some form such as

$$\iiint_\Omega (\nabla \cdot \mathbf{X})\, dV = \oiint_{\partial\Omega} \mathbf{X} \cdot d\mathbf{a}, \tag{12.8}$$

where $\Omega \subset \mathbb{R}^3$ is assumed to be a compact region bounded by a smooth surface $\partial\Omega \subset \mathbb{R}^3$, $\nabla \cdot \mathbf{X}$ is the divergence of a vector field $\mathbf{X} \in \mathfrak{X}(\Omega)$ with respect to the standard volume form $d\mathrm{vol}_{\mathbb{R}^3} := dx \wedge dy \wedge dz$, the "V" in $dV := d\mathrm{vol}_{\mathbb{R}^3}$ stands for "volume" and the "a" in $\mathbf{X} \cdot d\mathbf{a} := \langle \mathbf{X}, \nu\rangle\, d\mathrm{vol}_{\partial\Omega}$ stands for "area". (The symbol $d\mathbf{a}$ in this situation is thought of as a "vector-valued measure" that encodes not only the 2-dimensional measure on $\partial\Omega$ but also its normal vector field.) The repetition of the integral signs corresponds to the dimension of the manifold and can be seen as a reference to Fubini's theorem; the additional loop in $\oiint$ merely refers to the fact that $\partial\Omega$ is a "closed" surface (the 2-dimensional analogue of a closed loop), i.e. it is compact and has no boundary. Gauss's theorem has an important interpretation in electrostatics: if $\mathbf{X}$ represents the electric field on a

[43] or possibly "Gauss' divergence theorem", I don't know

region $\Omega \subset \mathbb{R}^3$, then its divergence is the electrical charge density, and (12.8) thus says that the total electrical charge in the region Ω is equal to the total *flux* of the electric field through the boundary of Ω.

12.5.2. *Curl.* The next example only makes sense in the case

$$\dim M = 3.$$

It relies on the observation that for any n-dimensional vector space V with a nontrivial top-dimensional form $\omega \in \Lambda^n V^*$, the map

$$V \to \Lambda^{n-1} V^* : v \mapsto \iota_v \omega$$

is an isomorphism. Indeed, it is clearly injective since $\omega \neq 0$ and any $v \neq 0$ can be extended to a basis of V, so surjectivity then follows from the fact that $\dim \Lambda^{n-1} V^* = \binom{n}{n-1} = n = \dim V$. With this understood, any volume form $d\mathrm{vol}_M$ on a 3-manifold M determines an isomorphism

$$\mathfrak{X}(M) \xrightarrow{\cong} \Omega^2(M) : X \mapsto \iota_X d\mathrm{vol}_M.$$

Let us now assume (M, g) is an oriented Riemannian 3-manifold and $d\mathrm{vol}_M$ is its Riemannian volume form. The metric $\langle \ , \ \rangle := g$ also determines an isomorphism

$$\mathfrak{X}(M) \xrightarrow{\cong} \Omega^1(M) : X \mapsto X_\flat := \langle X, \cdot \rangle.$$

The **curl** (*Rotation*) of $X \in \mathfrak{X}(M)$ is then defined as the unique vector field $\mathrm{curl}(X) \in \mathfrak{X}(M)$ such that

$$\iota_{\mathrm{curl}(X)} d\mathrm{vol}_M = d(X_\flat).$$

EXERCISE 12.17. Convince yourself that on $M := \mathbb{R}^3$ with its standard Riemannian metric defined via the Euclidean inner product, the curl of a vector field is the same thing that you learned about once upon a time in vector calculus.

Now if $\Sigma \subset M$ is an oriented 2-dimensional submanifold with boundary, Σ and $\partial\Sigma$ each inherit Riemannian metrics as submanifolds of M, and thus have canonical Riemannian volume forms $d\mathrm{vol}_\Sigma$ and $d\mathrm{vol}_{\partial\Sigma}$ respectively. For an appropriate choice[44] of normal vector field ν along Σ, Proposition 11.14 implies

$$d\mathrm{vol}_\Sigma = \iota_\nu d\mathrm{vol}_M|_{T\Sigma} \in \Omega^2(\Sigma),$$

and a repeat of the same argument we used for the divergence theorem then implies that for any $Y \in \mathfrak{X}(M)$,

$$\iota_Y d\mathrm{vol}_M|_{T\Sigma} = \langle Y, \nu \rangle \, d\mathrm{vol}_\Sigma.$$

If $Y = \mathrm{curl}(X)$ for some $X \in \mathfrak{X}(M)$, Stokes' theorem now implies

$$\int_\Sigma \langle \mathrm{curl}(X), \nu \rangle \, d\mathrm{vol}_\Sigma = \int_\Sigma d(X_\flat) = \int_{\partial\Sigma} X_\flat.$$

To understand the integral on the right, let $\tau \in \mathfrak{X}(\partial\Sigma)$ denote the unique positively-oriented unit vector field on $\partial\Sigma$, so $d\mathrm{vol}_{\partial\Sigma}(\tau) = 1$, and $X_\flat(\tau) = \langle X, \tau \rangle$ thus implies $X_\flat|_{T(\partial\Sigma)} = \langle X, \tau \rangle \, d\mathrm{vol}_{\partial\Sigma}$, and we obtain

$$\boxed{\int_\Sigma \langle \mathrm{curl}(X), \nu \rangle \, d\mathrm{vol}_\Sigma = \int_{\partial\Sigma} \langle X, \tau \rangle \, d\mathrm{vol}_{\partial\Sigma}}. \tag{12.9}$$

[44]One can deduce from the assumption that both M and Σ are oriented that a normal vector field ν along Σ exists, and there are multiple choices—if Σ is connected, then there are exactly two choices, differing by a sign. The *appropriate* choice is the one that makes the volume form $\iota_\nu d\mathrm{vol}_M$ on Σ positive.

This generalizes what is usually called the "classical" Stokes' theorem in vector calculus. In physics textbooks, one finds it written for the case $\Sigma \subset \mathbb{R}^3$ with the standard metric as

$$\iint_{\Sigma} (\nabla \times \mathbf{X}) \cdot d\mathbf{a} = \oint_{\partial\Sigma} \mathbf{X} \cdot d\mathbf{l},$$

where $\nabla \times \mathbf{X}$ denotes the curl of $\mathbf{X} \in \mathfrak{X}(\mathbb{R}^3)$, $d\mathbf{a}$ is the same "vector-valued measure" that appeared in (12.8), and $d\mathbf{l}$ similarly denotes a 1-dimensional vector-valued measure that encodes both the volume form $d\mathrm{vol}_{\partial\Sigma}$ and the tangent vector field τ.

13. Closed and exact forms

13.1. Some easy applications of Stokes. The following terminology is used consistently throughout differential geometry.

Definition 13.1. A manifold M is **closed** (*geschlossen*) if it is compact and $\partial M = \varnothing$. We say that M is **open** (*offen*) if none of its connected components are closed, i.e. they all are noncompact and/or have nonempty boundary.[45]

Example 13.2. Manifolds of dimension 0 never have boundary, so a 0-manifold is closed if and only if it is compact, i.e. it is a discrete finite set.

Example 13.3. If M is a compact manifold with boundary, then ∂M is a closed manifold.

Definition 13.4. A differential form $\omega \in \Omega^k(M)$ is called **closed** (*geschlossen*) if $d\omega = 0$, and it is called **exact** (*exakt*) of $\omega = d\alpha$ for some $\alpha \in \Omega^{k-1}(M)$. In the latter situation, the form α is called a **primitive** of ω.

Example 13.5. A closed 0-form is the same thing as a locally constant function, and an exact 1-form is the same thing as a differential. There are no exact 0-forms since there is no such thing as a (-1)-form.

Example 13.6. On an n-manifold, every n-form is closed since there are no nontrivial $(n+1)$-forms.

Example 13.7. Given a volume form $d\mathrm{vol} \in \Omega^n(M)$, a vector field $X \in \mathfrak{X}(M)$ has vanishing divergence if and only if the $(n-1)$-form $\iota_X d\mathrm{vol}$ is closed. Similarly, if (M, g) is an oriented Riemannian 3-manifold, $X \in \mathfrak{X}(M)$ has vanishing curl if and only if the 1-form $X_\flat := g(X, \cdot)$ is closed.

Here is a bit of low-hanging fruit that can be picked as soon as one understands the above definitions and the statement of Stokes' theorem.

Proposition 13.8. *If M is a closed oriented n-manifold and $\omega \in \Omega^n(M)$ is exact, then $\int_M \omega = 0$. Similarly, if M is a compact oriented n-manifold with boundary and $\alpha \in \Omega^{n-1}(M)$ is closed, then $\int_{\partial M} \alpha = 0$.*

[45] Be aware that the word "closed" has a different meaning when referring to a manifold than it does when referring to a subset of a topological space. For instance, if M is a manifold, then a compact submanifold $\Sigma \subset M$ with boundary is a closed subset of M, but it is not a closed manifold if $\partial\Sigma \neq \varnothing$. The German language uses two different words for these separate meanings of "closed": a subset in a topological space can be *abgeschlossen*, but a manifold can be *geschlossen*.

PROOF. If you review the proof of Stokes' theorem, you will find that it is valid in the case $\partial M = \varnothing$ so long as one understands every integral over $\varnothing$ to be 0 by definition. Thus $\partial M = \varnothing$ and $\omega = d\beta$ for some $\beta \in \Omega^{n-1}(M)$ implies

$$\int_M \omega = \int_M d\beta = \int_\varnothing \beta = 0,$$

and if ∂M is not assumed empty but $\alpha \in \Omega^{n-1}(M)$ is closed,

$$\int_{\partial M} \alpha = \int_M d\alpha = 0.$$

□

COROLLARY 13.9. *On a closed oriented n-manifold M, every n-form $\omega \in \Omega^n(M)$ with $\int_M \omega \neq 0$ is closed but not exact. In particular, this is true whenever ω is a volume form.* □

REMARK 13.10. One can show that Corollary 13.9 fails whenever either $\partial M \neq \varnothing$ or M is noncompact. In the former case, $\int_M \omega \neq 0$ for an exact form $\omega = d\alpha$ is not a contradiction, since $\int_{\partial M} \alpha$ might also be nonzero. There is a different problem if M has empty boundary but is noncompact: the use of Stokes' theorem to derive the contradiction $0 \neq \int_M d\alpha = \int_{\partial M} \alpha = 0$ is not valid unless α has compact support, so it can happen for instance that $\omega \in \Omega^n_c(M)$ satisfies $\int_M \omega \neq 0$ and is the exterior derivative of an $(n-1)$-form whose support is noncompact. We will see shortly that, indeed, *every* n-form on $\mathbb{R}^n$ for $n \geqslant 1$ is exact (see Corollary 13.34 below).

EXERCISE 13.11. Show that for each $k \geqslant 0$, a k-form $\omega \in \Omega^k(M)$ is closed if and only for every compact oriented $(k+1)$-dimensional submanifold $L \subset M$ with boundary, $\int_{\partial L} \omega = 0$.
Hint: For any point $p \in M$ and linearly-independent vectors $X_1, \dots, X_{k+1} \in T_pM$, you could choose $L \subset M$ to be a small $(k+1)$-disk through p tangent to the space spanned by $X_1, \dots, X_{k+1}$.

13.2. The Poincaré lemma and simple connectedness. The observation in Example 13.3 that boundaries of compact manifolds are closed has a dual statement for differential forms: since $d^2 := d \circ d = 0$, every exact differential form is also closed. Corollary 13.9 reveals however that the converse is generally false. Here is a more concrete example.

EXAMPLE 13.12. On $\mathbb{R}^2 \backslash \{0\}$, one can define a smooth 1-form in Cartesian coordinates (x, y) by

$$\lambda := \frac{1}{x^2 + y^2}\,(x\,dy - y\,dx)\,.$$

This expression takes a more revealing form of one rewrites it in polar coordinates: assume $\mathcal{U} \subset \mathbb{R}^2 \backslash \{0\}$ is a subset on which there is a well-defined chart of the form $(r, \theta) : \mathcal{U} \to \mathbb{R}^2$ such that r takes positive values and the relations $x = r\cos\theta$ and $y = r\sin\theta$ hold; concretely, we can take $\mathcal{U}$ to be the complement of a ray $\{tv \in \mathbb{R}^2 \mid t \in [0, \infty)\}$ for some $v \in \mathbb{R}^2 \backslash \{0\}$, and the image of θ is then an open interval of the form $(c, c + 2\pi)$. In terms of r and θ, we have $dx = (\cos\theta)\,dr - (r\sin\theta)\,d\theta$ and $dy = (\sin\theta)\,dr + (r\cos\theta)\,d\theta$, thus

$$\lambda = \frac{1}{r^2}\left[r\cos\theta\,(\sin\theta\,dr + r\cos\theta\,d\theta) - r\sin\theta\,(\cos\theta\,dr - r\sin\theta\,d\theta)\right] = d\theta,$$

so λ is exact on $\mathcal{U}$. Since this computation holds independently of the choice of domain $\mathcal{U} \subset \mathbb{R}^2 \backslash \{0\}$, it follows that $d\lambda = 0$ everywhere. But the restriction of $(\mathcal{U}, (r, \theta))$ to $\{r = 1\}$ now defines a chart on $S^1 \subset \mathbb{R}^2 \backslash \{0\}$ in the form $(S^1 \backslash \{q\}, \theta)$ for some point $q \in S^1$, which is a set of measure zero, thus $\int_{S^1} \lambda$ can be computed using the methods of §11.2, and the answer is

$$\int_{S^1} \lambda = \int_{(c, c+2\pi)} d\theta = 2\pi \neq 0.$$

This clearly could not happen if λ were df for some $f \in \Omega^0(\mathbb{R}^2\backslash\{0\}) = C^\infty(\mathbb{R}^2\backslash\{0\})$, as the restriction of λ to S^1 would then be $d(f|_{S^1})$ and we would have a contradiction to Proposition 13.8.

REMARK 13.13. It is conventional to denote the 1-form in Example 13.12 by

$$d\theta \in \Omega^1(\mathbb{R}^2\backslash\{0\})$$

even though, strictly speaking, it is not the differential of any smooth function $\theta : \mathbb{R}^2\backslash\{0\} \to \mathbb{R}$. One reasonable way to think about it is that while θ cannot be defined on this domain as a smooth *real*-valued function, it can be defined to take values in the quotient $\mathbb{R}/2\pi\mathbb{Z}$, which is a smooth manifold and $\theta : \mathbb{R}^2\backslash\{0\} \to \mathbb{R}/2\pi\mathbb{Z}$ in this sense is a smooth map. The latter means in practice that any point $p \in \mathbb{R}^2\backslash\{0\}$ admits a neighborhood $\mathcal{U} \subset \mathbb{R}^2\backslash\{0\}$ on which the smooth function $\theta : \mathcal{U} \to \mathbb{R}$ can be defined, though this function is not unique, as it can equally well be replaced by $\theta + 2\pi m$ for any $m \in \mathbb{Z}$. But modifying θ by addition of a constant does not change its differential, thus $d\theta$ is uniquely defined.

Remark 13.13 illustrates a phenomenon that is generalized in the following result: every closed differential form is "locally" exact.

THEOREM 13.14 (the Poincaré Lemma). *If $\omega \in \Omega^k(M)$ is closed and $k \geqslant 1$, then for every $p \in M$ there exists a neighborhood $\mathcal{U} \subset M$ of p and a $(k-1)$-form $\alpha \in \Omega^{k-1}(\mathcal{U})$ such that $d\alpha = \omega$ on $\mathcal{U}$.*

A proof of the Poincaré lemma will be given at the end of this lecture. The next two results are easier to prove, but imply a stronger statement for the case $k = 1$.

LEMMA 13.15. *A 1-form $\lambda \in \Omega^1(M)$ is exact if and only if $\int_{S^1} \gamma^*\lambda = 0$ for all smooth maps $\gamma : S^1 \to M$.*

PROOF. If $\lambda = df$ for some $f \in C^\infty(M)$, then Proposition 13.8 implies $\int_{S^1} \gamma^*\lambda = \int_{S^1} \gamma^* df = \int_{S^1} d(\gamma^* f) = 0$ for every smooth map $\gamma : S^1 \to M$. Conversely, assume $\int_{S^1} \gamma^*\lambda$ always vanishes. The following recipe for constructing a function $f : M \to \mathbb{R}$ with $df = \lambda$ can be applied on every connected component of M separately, so we may as well assume M is connected. We claim that if we fix a reference point $p_0 \in M$, then $f : M \to \mathbb{R}$ can be defined by

$$f(p) := \int_0^a \lambda(\dot\gamma(t))\,dt \quad \text{for any } a > 0,\ \gamma \in C^\infty([0,a], M) \text{ with } \gamma(0) = p_0,\ \gamma(a) = p. \tag{13.1}$$

We must first show that $f(p)$ is independent of the choice of the path $\gamma : [0,a] \to M$ from p_0 to p. To this end, here are two useful observations: first, by the substitution rule, the integral in (13.1) does not change if we replace $\gamma : [0,a] \to M$ with $\gamma \circ \psi : [0,1] \to M$ for any smooth map $\psi : [0,1] \to [0,a]$ with $\psi(0) = 0$ and $\psi(1) = a$. As a consequence, we lose no generality by restricting our attention to paths $\gamma : [0,1] \to M$ that are constant on neighborhoods of 0 and 1, with values p_0 and p respectively. The second observation is that if t denotes the standard coordinate on the 1-manifold $[0,1] \subset \mathbb{R}$, then $(\gamma^*\lambda)_t(\partial_t) = \lambda_{\gamma(t)}(\gamma_*\partial_t) = \lambda_{\gamma(t)}(\dot\gamma(t))$, thus we can also write

$$f(p) = \int_{[0,1]} \gamma^*\lambda.$$

Now if $\gamma_1, \gamma_2 : [0,1] \to M$ are two smooth paths from p_0 to p that are both constant near 0 and 1, we can concatenate γ_1 with the reversal of γ_2 to form a smooth loop $\varphi : S^1 \to M$ in the form

$$\varphi(e^{\pi i t}) = \begin{cases} \gamma_1(t) & \text{for } 0 \leqslant t \leqslant 1, \\ \gamma_2(2-t) & \text{for } 1 \leqslant t \leqslant 2, \end{cases}$$

where for convenience we are identifying $\mathbb{R}^2$ in the obvious way with $\mathbb{C}$ so that $S^1 \subset \mathbb{C}$. If we now split S^1 into its upper and lower semicircles $S^1_\pm$ with parametrizations $\psi_\pm : [0,1] \to S^1_\pm : t \mapsto e^{\pi i t}$, we have $\gamma_1 = \varphi \circ \psi_+$ and $\gamma_2 = \varphi \circ \psi_-$, but ψ_+ is orientation preserving while ψ_- is orientation reversing, thus

$$\begin{aligned} 0 &= \int_{S^1} \varphi^*\lambda = \int_{S^1_+} \varphi^*\lambda + \int_{S^1_-} \varphi^*\lambda = \int_{\psi_+([0,1])} \varphi^*\lambda + \int_{\psi_-([0,1])} \varphi^*\lambda \\ &= \int_{[0,1]} \psi_+^*\varphi^*\lambda - \int_{[0,1]} \psi_-^*\varphi^*\lambda = \int_{[0,1]} (\varphi\circ\psi_+)^*\lambda - \int_{[0,1]} (\varphi\circ\psi_-)^*\lambda = \int_{[0,1]} \gamma_1^*\lambda - \int_{[0,1]} \gamma_2^*\lambda. \end{aligned}$$

With independence of the choice of γ established, we observe that (13.1) implies $\frac{d}{dt} f(\gamma(t)) = \lambda(\dot\gamma(t))$ for every t and every smooth path γ starting at p_0, thus $df = \lambda$. □

Exercise 13.16. Use a slight modification of the proof of Lemma 13.15 to show that on S^1, a 1-form $\lambda \in \Omega^1(S^1)$ is exact if and only if $\int_{S^1} \lambda = 0$.

Definition 13.17. A smooth manifold M is **simply connected** (*einfach zusammenhängend*) if it is connected and every smooth map $\gamma : S^1 \to M$ admits a smooth extension over the 2-disk, i.e. a map $u : \mathbb{D}^2 \to M$ such that $u|_{\partial\mathbb{D}^2} = \gamma$.

Remark 13.18. In algebraic topology, a topological space is called simply connected if it is path-connected and its fundamental group vanishes, but for smooth manifolds, Definition 13.17 is equivalent to this condition. In particular, one could replace the word "smooth" by "continuous" without changing anything, because by general perturbation results in differential topology (see e.g. [**Hir94**]), continuous maps between smooth manifolds always admit smooth approximations.

Theorem 13.19. *If M is a simply connected manifold, then every closed 1-form $\lambda \in \Omega^1(M)$ is exact.*

Proof. If $\lambda \in \Omega^1(M)$ is closed and every smooth map $\gamma : S^1 \to M$ admits a smooth extension $u : \mathbb{D}^2 \to M$, then

$$\int_{S^1} \gamma^*\lambda = \int_{\partial\mathbb{D}^2} u^*\lambda = \int_{\mathbb{D}^2} d(u^*\lambda) = \int_{\mathbb{D}^2} u^*(d\lambda) = 0,$$

hence λ satisfies the criterion of Lemma 13.15 and is therefore exact. □

It should be easy to convince yourself that every convex subset of $\mathbb{R}^n$ is simply connected, and every point in a manifold has a neighborhood that looks like a convex subset of $\mathbb{R}^n$ in local coordinates, implying in turn that that neighborhood is simply connected. Theorem 13.19 thus implies the $k = 1$ case of the Poincaré lemma. But it also implies more, because there are many simply connected manifolds that are more interesting than convex sets.

Example 13.20. For each $n \geqslant 2$, the sphere S^n is simply connected. Here is an incomplete but (maybe?) believable proof: since $\dim S^n > \dim S^1$, no smooth map $\gamma : S^1 \to S^n$ can be surjective,[46] i.e. it must miss at least one point $p \in S^n$ and can thus be viewed as a map $S^1 \to S^n \setminus \{p\}$. But by stereographic projection, one can also find a diffeomorphism of $S^n \setminus \{p\}$ to $\mathbb{R}^n$ and then appeal to the fact that $\mathbb{R}^n$ (as a convex set) is simply connected. It follows that closed 1-forms on S^n for $n \geqslant 2$ are always exact.

[46] I'm pretty sure that you cannot visualize any surjective smooth map $f : M \to N$ when $\dim M < \dim N$, though actually proving they don't exist is not completely trivial. It follows easily from Sard's theorem, a fundamental result in differential topology stating that the set of critical values of a smooth map $f : M \to N$ always has measure zero. This means that for almost every $q \in N$, $T_p f : T_p M \to T_q N$ is surjective for every $p \in f^{-1}(q)$; the only way for this to hold when $\dim M < \dim N$ is if $f^{-1}(q) = \varnothing$. The much more surprising fact is that *continuous* maps $f : M \to N$ can be surjective, even when $\dim N > \dim M$; look up the term "space-filling curve". Such maps can never be smooth.

REMARK 13.21. You may have noticed that in Theorem 13.19, it would have sufficed to assume that every smooth map $\gamma : S^1 \to M$ admits a smooth extension $u : \Sigma \to M$ over *some* compact, smooth, oriented surface Σ with boundary $\partial\Sigma = S^1$, i.e. not necessarily the disk, but any surface whose boundary is a circle. (An easy example would be obtained by cutting a hole out of the 2-torus $\mathbb{T}^2$.) This means that Theorem 13.19 is true under a somewhat more general hypothesis than simple connectedness. The natural language for this generalization is homology, i.e. the theorem holds for any manifold M whose first homology group with real coefficients vanishes. A full explanation of this statement would require a major digression into algebraic topology, so we will not discuss it any further here, but suffice it to say that in dimension 2, there are no examples for which this distinction makes a difference, but in dimension 3 there are. Poincaré famously conjectured that every closed 3-manifold with vanishing first homology group is homeomorphic to S^3, but later found an example—now known as the *Poincaré homology sphere*—that satisfies this hypothesis but (unlike S^3) is not simply connected, and thus had to revise his conjecture. The revised conjecture was proved over 100 years later.

EXAMPLE 13.22. On a Riemannian manifold (M, g), the inner product $\langle\ ,\ \rangle := g$ determines an isomorphism $T_pM \to T_p^*M : X \mapsto X_\flat := \langle X, \cdot\rangle$ at every point $p \in M$, which can be used to associate to any smooth function $f : M \to \mathbb{R}$ its **gradient** vector field $\nabla f \in \mathfrak{X}(M)$, uniquely determined by

$$df = \langle \nabla f, \cdot\rangle.$$

A vector field $X \in \mathfrak{X}(M)$ cannot be the gradient of a function unless the 1-form $X_\flat \in \Omega^1(M)$ is closed, and conversely, the Poincaré lemma implies that every vector field satisfying this condition is *locally* the gradient of a function, though perhaps not globally (unless M is simply connected). If M is oriented and 3-dimensional, then this result can also be expressed in terms of the curl (cf. §12.5.2): any gradient $X = \nabla f$ satisfies $\iota_{\mathrm{curl}(X)} d\mathrm{vol}_M = d(df) = 0$, implying

$$\mathrm{curl}(\nabla f) \equiv 0,$$

and conversely, any vector field $X \in \mathfrak{X}(M)$ with $\mathrm{curl}(X) \equiv 0$ is locally the gradient of a function.

In the same context, the curl of any vector field $X \in \mathfrak{X}(M)$ satisfies $\iota_{\mathrm{curl}(X)} d\mathrm{vol}_M = d(X_\flat)$ and thus $d(\iota_{\mathrm{curl}(X)} d\mathrm{vol}_M) = d^2(X_\flat) = 0$, implying

$$\mathrm{div}(\mathrm{curl}(X)) \equiv 0.$$

Conversely, any divergenceless vector field $Y \in \mathfrak{X}(M)$ satisfies $d(\iota_Y d\mathrm{vol}_M) = 0$, so that by the Poincaré lemma, $\iota_Y d\mathrm{vol}_M \in \Omega^2(M)$ can be written on any sufficienly small neighborhood $\mathcal{U}$ as $d\lambda$ for some $\lambda \in \Omega^1(\mathcal{U})$. The latter is also $X_\flat$ for a unique vector field $X \in \mathfrak{X}(\mathcal{U})$, whose curl is therefore Y: in other words, any divergenceless vector field is locally the curl of another vector field.

While (13.1) provides a fairly straightforward recipe to find a local primitive of any closed 1-form, it is not as easy to derive local primitives for closed k-forms when $k \geqslant 2$. One possible approach is to work on "boxes" of the form $M := (a_1, b_1) \times \ldots \times (a_n, b_n)$ and proceed by induction on the number of dimensions, showing that if one can already find primitives for closed k-forms on the hypersurface $\Sigma_c := (a_1, b_1) \times \ldots \times (a_{n-1}, b_{n-1}) \times \{c\}$ for some constant $c \in (a_n, b_n)$, then primitives on Σ_c can be extended to primitives on M by integrating in the nth direction. I have proved the Poincaré lemma in this way when I've taught analysis courses (see [Wen19]), but the idea behind the argument has a tendency to get lost behind computational details. We will adopt a different approach in these notes, and deduce the Poincaré lemma from a deeper theorem about the homotopy-invariance of de Rham cohomology. We will see at the end that this approach does lead to an explicit formula generalizing (13.1) to produce local primitives of closed k-forms (see in partiular Remark 13.39), but in contrast with (13.1), one would be very unlikely to find this formula from an educated guess.

13.3. De Rham cohomology. By now we have gathered some evidence that the distinction between closed and exact forms on a manifold M has something to do with the topology of M. We shall now formalize this relation by defining an algebraic invariant of smooth manifolds.

Definition 13.23. For a smooth n-manifold M and each integer $k \in \mathbb{Z}$, let $d_k : \Omega^k(M) \to \Omega^{k+1}(M)$ denote the restriction of the exterior derivative $d : \Omega^*(M) \to \Omega^*(M)$ to the subspace $\Omega^k(M) \subset \Omega^*(M)$, with the convention that for $k < 0$, $\Omega^k(M)$ is the trivial subspace (hence d_{-1} is the trivial map into $\Omega^0(M)$). The k**th de Rham cohomology** of M is the vector space

$$H^k_{\mathrm{dR}}(M) := \ker(d_k) \big/ \operatorname{im}(d_{k-1}),$$

i.e. it is the quotient of the space of closed k-forms by the subspace of exact k-forms. We write

$$H^*_{\mathrm{dR}}(M) := \bigoplus_{k \in \mathbb{Z}} H^k_{\mathrm{dR}}(M).$$

Remark 13.24. The case $k < 0$ was included in Definition 13.23 only in order to make sure that the definition of $H^0_{\mathrm{dR}}(M)$ makes sense, but $H^k_{\mathrm{dR}}(M)$ for $k < 0$ is just the trivial vector space, and we will have no need to mention it again. It is similarly easy to see that $H^k_{\mathrm{dR}}(M) = 0$ whenever $k > \dim M$, since the space of k-forms is already trivial in this case. Thus in practice, $H^k_{\mathrm{dR}}(M)$ is potentially interesting only for k in the range $0 \leqslant k \leqslant \dim M$.

It may seem surprising at first glance that $H^k_{\mathrm{dR}}(M)$ is useful or computable: in typical cases both $\ker(d_k)$ and $\operatorname{im}(d_{k-1})$ are infinite-dimensional vector spaces, and one would not normally expect the quotient of one infinite-dimensional space by another one to carry interesting information. It turns out however that in almost all interesting cases, the quotient is finite dimensional, and its dimension is a useful numerical invariant of manifolds. Let us first clarify what is meant by the word "invariant".

Proposition 13.25. *For smooth maps $f : M \to N$, the linear map $f^* : \Omega^k(N) \to \Omega^k(M)$ sends closed forms on N to closed forms on M, and it also descends[47] to the quotients to define a linear map $f^* : H^k_{\mathrm{dR}}(N) \to H^k_{\mathrm{dR}}(M)$ that satisfies the following properties:*

(1) For another smooth map $g : N \to Q$, $(g \circ f) = f^* g^* : H^k_{\mathrm{dR}}(Q) \to H^k_{\mathrm{dR}}(M)$;
(2) For the identity map $\mathrm{Id} : M \to M$, $\mathrm{Id}^* : H^k_{\mathrm{dR}}(M) \to H^k_{\mathrm{dR}}(M)$ is the identity map.

It follows in particular that whenever $f : M \to N$ is a diffeomorphism, $f^* : H^k_{\mathrm{dR}}(N) \to H^k_{\mathrm{dR}}(M)$ is a vector space isomorphism for each k.

Proof. The relation $f^*(d\omega) = d(f^*\omega)$ implies that f^* preserves both the spaces of closed forms and exact forms, and thus descends to their quotient. The rest of the statement follows immediately from the basic properties of pullbacks. □

Remark 13.26. For those who enjoy this kind of language, Proposition 13.25 says that H^k_{dR} for each $k \in \mathbb{Z}$ defines a contravariant functor from the category of smooth manifolds and smooth maps to the category of real vector spaces and linear maps.

Example 13.27. The closed 0-forms on M are the locally constant functions, which can take independent but constant values on each connected component of M, while the subspace of exact 0-forms is trivial, thus if M has $N \in \mathbb{N}$ connected components, $H^0_{\mathrm{dR}}(M) \cong \mathbb{R}^N$.

[47] Recall that if $A : V \to W$ is a linear map between vector spaces and $X \subset V$ and $Y \subset W$ are linear subspaces such that $A(X) \subset Y$, then there is a well-defined linear map $V/X \to W/Y$ sending the equivalence class $[x] \in V/X$ of each $x \in V$ to the equivalence class $[Ax] \in W/Y$ of $Ax \in W$. One says in this situation that $A : V \to W$ **descends** to a map $V/X \to W/Y$.

EXAMPLE 13.28. If $M := \{\mathrm{pt}\}$ is the 0-manifold consisting of a single point, then $\Omega^0(\{\mathrm{pt}\}) \cong \mathbb{R}$, $\Omega^k(\{\mathrm{pt}\}) = 0$ for each $k > 0$, and the exterior derivative is the trivial map, implying

$$H^k_{\mathrm{dR}}(\{\mathrm{pt}\}) \cong \begin{cases} \mathbb{R} & \text{for } k = 0, \\ 0 & \text{for } k > 0. \end{cases}$$

EXAMPLE 13.29. Theorem 13.19 implies that $H^1_{\mathrm{dR}}(M) = 0$ whenever M is simply connected.

EXAMPLE 13.30. Corollary 13.9 implies that $H^n_{\mathrm{dR}}(M) \neq 0$ whenever M is a closed oriented n-manifold.

Diffeomorphism-invariance is a nice property, but de Rham cohomology also satisfies a stronger invariance property that makes it much easier to compute.

DEFINITION 13.31. Two smooth maps $f_0, f_1 : M \to N$ are called **smoothly homotopic** (*glatt homotop*) if there exists a smooth map $h : [0,1] \times M \to N$ such that $h(0,\cdot) = f_0$ and $h(1,\cdot) = f_1$.

THEOREM 13.32. *If $f_0, f_1 : M \to N$ are smoothly homotopic maps, then for each k, the linear maps $H^k_{\mathrm{dR}}(N) \to H^k_{\mathrm{dR}}(M)$ defined by f_0^* and f_1^* are identical.*

Before proving this, let's think through some of the consequences. A map $f : M \to N$ is called a **smooth homotopy equivalence** (*glatte Homotopieäquivalenz*) if there exists another smooth map $g : N \to M$ such that $f \circ g : N \to N$ and $g \circ f : M \to M$ are each smoothly homotopic to the identity map. Combining Proposition 13.25 with Theorem 13.32 in this situation implies that $f^* : H^*_{\mathrm{dR}}(N) \to H^*_{\mathrm{dR}}(M)$ and $g^* : H^*_{\mathrm{dR}}(M) \to H^*_{\mathrm{dR}}(N)$ are inverses; in particular, f^* is an isomorphism:

COROLLARY 13.33. *If two manifolds M and N are smoothly homotopy equivalent, then their de Rham cohomologies are isomorphic.* □

The power of Corollary 13.33 lies in the fact that two manifolds can easily be homotopy equivalent without being diffeomorphic; in fact, homotopy equivalence does not even imply that they have the same dimension. Here is an extreme example: a manifold M is called **smoothly contractible** (*glatt zusammenziehbar*) if there exists a smooth homotopy of the identity map $M \to M$ to a constant map. It is easy to see for instance that $\mathbb{R}^n$ is smoothly contractible, and so is any convex subset of $\mathbb{R}^n$. Given a smooth homotopy $h : [0,1] \times M \to M$ with $h(1,\cdot) = \mathrm{Id}_M$ and $h(0,\cdot) \equiv p \in M$ for some fixed point $p \in M$, consider the maps

$$\pi : M \to \{p\}, \qquad i : \{p\} \hookrightarrow M,$$

where π is the unique map and i is the natural inclusion. Now $\pi \circ i$ is the identity map on $\{p\}$, and $i \circ \pi : M \to M$ is $h(0,\cdot)$, which is therefore smoothly homotopic to Id_M. This proves that M is smoothly homotopy equivalent to the one-point manifold $\{p\}$, so combining Corollary 13.33 with Example 13.28 gives:

COROLLARY 13.34. *If M is smoothly contractible, then $H^k_{\mathrm{dR}}(M) = 0$ for all $k > 0$ and $H^0_{\mathrm{dR}}(M) \cong \mathbb{R}$.*

PROOF OF THE POINCARÉ LEMMA. Every point $p \in M$ has a neighborhood $\mathcal{U} \subset M$ that looks like a convex set in some coordinate chart and is thus smoothly contractible. For $k > 0$, it now follows from $H^k_{\mathrm{dR}}(\mathcal{U}) = 0$ that the spaces of closed and exact k-forms on $\mathcal{U}$ are identical. □

PROOF OF THEOREM 13.32. We assume $h : [0,1] \times M \to N$ satisfies $h(0,\cdot) = f_0$ and $h(1,\cdot) = f_1$. Given $\omega \in \Omega^k(N)$, let us assume $L \subset M$ is a compact oriented k-dimensional submanifold with boundary and consider the integral of $h^* d\omega \in \Omega^{k+1}([0,1] \times M)$ over the domain $[0,1] \times L$. Note that the latter is not a smooth manifold with boundary unless $\partial L = \emptyset$; in general $[0,1] \times L$ can be understood as a manifold with boundary *and corners.* Nonetheless, one can make sense of Stokes' theorem on this domain as described in Example 12.14, leading to the relation

$$\begin{aligned} \int_{[0,1]\times L} h^*(d\omega) &= \int_{[0,1]\times L} d(h^*\omega) = \int_{\partial([0,1]\times L)} h^*\omega := \int_{\partial[0,1]\times L} h^*\omega - \int_{[0,1]\times\partial L} h^*\omega \\ &= \int_{\{1\}\times L} h^*\omega - \int_{\{0\}\times L} h^*\omega - \int_{[0,1]\times\partial L} h^*\omega \\ &= \int_L f_1^*\omega - \int_L f_0^*\omega - \int_{[0,1]\times\partial L} h^*\omega, \end{aligned} \tag{13.2}$$

where in the last line we have used the obvious identifications of $\{1\}\times L$ and $\{0\}\times L$ with L, so that the restrictions of $h^*\omega$ to these two submanifolds become $f_1^*\omega$ and $f_0^*\omega$ respectively. Now observe that for any compact oriented m-dimensional submanifold $Q \subset M$ and an $(m+1)$-form $\alpha \in \Omega^{m+1}(N)$, there is a natural way of presenting $\int_{[0,1]\times Q} h^*\alpha$ as the integral of an m-form over Q: we define $P\alpha \in \Omega^m(M)$ namely via the formula

$$(P\alpha)_p(X_1,\dots,X_m) := \int_0^1 (h^*\alpha)_{(t,p)}(\partial_t, X_1,\dots,X_m)\,dt \in \mathbb{R},$$

where ∂_t here denotes the obvious unit vector field on $[0,1]\times M$ pointing in the positive direction on the first factor, and each $X_1,\dots,X_m \in T_pM$ is regarded as living in the subspace $\{0\}\times T_pM \subset T_t[0,1]\times T_pM = T_{(t,p)}([0,1]\times M)$. In this way we have defined a linear operator

$$P : \Omega^{m+1}(N) \to \Omega^m(M) \qquad \text{such that} \qquad \int_{[0,1]\times Q} h^*\alpha = \int_Q P\alpha$$

for all $\alpha \in \Omega^{m+1}(N)$ and compact oriented m-dimensional submanifolds $Q \subset M$. We can use this to transform (13.2) into the relation

$$\int_L (f_1^*\omega - f_0^*\omega) = \int_L P(d\omega) + \int_{\partial L} P\omega = \int_L \left[P(d\omega) + d(P\omega)\right],$$

where we have again applied Stokes' theorem to transform the integral over ∂L into one over L. We now have an equality of the integrals of two k-forms over an arbitrary compact oriented k-dimensional submanifold with boundary: in particular, one could pick any point $p \in M$ and any vectors $X_1,\dots,X_k \in T_pM$ and then approximate the evaluation of both k-forms on $(X_1,\dots,X_k)$ arbitrarily well by integrating them over a submanifold L that is chosen to be a small k-disk through p tangent to the space spanned by $X_1,\dots,X_k$. The conclusion is that these two k-forms must be identical, so we have proved that $f_1^*\omega - f_0^*\omega = P(d\omega) + d(P\omega)$, or rewriting it as an equality between two linear maps $H^k_{\mathrm{dR}}(N) \to H^k_{\mathrm{dR}}(M)$,

$$f_1^* - f_0^* = P\circ d + d\circ P. \tag{13.3}$$

This formula is well known in homological algebra: it is called the **chain homotopy relation**, and the operator $P : \Omega^*(N) \to \Omega^*(M)$ of degree -1 is consequently called a **chain homotopy** (*Kettenhomotopie*). Its existence has the following consequence: if $\omega \in \Omega^k(N)$ is closed, then

$$f_1^*\omega = f_0^*\omega + d(P\omega),$$

implying that $f_1^*\omega$ and $f_0^*\omega$ represent the same element in the quotient $H^k_{\mathrm{dR}}(M)$. □

EXERCISE 13.35. Suppose $\mathcal{O}$ is an open subset of either $\mathbb{H}^n$ or $\mathbb{R}^n$. We call $\mathcal{O}$ a **star-shaped** domain if for every $p \in \mathcal{O}$, it also contains the points $tp \in \mathbb{R}^n$ for all $t \in [0,1]$. It follows that $h(t,p) := tp$ defines a smooth homotopy $h : [0,1]\times\mathcal{O} \to \mathcal{O}$ between the identity and the constant map whose value is the origin, making $\mathcal{O}$ smoothly contractible. Use this homotopy to extract from the proof of Theorem 13.32 an explicit formula for a linear operator $P : \Omega^k(\mathcal{O}) \to \Omega^{k-1}(\mathcal{O})$ for each $k \geqslant 1$ satisfying

$$\omega = P(d\omega) + d(P\omega)$$

for all $\omega \in \Omega^k(\mathcal{O})$. In particular, whenever ω is a closed k-form, $P\omega$ is a primitive of ω. (As a sanity check, a formula for P is given in Remark 13.39 at the end of this lecture, but try to derive it without knowing it in advance.)

One further property of $H^*_{\mathrm{dR}}(M)$ deserves to be mentioned, though a full explanation of it would fall far outside the scope of this course. By a result known as *de Rham's theorem*, $H^k_{\mathrm{dR}}(M)$ is naturally isomorphic to another invariant that is a standard topic in algebraic topology, namely the kth *singular cohomology* with real coefficients:

$$H^k_{\mathrm{dR}}(M) \cong H^k(M; \mathbb{R}).$$

The latter is defined for all topological spaces, not just smooth manifolds, and spaces that are *homeomorphic* always have isomorphic singular cohomologies, implying that $H^k_{\mathrm{dR}}(M)$ is actually a topological invariant. The topological invariance of $H^*_{\mathrm{dR}}(M)$ cannot be seen directly from its definition, since pullbacks of differential forms via maps $f : M \to N$ do not make sense in general when f is continuous but not differentiable. As one learns in algebraic topology, $H^k(M; \mathbb{R})$ is often surprisingly easy to compute, and for instance when M is compact, it can be derived from a finite-dimensional chain complex, implying the highly non-obvious fact that

$$\dim H^k_{\mathrm{dR}}(M) < \infty$$

whenever M is compact.

EXERCISE 13.36. Here is the most basic computation of $H^*_{\mathrm{dR}}(M)$ for a non-contractible manifold: we will show in this exercise that for every $n \in \mathbb{N}$ and $k \in \{0, \dots, n\}$,

$$\dim H^k_{\mathrm{dR}}(S^n) = \begin{cases} 1 & \text{if } k = 0 \text{ or } k = n, \\ 0 & \text{otherwise.} \end{cases} \tag{13.4}$$

Clearly every sphere S^n for $n \geqslant 1$ is connected,[48] so Example 13.27 establishes $H^0_{\mathrm{dR}}(S^n) \cong \mathbb{R}$. For the computation of $H^k_{\mathrm{dR}}(S^n)$ when $k \geqslant 1$, we proceed by induction on n.

(a) Show that if M is a closed oriented n-manifold, then there is a well-defined linear map

$$H^n_{\mathrm{dR}}(M) \to \mathbb{R} : [\omega] \mapsto \int_M \omega, \tag{13.5}$$

and the following conditions are equivalent:
  (i) $H^n_{\mathrm{dR}}(M) \cong \mathbb{R}$;
  (ii) The map (13.5) is an isomorphism;
  (iii) Every $\omega \in \Omega^n(M)$ satisfying $\int_M \omega = 0$ is exact.

(b) Deduce via Exercise 13.16 that (13.4) is correct for $n = 1$.

(c) Suppose M is a closed n-manifold and ω_+, ω_- is a pair of k-forms on $M \times [-1, 1]$ such that $d\omega_+ = d\omega_-$. Show that the following conditions are equivalent:
  (i) $\omega_+ - \omega_-$ is exact;
  (ii) $i_t^*\omega_+ - i_t^*\omega_-$ is an exact k-form on M for every $t \in [-1, 1]$, where $i_t : M \hookrightarrow M \times [-1, 1]$ denotes the inclusion $p \mapsto (p, t)$.
  (iii) There exists a k-form ω on $M \times [-1, 1]$ which matches $\omega_\pm$ near $M \times \{\pm 1\}$ and satisfies $d\omega = d\omega_+ = d\omega_-$.

[48] The 0-sphere is a discrete set of two points $S^0 = \{1, -1\} \subset \mathbb{R}$, and is thus not connected. That's why we excluded the case $n = 0$ from (13.4).

Hint: First prove the equivalence of (i) and (ii), after convincing yourself that $i_t : M \hookrightarrow M \times [-1, 1]$ is a smooth homotopy equivalence for each t.

(d) Under the same assumptions as in part (c), suppose also that M is oriented and $k = n$. Show that the number $\int_{M \times \{t\}} \omega_+ - \int_{M \times \{t\}} \omega_- \in \mathbb{R}$ is the same for any choice of $t \in [-1, 1]$.
Hint: Given $-1 \leqslant t_- < t_+ \leqslant 1$, integrate something over $M \times [t_-, t_+]$ and apply Stokes' theorem.

(e) Now given an integer $n \geqslant 2$, assume (13.4) is true for S^{n-1}, and fix $k \in \{1, \ldots, n\}$. Regarding S^n as the unit sphere in $\mathbb{R}^{n+1}$ with standard coordinates $(x^1, \ldots, x^{n+1})$, we can decompose it into two overlapping n-dimensional disks $S^n = D_+ \cup D_-$ whose intersection looks like $S^{n-1} \times [-1, 1]$; specifically, define

$$D_+ := \{x^1 \geqslant -1/2\} \cap S^n, \qquad D_- := \{x^1 \leqslant 1/2\} \cap S^n.$$

Take a moment to convince yourself that there is a diffeomorphism $D_+ \cap D_- \cong S^{n-1} \times [-1, 1]$. Observe next that D_+ and D_- are each smoothly contractible, thus any closed k-form ω on S^n will then by exact over each of D_+ and D_-, giving $\alpha_\pm \in \Omega^{k-1}(D_\pm)$ such that $d\alpha_\pm = \omega$ on $D_\pm$. The difficulty is that α_+ and α_- need not match on $D_+ \cap D_-$. Use the inductive hypothesis and the previous steps in this problem to show that if either $1 \leqslant k \leqslant n-1$ or $k = n$ with $\int_{S^n} \omega = 0$, then there exists $\alpha \in \Omega^{k-1}(S^n)$ satisfying $d\alpha = \omega$; show in fact that α can be chosen to match $\alpha_\pm$ on the portions of $D_\pm$ where D_+ and D_- do not overlap. This completes the inductive proof of (13.4).
Hint: The case $k = n$ is trickiest, as you need to use the hypothesis $\int_{S^n} \omega = 0$ to deduce something about α_+ and α_-. What can you say about the integrals of $\alpha_\pm$ over the "equator" $S^{n-1} \cong \{x^1 = 0\} \subset S^n$? Try Stokes' theorem, but be careful with orientations!

Exercise 13.37. Show that the wedge product descends to an associative and graded-commutative product $\cup : H^k_{\mathrm{dR}}(M) \times H^\ell_{\mathrm{dR}}(M) \to H^{k+\ell}_{\mathrm{dR}}(M)$, defined by

$$[\alpha] \cup [\beta] := [\alpha \wedge \beta].$$

This is called the **cup product** on de Rham cohomology.
Remark: There is similarly a cup product on singular cohomology, to which this one is isomorphic via de Rham's theorem. But this one is easier to define, and is thus often used in practice as a surrogate for the singular cup product.

Exercise 13.38. For this exercise, identify the n-torus $\mathbb{T}^n$ with the quotient $\mathbb{R}^n/\mathbb{Z}^n$ (recall from Exercise 3.4 that there is a natural diffeomorphism). For any sufficiently small open set $\widetilde{\mathcal{U}} \subset \mathbb{R}^n$, the usual Cartesian coordinates $x^1, \ldots, x^n : \widetilde{\mathcal{U}} \to \mathbb{R}$ can be used to define a smooth chart $(\mathcal{U}, x)$ on $\mathbb{T}^n$ where

$$\mathcal{U} := \left\{ [p] \in \mathbb{T}^n \;\middle|\; p \in \widetilde{\mathcal{U}} \right\}, \qquad x([p]) := (x^1(p), \ldots, x^n(p)) \text{ for } p \in \widetilde{\mathcal{U}}.$$

(a) Show that the coordinate differentials $dx^1, \ldots, dx^n \in \Omega^1(\mathcal{U})$ arising from the chart $(\mathcal{U}, x)$ described above are independent of the choice of the set $\widetilde{\mathcal{U}} \subset \mathbb{R}^n$, i.e. the definitions of the coordinate differentials obtained from two different choices $\widetilde{\mathcal{U}}_1, \widetilde{\mathcal{U}}_2 \subset \mathbb{R}^n$ coincide on the region $\mathcal{U}_1 \cap \mathcal{U}_2 \subset \mathbb{T}^n$ where they overlap.

(b) As a consequence of part (a), the 1-forms $dx^1, \ldots, dx^n \in \Omega^1(\mathbb{T}^n)$ are well-defined on the entire torus, and they are obviously locally exact and therefore closed, but they might not actually be exact since none of the coordinates $x^1, \ldots, x^n$ admit smooth definitions globally on $\mathbb{T}^n$. (This is another example of the phenomenon we saw with $d\theta \in \Omega^1(\mathbb{R}^2 \backslash \{0\})$ in Remark 13.13.) Show in fact that for any vector $(a_1, \ldots, a_n) \in \mathbb{R}^n \backslash \{0\}$, the 1-form

$$\lambda := a_i \, dx^i \in \Omega^1(\mathbb{T}^n)$$

is closed but not exact.
Hint: You only need to find one smooth map $\gamma : S^1 \to \mathbb{T}^n$ such that $\int_{S^1} \gamma^*\lambda \neq 0$.

(c) One can similarly produce closed k-forms $\omega \in \Omega^k(\mathbb{T}^n)$ for any $k \leqslant n$ by choosing constants $a_{i_1\dots i_k} \in \mathbb{R}$ and writing

$$\omega = \sum_{i_1<\dots<i_k} a_{i_1\dots i_k}\, dx^{i_1} \wedge \dots \wedge dx^{i_k} \in \Omega^k(\mathbb{T}^n). \tag{13.6}$$

Show that for every nontrivial k-form of this type, one can find a cohomology class $[\alpha] \in H^{n-k}_{\mathrm{dR}}(\mathbb{T}^n)$ such that the cup product $[\omega] \cup [\alpha] \in H^n_{\mathrm{dR}}(\mathbb{T}^n)$ defined in Exercise 13.37 is nontrivial, and deduce from this that ω is not exact.
Hint: Can you choose $\alpha \in \Omega^{n-k}(\mathbb{T}^n)$ so that $\omega \wedge \alpha$ is a volume form?

Remark: One can show that all cohomology classes in $H^k_{\mathrm{dR}}(\mathbb{T}^n)$ are representable by k-forms with constant coefficients as in (13.6), thus $\dim H^k_{\mathrm{dR}}(\mathbb{T}^n) = \binom{n}{k}$.

REMARK 13.39. Here is a formula for the operator $P : \Omega^k(\mathcal{O}) \to \Omega^{k-1}(\mathcal{O})$ promised in Exercise 13.35 on a star-shaped domain $\mathcal{O}$ in $\mathbb{H}^n$ or $\mathbb{R}^n$:

$$(P\omega)_p(X_1, \dots, X_{k-1}) := \int_0^1 t^{k-1}\omega_{tp}(p, X_1, \dots, X_{k-1})\, dt,$$

where since $\mathcal{O}$ is a subset of $\mathbb{R}^n$, we are using the natural isomorphisms $T_p\mathcal{O} = \mathbb{R}^n$ at every point. (Otherwise the expression $\omega_{tp}(p, X_1, \dots, X_{k-1})$ would not generally make sense because $X_1, \dots, X_{k-1} \in T_p\mathcal{O} \neq T_{tp}\mathcal{O}$.) In applications, it is occasionally useful to observe that $P\omega$ depends *continuously* on ω, i.e. one obtains in this way a continuous right-inverse of the operator $d_{k-1} : \Omega^{k-1}(\mathcal{O}) \to \mathrm{im}(d_{k-1}) \subset \Omega^k(\mathcal{O})$.

14. Volume-preserving and symplectic maps

14.1. Volume-preserving flows. Assume M is an oriented n-manifold with a fixed positive volume form $d\mathrm{vol} \in \Omega^n(M)$. In §12.5, we defined the divergence of a vector field $X \in \mathfrak{X}(M)$ in this context as the unique function $\mathrm{div}(X) : M \to \mathbb{R}$ such that

$$d(\iota_X d\mathrm{vol}) = \mathrm{div}(X) \cdot d\mathrm{vol}.$$

A partial justification for this definition was furnished by the Gauss divergence theorem,

$$\int_M \mathrm{div}(X)\, d\mathrm{vol}_M = \int_{\partial M} \langle X, \nu \rangle\, d\mathrm{vol}_{\partial M}, \tag{14.1}$$

a corollary of Stokes' theorem that equates the total divergence of a vector field on a Riemannian manifold with boundary to its total *flux* through the boundary (see §12.5.1). We would now like to explain a more fundamental interpretation of the divergence: it measures the extent to which the flow of X changes volume.

Writing $\mathrm{Vol}(A) := \int_A d\mathrm{vol}$, a diffeomorphism $\varphi : M \to M$ is called **volume preserving** if

$$\mathrm{Vol}(\varphi(A)) = \mathrm{Vol}(A) \qquad \text{for all measurable sets } A \subset M.$$

For a vector field $X \in \mathfrak{X}(M)$ admitting a global flow, we say that its flow is volume preserving if φ^t_X is volume preserving for every $t \in \mathbb{R}$. Without assuming there is a global flow, this condition can still be generalized as follows: for every measurable set $A \subset M$ and every $t \in \mathbb{R}$ for which the domain of φ^t_X contains A, $\mathrm{Vol}(\varphi^t_X(A)) = \mathrm{Vol}(A)$. Note that if A has compact closure, then this condition always makes sense at least for t close to 0. For simplicity we will assume in the following discussion that there is always a global flow, but this condition can be lifted by paying more careful attention to the domains of the flow maps φ^t_X.

The diffeomorphisms $\varphi_X^t : M \to M$ defined via the flow of a vector field are always orientation preserving—this results from the fact that $\varphi_X^0 : M \to M$ is the identity map, so for any $p \in M$, any positively oriented basis $Y_1, \dots, Y_n$ of T_pM gives rise to a continuous 1-parameter family of bases

$$(T\varphi_X^t(Y_1), \dots, T\varphi_X^t(Y_n))$$

for the tangent spaces $T_{\varphi_X^t(p)}M$, and continuity dictates that they must all be positively oriented. We therefore have

$$\mathrm{Vol}(\varphi_X^t(A)) = \int_{\varphi_X^t(A)} d\mathrm{vol} = \int_A (\varphi_X^t)^* d\mathrm{vol}$$

for every $A \subset M$, and the rate of change of this volume is

$$\frac{d}{dt}\mathrm{Vol}(\varphi_X^t(A)) = \frac{d}{dt}\int_A (\varphi_X^t)^* d\mathrm{vol} = \int_A \partial_t(\varphi_X^t)^* d\mathrm{vol}. \tag{14.2}$$

The next step in the calculation works in more general contexts: in place of the volume form $d$vol, we can consider an arbitrary tensor field $S \in \Gamma(T^k_\ell M)$. Recall that $\varphi_X^{s+t} = \varphi_X^s \circ \varphi_X^t$, thus $(\varphi_X^{s+t})^* = (\varphi_X^t)^*(\varphi_X^s)^*$, and

$$\begin{aligned} \partial_t(\varphi_X^t)^* S &= \partial_s(\varphi_X^{s+t})^* S\big|_{s=0} = \partial_s(\varphi_X^t)^*(\varphi_X^s)^* S\big|_{s=0} \\ &= (\varphi_X^t)^*\left(\partial_s(\varphi_X^s)^* S\big|_{s=0}\right) = (\varphi_X^t)^*\left(\mathcal{L}_X S\right). \end{aligned} \tag{14.3}$$

Applying this to (14.2) gives

$$\frac{d}{dt}\mathrm{Vol}(\varphi_X^t(A)) = \int_A (\varphi_X^t)^*\left(\mathcal{L}_X d\mathrm{vol}\right) = \int_{\varphi_X^t(A)} \mathcal{L}_X d\mathrm{vol}.$$

It follows that the flow is volume preserving if the Lie derivative of the volume form $d$vol with respect to X vanishes, and conversely, the derivative of $\mathrm{Vol}(\varphi_X^t(A))$ can only vanish for every measurable set $A \subset M$ if the n-form $(\varphi_X^t)^*(\mathcal{L}_X d\mathrm{vol})$ vanishes identically for every t, which is equivalent to the condition $\mathcal{L}_X d\mathrm{vol} \equiv 0$ since $(\varphi_X^t)^* : \Omega^n(M) \to \Omega^n(M)$ is a bijection.

LEMMA 14.1. *For any volume form* $d\mathrm{vol} \in \Omega^n(M)$ *and vector field* $X \in \mathfrak{X}(M)$*,*

$$\mathcal{L}_X d\mathrm{vol} = d(\iota_X d\mathrm{vol}).$$

This relation will follow from the more general formula of Cartan for Lie derivatives of differential forms, to be proved in the next section. We can now alternatively characterize the divergence of X as the unique function such that

$$\mathcal{L}_X d\mathrm{vol} = \mathrm{div}(X) \cdot d\mathrm{vol}, \tag{14.4}$$

and the discussion above implies:

THEOREM 14.2. *On a manifold* M *with volume form* $d$vol*, a vector field* $X \in \mathfrak{X}(M)$ *has a volume-preserving flow if and only if* $\mathrm{div}(X) \equiv 0$. ☐

The divergence theorem (14.1) now admits a new geometric interpretation whenever M is a compact submanifold with boundary in a larger n-manifold N on which the vector field X and volume form $d$vol are defined. In this case, the flow φ_X^t of X is well defined on M for all t sufficiently close to zero, and the left hand side of (14.1) then becomes

$$\begin{aligned} \int_M \mathrm{div}(X)\, d\mathrm{vol}_N &= \int_M \mathcal{L}_X(d\mathrm{vol}_N) = \frac{d}{dt}\int_M (\varphi_X^t)^* d\mathrm{vol}_N\bigg|_{t=0} = \frac{d}{dt}\int_{\varphi_X^t(M)} d\mathrm{vol}_N\bigg|_{t=0} \\ &= \frac{d}{dt}\mathrm{Vol}(\varphi_X^t(M))\bigg|_{t=0}. \end{aligned}$$

The divergence theorem thus relates the rate of change of the volume of M under the flow of X to the average of $\langle X, \nu\rangle$ along ∂M, which measures the extent to which X flows out of M vs. into M through its boundary.

14.2. Cartan's formula for the Lie derivative. The following practical tool for computing Lie derivatives of forms is sometimes called *Cartan's magic formula*.

THEOREM 14.3. *For any $X \in \mathfrak{X}(M)$ and $\omega \in \Omega^k(M)$,*

$$\mathcal{L}_X \omega = d(\iota_X \omega) + \iota_X(d\omega).$$

An immediate application is Lemma 14.1 above: if $d\mathrm{vol} \in \Omega^n(M)$ is a volume form, then

$$\mathcal{L}_X d\mathrm{vol} = d(\iota_X d\mathrm{vol}) + \iota_X d(d\mathrm{vol}) = d(\iota_X d\mathrm{vol})$$

since $d(d\mathrm{vol})$ is an $(n+1)$-form on an n-manifold and therefore vanishes.[49]

The following sequence of exercises sums up to a proof of Cartan's formula, the idea behind it being to show that for any given $X \in \mathfrak{X}(M)$, both of the operators $\mathcal{L}_X$ and $d\iota_X + \iota_X d$ define derivations on the exterior algebra $\Omega^*(M)$ that match when applied to functions or differentials of functions. This is sufficient for the same reason that a few formal properties centered around the graded Leibniz rule sufficed in Proposition 9.16 for characterizing the exterior derivaive: both are clearly local operators, and locally, every differential form is a finite sum of wedge products of functions and differentials.

EXERCISE 14.4 (easy). Show that Theorem 14.3 holds for all $\omega = f \in C^\infty(M) = \Omega^0(M)$.

LEMMA 14.5. *Theorem 14.3 holds for all $\omega = df \in \Omega^1(M)$ with $f \in C^\infty(M)$.*

PROOF. Since $d^2 = 0$, $d\iota_X df + \iota_X d(df) = d(\iota_X df)$, where $\iota_X df$ is the real-valued function $p \mapsto df(X(p))$. To evaluate $\mathcal{L}_X(df) \in \Omega^1(M)$ on some $Y \in T_pM$ at a point $p \in M$, choose a smooth path $\gamma : (-\epsilon, \epsilon) \to M$ with $\gamma(0) = p$ and $\dot\gamma(0) = Y$. Then using Proposition 9.18,

$$\begin{aligned}\mathcal{L}_X(df)(Y) &= \partial_t(\varphi_X^t)^*(df)(Y)\big|_{t=0} = \partial_t d(f \circ \varphi_X^t)(Y)\big|_{t=0} = \partial_t\partial_s f(\varphi_X^t(\gamma(s)))\big|_{s=t=0} \\ &= \partial_s\partial_t f(\varphi_X^t(\gamma(s)))\big|_{s=t=0} = \partial_s df(X(\gamma(s))|_{s=0} = \partial_s \iota_X(df)(\gamma(s))\big|_{s=0} = d(\iota_X df)(Y).\end{aligned}$$

□

The next exercise follows also quite easily from the definition of the Lie derivative, plus Proposition 9.18 and the fact that the wedge product is bilinear. Notice that in contrast to the exterior derivative, no annoying sign appears in the Leibniz rule for $\mathcal{L}_X$. Formally, the reason is because $\mathcal{L}_X$ sends k-forms to k-forms for each $k \geqslant 0$, and is thus an operator of "degree 0", i.e. it is even, while the exterior derivative is odd.

EXERCISE 14.6. Show that $\mathcal{L}_X : \Omega^*(M) \to \Omega^*(M)$ is a derivation with respect to the wedge product, meaning

$$\mathcal{L}_X(\alpha \wedge \beta) = \mathcal{L}_X\alpha \wedge \beta + \alpha \wedge \mathcal{L}_X\beta.$$

We now turn our attention fully to the operator

$$P_X := d\iota_X + \iota_X d : \Omega^*(M) \to \Omega^*(M), \tag{14.5}$$

in which each term is a composition of operators with degrees 1 and -1, so P_X itself also has degree 0. We've seen already that d satisfies a graded Leibniz rule; it turns out that ι_X does as well:

[49] Here is another cautionary reminder about the oddity of our notation for volume forms: we have not defined any $(n-1)$-form "$\mathrm{vol} \in \Omega^{n-1}(M)$" for $d\mathrm{vol}$ to be the exterior derivative of, and we have seen for instance that when M is a closed manifold, $d\mathrm{vol}$ is definitely not the exterior derivative of anything. The vanishing of $d(d\mathrm{vol})$ thus has nothing to do with the relation $d \circ d = 0$; it vanishes for a completely different reason.

EXERCISE 14.7. For V an n-dimensional vector space, the goal of this exercise is to show that for every $v \in V$, the operator $\iota_v : \Lambda^* V^* \to \Lambda^* V^*$ satisfies the graded Leibniz rule

$$\iota_v(\alpha \wedge \beta) = (\iota_v \alpha) \wedge \beta + (-1)^k \alpha \wedge (\iota_v \beta) \tag{14.6}$$

for all $\alpha \in \Lambda^k V^*$ and $\beta \in \Lambda^\ell V^*$. The statement is trivial if $v = 0$, so assume otherwise, in which case we may as well assume v is the first element e_1 of a basis $e_1, \ldots, e_n \in V$, whose dual basis we can denote by $e_*^1, \ldots, e_*^n \in V^* = \Lambda^1 V^*$.

(a) Prove that (14.6) holds whenever α and β are both products of the form $\alpha = e_*^{i_1} \wedge \ldots \wedge e_*^{i_k}$ and $\beta = e_*^{j_1} \wedge \ldots \wedge e_*^{j_\ell}$ with $i_1 < \ldots < i_k$ and $j_1 < \ldots < j_\ell$.
Hint: Consider separately a short list of cases depending on whether each of i_1 and j_1 are 1 and whether the sets $\{i_1, \ldots, i_k\}$ and $\{j_1, \ldots, j_\ell\}$ are disjoint.

(b) Deduce via linearity that (14.6) holds always.

EXERCISE 14.8. Prove that the operator P_X in (14.5) is also a derivation on $\Omega^*(M)$, and deduce that $P_X = \mathcal{L}_X$, thus proving Theorem 14.3.

14.3. Symplectic manifolds and Hamiltonian systems. Volume-preserving flows arise naturally in the context of Hamiltonian systems, a special class of dynamical systems that originate in classical mechanics. From a mathematical perspective, the most natural language for this discussion is that of *symplectic* geometry.

DEFINITION 14.9. Assume M is a smooth manifold of even dimension $2n$ for some $n \in \mathbb{N}$. A 2-form $\omega \in \Omega^2(M)$ is called **symplectic** (*symplektisch*) if every point $x \in M$ admits a neighborhood $x \in \mathcal{U} \subset M$ with a coordinate chart of the form $(\mathcal{U}, (p^1, q^1, \ldots, p^n, q^n))$ such that

$$\omega = \sum_{j=1}^n dp^j \wedge dq^j \qquad \text{on } \mathcal{U}. \tag{14.7}$$

A 2-form with this property is also sometimes called a **symplectic structure** (*symplektische Struktur*) on M, and the pair (M, ω) in this situation is called a **symplectic manifold** (*symplektische Mannigfaltigkeit*).

Observe that the coordinates $(p^1, q^1, \ldots, p^n, q^n)$ appearing in (14.7) are special; it would certainly be impossible to demand that any 2-form satisfy (14.7) for *every* choice of chart, but the definition only requires the existence of *some* chart near every point so that ω takes this form. In this sense, a symplectic structure is somewhat analogous to an orientation: it is equivalent in fact to a maximal atlas of compatible charts in which the word "compatible" has been given a new and much stricter definition, requiring all transition maps to not only be smooth but also to preserve the relation (14.7). Physicists sometimes refer to coordinates $(p^1, q^1, \ldots, p^n, q^n)$ of this type as *canonical coordinates* and call the corresponding transition maps *canonical transformations*. Mathematicians prefer to call them *Darboux coordinates*, after Darboux's theorem (see Remark 14.11 below).

EXERCISE 14.10. Show that a symplectic form $\omega \in \Omega^2(M)$ always has the following properties:

(a) ω is closed: $d\omega = 0$.

(b) For every $x \in M$, the linear map $T_x M \to T_x^* M : X \mapsto \omega(X, \cdot)$ is an isomorphism. (Bilinear forms with this property are called **nondegenerate**).

(c) The "top" exterior power of ω,

$$\omega^n := \underbrace{\omega \wedge \ldots \wedge \omega}_{n} \in \Omega^{2n}(M)$$

is a volume form on M. It follows in particular that M is orientable.

(d) Is M is closed, then ω represents a nontrivial cohomology class $[\omega] \in H^2_{\mathrm{dR}}(M)$.
Hint: Recall the cup product from Exercise 13.37. What can you say about the n-fold cup product of $[\omega]$ with itself?

Remark 14.11. A fundamental result known as *Darboux's theorem* says that symplectic forms can in fact be characterized fully in terms of the first two properties in Exercise 14.10, i.e. every 2-form that is both closed and nondegenerate admits an atlas of charts satisfying (14.7) and is thus a symplectic form. This reveals for instance that every volume form on a surface[50] is a symplectic form. We will not make use of these facts here, but it is important to be aware of them since most textbooks prefer to *define* the term "symplectic form" to mean a closed and nondegenerate 2-form.

Given a smooth function $H : M \to \mathbb{R}$ on a symplectic manifold (M, ω), the nondegeneracy of ω implies that there is a unique vector field $X_H \in \mathfrak{X}(M)$ satisfying

$$\omega(X_H, \cdot) = -dH \in \Omega^1(M). \tag{14.8}$$

We call X_H the **Hamiltonian vector field** determined by H, and in this context, the function H itself is often called a **Hamiltonian**. In Darboux coordinates, it is not hard to derive an explicit formula for the Hamiltonian vector field: writing $X_H = A^j \frac{\partial}{\partial q^j} + B^j \frac{\partial}{\partial p^j}$, we find

$$\begin{aligned} dH = \frac{\partial H}{\partial q^j}\, dq^j + \frac{\partial H}{\partial p^j}\, dp^j &= -\omega(X_H, \cdot) = -\sum_{i=1}^n (dp^i \wedge dq^i)\left(A^j \frac{\partial}{\partial q^j} + B^j \frac{\partial}{\partial p^j}, \cdot\right) \\ &= \sum_{i=1}^n \left(-B^i\, dq^i + A^i\, dp^i\right), \end{aligned}$$

implying

$$X_H = \sum_{i=1}^n \left(\frac{\partial H}{\partial p^i}\frac{\partial}{\partial q^i} - \frac{\partial H}{\partial q^i}\frac{\partial}{\partial p^i}\right). \tag{14.9}$$

In other words, if $x(t) \in M$ denotes a smooth path passing through the domain of a Darboux chart and its coordinates in this chart at time t are written as $(p^1(t), q^1(t), \dots, p^n(t), q^n(t))$, then x is an orbit of X_H if and only if its coordinates satisfy the following system of $2n$ first-order ODEs:

$$\dot{q}^i(t) = \frac{\partial H}{\partial p^i}(x(t)), \qquad \dot{p}^i(t) = -\frac{\partial H}{\partial q^i}(x(t)) \qquad i = 1, \dots, n. \tag{14.10}$$

This system is known as *Hamilton's equations*, and the dynamical system defined by the flow of X_H is called a *Hamiltonian system.*

The study of Hamiltonian systems originates with the following example.

Example 14.12. In classical mechanics, the motion in $\mathbb{R}^3$ of a single particle with mass $m > 0$ under the influence of a force is described by a path $\mathbf{q}(t) = (q^1(t), q^2(t), q^3(t)) \in \mathbb{R}^3$ that obeys Newton's second law,

$$\mathbf{F}(\mathbf{q}(t)) = m\ddot{\mathbf{q}}(t),$$

where $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^3$ is a vector field representing the force. In standard examples, $\mathbf{F}$ is determined by a *potential* $V : \mathbb{R}^3 \to \mathbb{R}$ via the relation

$$\mathbf{F} = -\nabla V,$$

[50] On a manifold of dimension 2, it is also common to refer to volume forms as **area forms**.

hence the individual coordinates satisfy $m\ddot{q}^i(t) = -\frac{\partial V}{\partial q^i}(\mathbf{q}(t))$. There is a popular trick for turning second-order systems of ODEs like this one into first-order systems with twice as many degrees of freedom: we associate to the position variables q^1, q^2, q^3 the corresponding *momentum* variables

$$p^i(t) := m\dot{q}^i(t), \qquad \mathbf{p} := (p^1, p^2, p^3)$$

and observe that the path $(\mathbf{q}(t), \mathbf{p}(t)) \in \mathbb{R}^6$ now satisfies the first-order system of equations

$$\dot{q}^i(t) = \frac{1}{m}p^i(t), \qquad \dot{p}^i(t) = -\frac{\partial V}{\partial q^i}(\mathbf{q}(t)), \qquad i = 1, 2, 3.$$

As it happens, this is the Hamiltonian system determined by the function $H : \mathbb{R}^6 \to \mathbb{R}$ given by

$$H(\mathbf{q}, \mathbf{p}) := \frac{|\mathbf{p}|^2}{2m} + V(\mathbf{q}).$$

Rewriting this as a function of $\mathbf{q}$ and $\dot{\mathbf{q}} := \frac{1}{m}\mathbf{p}$, the first term becomes $\frac{1}{2}m|\dot{\mathbf{q}}|^2$, which physicists call the *kinetic energy* of the moving particle. This is summed with the potential energy $V(\mathbf{q})$ to produce the Hamiltonian, which therefore has an interpretation as the *total energy* of the particle.

The Hamiltonian formalism lends itself to generalization: to turn the example above into a system of $N > 1$ moving particles, one can package the coordinates of all particles together to form a path in $\mathbb{R}^{3N}$, define corresponding momenta to produce a path in the so-called **phase space** $\mathbb{R}^{6N}$, write the total energy of the system as a function of all its position and momentum variables, and then write down Hamilton's equations (14.10). More generally, one can consider systems with constraints that prevent their positions from moving freely in Euclidean space, but confine them instead to a submanifold. In this situation there might not exist any global coordinate system in which Hamilton's equations (14.10) make sense, but if we have a symplectic form and a Hamiltonian function, then (14.8) defines the Hamiltonian vector field in a way that is independent of coordinates. We will see for instance that on any n-dimensional Riemannian manifold, the geodesic equation can be identified with a Hamiltonian system on a manifold of dimension $2n$.

If you've wondered why we are discussing symplectic manifolds in the same lecture with volume-preserving flows, here is the reasons:

THEOREM 14.13 (Liouville's theorem). *For any symplectic manifold (M, ω) and Hamiltonian $H \in C^\infty(M)$, the flow of the Hamiltonian vector field X_H is volume preserving with respect to the volume form $\omega^n \in \Omega^{2n}(M)$.*

PROOF. Let's do two proofs. The first is a coordinate-based computation: in any Darboux chart on some region in M, ω^n becomes a constant multiple of the standard volume form

$$\omega^n = \left(\sum_{i_1=1}^n dp^{i_1} \wedge dq^{i_1}\right) \wedge \ldots \wedge \left(\sum_{i_n=1}^n dp^{i_n} \wedge dq^{i_n}\right) = n\, dp^1 \wedge dq^1 \wedge \ldots \wedge dp^n \wedge dq^n,$$

and according to Exercise 12.16 and (14.9), the divergence of X_H is thus

$$\operatorname{div}(X_H) = \sum_{i=1}^n \left(\frac{\partial}{\partial q^i}\frac{\partial H}{\partial p^i} - \frac{\partial}{\partial p^i}\frac{\partial H}{\partial q^i}\right) = 0.$$

The result now follows from Theorem 14.2.

The second proof is more elegant, because it does not require coordinates, and it also proves a stronger result. Using Cartan's formula and the defining property of the vector field X_H, we find

$$\mathcal{L}_{X_H}\omega = d(\iota_{X_H}\omega) + \iota_{X_H}(d\omega) = -d(dH) = 0.$$

It follows via (14.3) that the 2-forms $(\varphi^t_{X_H})^*\omega$ are independent of t, and thus

$$(\varphi^t_{X_H})^*\omega = \omega \text{ for all } t. \tag{14.11}$$

It follows that for each t, $\varphi := \varphi^t_{X_H}$ also preserves the volume form ω^n, since

$$\varphi^*(\omega \wedge \ldots \wedge \omega) = \varphi^*\omega \wedge \ldots \wedge \varphi^*\omega = \omega \wedge \ldots \wedge \omega. \tag{14.12}$$

$\square$

I mentioned that our second proof of Liouville's theorem actually proves a stronger result. On a symplectic manifold (M, ω), a diffeomorphism $\psi : M \to M$ that satisfies

$$\psi^*\omega = \omega$$

is called a **symplectomorphism** (*Symplektomorphismus*), which can be viewed as an abbreviation for **symplectic diffeomorphism**. We see from (14.11) that Hamiltonian flows $\varphi^t_{X_H}$ have this property for every t, and by (14.12), all symplectomorphisms are also volume preserving.

While the subject of symplectic geometry has existed since the beginning of the 20th century, it was unknown for many decades whether the condition of being a symplectomorphism is truly more restrictive than being volume preserving. The following answer to this question emerged in 1985 and opened up a whole new subfield of geometry, known as *symplectic topology*:

Theorem (Gromov's non-squeezing theorem [Gro85]). *Fix the global coordinates $(p^1, q^1, \ldots, p^n, q^n)$ on $\mathbb{R}^{2n}$ with the "standard" symplectic form $\omega = \sum_{i=1}^n dp^i \wedge dq^i$, and let $B^k_r \subset \mathbb{R}^k$ denote the open ball of radius r. Then for two constants $r, R > 0$, the $2n$-ball $B^{2n}_r \subset \mathbb{R}^{2n}$ is symplectomorphic to a subset of the "cylinder"*

$$Z^{2n}_R := B^2_R \times \mathbb{R}^{2n-2} \subset \mathbb{R}^{2n}$$

if and only if $r \leqslant R$.

This is a hard theorem; various proofs are known, but all of them require a substantial amount of analytical machinery which cannot be fit into an introductory course. The significance of the non-squeezing theorem is that if $n \geqslant 2$, then no matter how small $R > 0$ may be, the cylinder Z^{2n}_R contains unlimited space in $2n-2$ of its $2n$ dimensions, and it is never difficult to find a volume-preserving embedding $B^{2n}_r \hookrightarrow Z^{2n}_R$ that compresses the first two dimensions as much as needed while expanding the others to compensate. The fact that *symplectic* embeddings cannot do this when $R < r$ means that there are meaningful restrictions on symplectic maps beyond the requirement that they must preserve volume. That subject is still an active area of research today.

Exercise 14.14. In 1915, Emmy Noether established a beautiful correspondence between the conserved quantities of a mechanical system and its symmetries. A simple version of this theorem in the Hamiltonian context takes the following form. Assume (M, ω) is a symplectic manifold, and $H : M \to \mathbb{R}$ and $F : M \to \mathbb{R}$ are two functions such that the corresponding Hamiltonian vector fields X_H and X_F have global flows. We say that F is *conserved* under the flow of X_H if F is constant along every orbit of X_H.

(a) Show that F is conversed under the flow of X_H if and only if H is conserved under the flow of X_F.

(b) In some settings, there is a converse to the result proved in part (a). Suppose M is simply connected, and $Y \in \mathfrak{X}(M)$ is a vector field with a global flow that is symplectic and preserves H, i.e.

$$(\varphi^t_Y)^*\omega = \omega \qquad \text{and} \qquad H \circ \varphi^t_Y = H \tag{14.13}$$

for all t. One says in this situation that Y determines a *symmetry* of the Hamiltonian system on (M, ω) defined by H. Under these assumptions, show that there exists a function $F : M \to \mathbb{R}$, uniquely defined up to addition of a constant, such that $Y = X_F$, and F is then conserved under the flow of X_H.

Let's work out a concrete example. Let $M = \mathbb{R}^4$ with coordinates (p_x, x, p_y, y) and the standard symplectic form

$$\omega = dp_x \wedge dx + dp_y \wedge dy \in \Omega^2(\mathbb{R}^4).$$

We can think of $\mathbb{R}^4$ as the "position-momentum space" (also known as *phase space*) representing the motion of a single particle of mass $m > 0$ in a plane: its position is given by $\mathbf{q} := (x, y) \in \mathbb{R}^2$, and $\mathbf{p} := (p_x, p_y) \in \mathbb{R}^2$ are the corresponding momentum variables. Given a "potential" function $V : \mathbb{R}^2 \to \mathbb{R}$, the total energy of the system is given by the function $H : \mathbb{R}^4 \to \mathbb{R}$,

$$H = \frac{|\mathbf{p}|^2}{2m} + V(\mathbf{q}).$$

Suppose now that the potential V is chosen to be *rotationally symmetric*, e.g. this is the case if $\mathbf{q}$ represents the position of the Earth moving around the sun (with the latter positioned at the origin). To express this condition succinctly, one can transform to polar coordinates (r, θ) on a suitable subset of $\mathbb{R}^2$, related to the (x, y)-coordinates as usual by $x = r\cos\theta$ and $y = r\sin\theta$. The condition imposed on V is then $\partial_\theta V \equiv 0$.

(c) Regarding r and θ as real-valued functions on (a suitable subdomain of) $\mathbb{R}^4$ that depend on the coordinates x and y but not on p_x and p_y, define two additional functions on the same domain by

$$p_r := \frac{x}{r}p_x + \frac{y}{r}p_y, \qquad p_\theta := yp_x - xp_y.$$

Show that $(p_r, r, p_\theta, \theta)$ is then a Darboux chart for the symplectic form ω.
Hint: It suffices to compute ω in the new coordinates and show that it satisfies the right formula, but this computation is a bit long. You could make your life easier by observing that $\omega = d\lambda$ for $\lambda := p_x\, dx + p_y\, dy$, and then computing λ in the new coordinates.

(d) Write down H as a function of $(p_r, r, p_\theta, \theta)$ and show that the vector field $Y := \partial_\theta$ defined in these coordinates on $\mathbb{R}^4 \backslash \{r = 0\}$ satisfies (14.13). Derive a formula for the corresponding conserved quantity F as promised by part (b). It is called the angular momentum of the system.

15. Partitions of unity

In Lecture 11, we constructed partitions of unity subordinate to finite open covers of compact manifolds: more precisely, if $\{\mathcal{U}_\alpha\}_{\alpha \in I}$ is a finite collection of open sets in a manifold M whose union contains the compact subset $K \subset M$, then there exists an associated collection of smooth functions $\{\varphi_\alpha : M \to [0, 1]\}_{\alpha \in I}$ such that

$$\sum_{\alpha \in I} \varphi_\alpha \equiv 1 \text{ on } K, \qquad \text{and} \qquad \operatorname{supp}(\varphi_\alpha) \subset \mathcal{U}_\alpha \text{ is compact for every } \alpha \in I.$$

This was used in order to "localize" the problem of defining integrals $\int_A \omega$, and we used the same localization trick again to prove Stokes' theorem in Lecture 12. In this lecture, we will use a more general localization trick to prove that Riemannian metrics exist on all smooth manifolds M. Unless M happens to be compact, we will not be able to get away with considering only finite open covers or functions with compact support. We will therefore need a more general notion of partitions of unity and an extension of the previous construction. This turns out to be the point where one must finally make explicit use of the assumption that manifolds are metrizable.

15.1. Local finiteness. A collection of subsets $\{\mathcal{U}_\alpha \subset X\}_{\alpha \in I}$ in a topological space X is called **locally finite** if every point $p \in X$ has a neighborhood that intersects at most finitely many of the sets $\mathcal{U}_\alpha$. Similarly, a collection of functions $\{f_\alpha : X \to \mathbb{R}\}_{\alpha \in I}$ is called **locally finite** if the sets $\{f_\alpha^{-1}(\mathbb{R}\backslash\{0\}) \subset X\}_{\alpha \in I}$ form a locally finite collection. This condition has the following advantage: if $\{f_\alpha : M \to \mathbb{R}\}_{\alpha \in I}$ is a locally finite collection of *smooth* functions on a manifold M, then one can make sense of the sum

$$\sum_{\alpha \in I} f_\alpha(p) \in \mathbb{R}$$

for every $p \in M$ since, even if I is an uncountably infinite set, at most finitely many terms in this sum are nonzero. Even better, p admits a neighborhood $\mathcal{V} \subset M$ that intersects at most finitely many of the sets $f_\alpha^{-1}(\mathbb{R}\backslash\{0\})$, implying that at most finitely many of the functions f_α can have nonzero values anywhere on $\mathcal{V}$, and $\sum_{\alpha \in I} f_\alpha$ therefore makes sense as a *smooth* function on $\mathcal{V}$. We therefore obtain a global smooth function

$$\sum_{\alpha \in I} f_\alpha \in C^\infty(M),$$

even if the sum contains uncountably many terms that are (somewhere) nontrivial functions on M.

Exercise 15.1. Show that if X is a topological space with open subset $\mathcal{U} \subset X$ and a locally finite collection of continuous functions $\{f_\alpha : X \to \mathbb{R}\}_{\alpha \in I}$ satisfying $\operatorname{supp}(f_\alpha) \subset \mathcal{U}$ for every $\alpha \in \mathcal{U}$, then $\sum_{\alpha \in I} f_\alpha$ also has support in $\mathcal{U}$.

Definition 15.2. Given an open cover $\{\mathcal{U}_\alpha\}_{\alpha \in I}$ of a smooth manifold M, a **partition of unity** subordinate to $\{\mathcal{U}_\alpha\}_{\alpha \in I}$ is a locally finite collection of smooth functions $\{\varphi_\alpha : M \to [0,1]\}_{\alpha \in I}$ which satisfy the following assumptions:

(1) For each $\alpha \in I$, $\operatorname{supp}(\varphi_\alpha) \subset \mathcal{U}_\alpha$;
(2) $\sum_{\alpha \in I} \varphi_\alpha \equiv 1$.

Note that in Definition 15.2, the condition $\sum_{\alpha \in I} \varphi_\alpha \equiv 1$ makes sense due to the local finiteness assumption; this condition was automatic in Lecture 11 because we were considering only a finite collection of functions, but here we are not assuming the collection is finite, nor that the functions have compact support. This relaxation of assumptions makes it possible to prove the following without assuming M is compact:

Theorem 15.3. *Every open cover of a smooth manifold admits a subordinate partition of unity.*

This theorem will be proved in §15.4.

15.2. Existence of Riemannian metrics and volume forms. Before proving that partitions of unity always exist, we shall demonstrate their usefulness by proving the following:

Theorem 15.4. *Every smooth manifold admits a Riemannian metric.*

As a preliminary remark relevant to the proof, we observe that on any vector space V, the set of inner products on V forms a *convex* subset of the vector space of bilinear maps $V \times V \to \mathbb{R}$. Indeed, the symmetric bilinear maps form a linear subspace, and whenever $\langle\ ,\ \rangle_0$ and $\langle\ ,\ \rangle_1$ are two inner products on V, the interpolation $\langle\ ,\ \rangle_t := t\langle\ ,\ \rangle_1 + (1-t)\langle\ ,\ \rangle_0$ for $t \in [0,1]$ also satisfies

$$\langle v, v\rangle_t = t\langle v, v\rangle_1 + (1-t)\langle v, v\rangle_0 > 0$$

for every nonzero $v \in V$. More generally, any *convex combination* of finitely many inner products on V is also an inner product, i.e. for any finite collection of inner products $\langle\ ,\ \rangle_i$ and numbers

$\tau_i \in [0,1]$ for $i = 1, \ldots, k$ with $\sum_{i=1}^k \tau_i = 1$,

$$\sum_{i=1}^k \tau_i \langle\ ,\ \rangle_i$$

is an inner product.

LEMMA 15.5. *Suppose $\{\mathcal{U}_\alpha\}_{\alpha\in I}$ is an open cover of a smooth manifold M with subordinate partition of unity $\{\varphi_\alpha\}_{\alpha\in I}$, and for each $\alpha \in I$, $g_\alpha \in \Gamma(T_2^0\mathcal{U}_\alpha)$ is a Riemannian metric on the open subset $\mathcal{U}_\alpha$. Then the formula*

$$g := \sum_{\alpha\in I} \varphi_\alpha g_\alpha$$

defines a Riemannian metric on M, where in this sum, the term $\varphi_\alpha g_\alpha$ is interpreted as an element of $\Gamma(T_2^0 M)$ that vanishes outside of $\mathcal{U}_\alpha$.

PROOF. Since $\operatorname{supp}(\varphi_\alpha) \subset \mathcal{U}_\alpha$, the tensor field $\varphi_\alpha g_\alpha \in \Gamma(T_2^0\mathcal{U}_\alpha)$ can be extended to a smooth tensor field on M that vanishes outside of $\mathcal{U}_\alpha$, and we will continue to denote the extension by $\varphi_\alpha g_\alpha \in \Gamma(T_2^0 M)$. The sum then makes sense and is smooth due to local finiteness, as every point is contained in a neighborhood on which only finitely many terms of the sum are nontrivial. Moreover, at each individual point $p \in M$, $g_p : T_pM \times T_pM \to \mathbb{R}$ is a convex combination of inner products, and is therefore also an inner product. □

PROOF OF THEOREM 15.4. Choose an open cover $\{\mathcal{U}_\alpha\}_{\alpha\in I}$ of M such that each $\mathcal{U}_\alpha$ is the domain of a chart x_α, and define a Riemannian metric g_α on $\mathcal{U}_\alpha$ that looks like the standard Euclidean inner product in the chosen coordinates. A global Riemannian metric $g \in \Gamma(T_2^0 M)$ can then be defined via Lemma 15.5 after choosing a subordinate partition of unity. □

In light of Corollary 11.10 on the Riemannian volume form associated to a Riemannian metric, Theorem 15.4 implies:

COROLLARY 15.6. *Every smooth oriented manifold admits a volume form.* □

EXERCISE 15.7. Use a partition of unity to prove Corollary 15.6 without mentioning Theorem 15.4 or Riemannian metrics. Use instead the fact that for any oriented n-dimensional vector space V, the set

$$\left\{\omega \in \Lambda^n V^* \;\middle|\; \omega(v_1, \ldots, v_n) > 0 \text{ for some positively-oriented basis } v_1, \ldots, v_n \in V\right\}$$

is convex.

REMARK 15.8. Without assuming M is oriented, Theorem 15.4 also implies that every smooth manifold admits a volume element (see §11.4).

15.3. Paracompactness. Any Riemannian manifold (M, g) is also a metric space in a natural way, at least if it is connected, because one can define the distance between two points $p, q \in M$ by

$$\operatorname{dist}(p, q) := \inf_\gamma \int_a^b \sqrt{g(\dot\gamma(t), \dot\gamma(t))}\, dt, \tag{15.1}$$

where the infimum is over all intervals $[a, b] \subset \mathbb{R}$ and smooth paths $\gamma : [a, b] \to M$ with $\gamma(a) = p$ and $\gamma(b) = q$. For a Riemannian manifold with multiple connected components, each component has a natural metric defined in this way, and there are standard tricks for defining metrics on any disjoint union of metric spaces (see e.g. Exercise 2.23). The point is: if you hadn't already assumed that smooth manifolds are metrizable but you assumed that Theorem 15.4 is true, then the theorem would imply metrizability.

EXERCISE 15.9. Take a moment to convince yourself that (15.1) really does define a metric, in particular that it satisfies the triangle inequality.
Hint: One can reparametrize the path $\gamma : [a, b] \to M$ quite freely without changing the integral. If you take advantage of this freedom, then a path from p to q and a path from q to r can always be concatenated smoothly.

The existence of the metric (15.1) is a dead giveaway that something about Theorem 15.4 depends on our assumption that all manifolds are metrizable. We haven't used that assumption in this course until now. But we will need it for constructing the partition of unity.

Recall that a **refinement** of an open cover $\{\mathcal{U}_\alpha\}_{\alpha \in I}$ is another open cover $\{\mathcal{V}_\beta\}_{\beta \in J}$ such that for every $\beta \in J$, $\mathcal{V}_\beta$ is contained in $\mathcal{U}_\alpha$ for some $\alpha \in I$.

DEFINITION 15.10. A topological space X is **paracompact** if every open cover of X admits a locally finite refinement.

Compact topological spaces are obviously paracompact since a finite subcover can also be viewed as a locally finite refinement. I can now tell you the true reason why we need to assume manifolds are metrizable: *all metrizable spaces are paracompact.* We will not prove quite such a general statement here, but we will make use of the metrizability assumption in the following to prove that manifolds are always paracompact.

LEMMA 15.11. *Every manifold M is σ-compact, i.e. it is the union of countably many compact subsets.*

PROOF. The result is true for every connected locally compact metric space (see e.g. [Spi99a, Theorem 1.2]), but for our purposes it will be more convenient to drop connectedness and instead assume separability, which holds in any case on all manifolds. Fix a metric d on M that is compatible with its topology. The term "locally compact" refers to the following observation: for every $p \in M$, the closed ball

$$\bar{B}_r(p) := \{q \in M \mid d(p, q) \leqslant r\}$$

is compact for every $r > 0$ sufficiently small. This holds because whenever r is sufficiently small, $\bar{B}_r(p)$ lies in the domain of a chart that identifies it with a closed and bounded (and therefore compact) subset of Euclidean space. On the other hand, closed and bounded subsets of arbitrary metric spaces are not always compact, so we cannot assume $\bar{B}_r(p)$ is compact for every $r > 0$, but there is a positive (if not infinite) upper bound

$$\kappa(p) := \sup \{r > 0 \mid \bar{B}_r(p) \text{ is compact}\} \in (0, \infty].$$

If $\kappa(p) = \infty$ at any point p, then M is exhausted by the sequence of compact sets $\bar{B}_k(p)$ for $k = 1, 2, 3, \dots$ and we are therefore done. Otherwise, observe that by the triangle inequality, every $q \in \bar{B}_{\frac{1}{3}\kappa(p)}(p)$ satisfies

$$\bar{B}_{\frac{1}{3}\kappa(p)}(q) \subset \bar{B}_{\frac{2}{3}\kappa(p)}(p),$$

implying that $\bar{B}_{\frac{1}{3}\kappa(p)}(q)$ is also compact and thus

$$\kappa(q) \geqslant \frac{\kappa(p)}{3} \quad \text{for all} \quad q \in \bar{B}_{\frac{1}{3}\kappa(p)}(p). \tag{15.2}$$

Now for any dense sequence $p_1, p_2, p_3, \dots \in M$, we claim that

$$M = \bigcup_{k=1}^{\infty} \bar{B}_{\frac{2}{3}\kappa(p_k)}(p_k),$$

where the sets on the right hand side are clearly all compact. Indeed, for any $p \in M$, we can replace $p_1, p_2, p_3, \dots$ with a subsequence such that $p_k \to p$ as $k \to \infty$, and it follows from (15.2) that $\kappa(p_k) \geqslant \kappa(p)/3$ for all k sufficiently large, so that eventually $p \in \bar{B}_{\frac{2}{3}\kappa(p_k)}$. □

Exercise 15.12. Show that if X is a topological space with a locally finite open cover $\{\mathcal{U}_\alpha\}_{\alpha\in I}$ and $K \subset X$ is a compact subset, then K intersects only finitely many of the sets $\mathcal{U}_\alpha$. (It follows from this that if X is σ-compact, then the set I cannot be uncountable, i.e. all locally finite open covers are at most countable. By Lemma 15.11, this applies in particular to all manifolds.

Theorem 15.13. *Every smooth manifold is paracompact.*

Proof. Assume $\{\mathcal{U}_\alpha\}_{\alpha\in I}$ is an open cover of M, and using Lemma 15.11, write $M = \bigcup_{j=1}^\infty K_j$ for compact subsets $K_1, K_2, K_3, \ldots$. Choose an open neighborhood $\mathcal{V}_1 \subset M$ of K_1 whose closure is compact, so the set $\bar{\mathcal{V}}_1 \cup K_2$ is also compact. Next, choose $\mathcal{V}_2 \subset M$ to be an open neighborhood of $\bar{\mathcal{V}}_1 \cup K_2$ whose closure is compact, so $\bar{\mathcal{V}}_2 \cup K_3$ is compact. Continuing in this way, one obtains a nested sequence

$$\varnothing =: \mathcal{V}_0 \subset \mathcal{V}_1 \subset \bar{\mathcal{V}}_1 \subset \mathcal{V}_2 \subset \bar{\mathcal{V}}_2 \subset \mathcal{V}_3 \subset \bar{\mathcal{V}}_3 \subset \ldots \subset \bigcup_{j=1}^\infty \mathcal{V}_j = M$$

such that each $\mathcal{V}_j$ is open and each $\bar{\mathcal{V}}_j$ is compact. We will now construct a locally finite refinement of $\{\mathcal{U}_\alpha\}_{\alpha\in I}$ by using the "annular" regions

$$A_j := \bar{\mathcal{V}}_j \backslash \mathcal{V}_{j-1} \subset M, \qquad j = 1, 2, 3, \ldots,$$

which are all compact, and their union is also M. For each $j \in \mathbb{N}$, pick a finite open covering $\{\mathcal{O}^j_\beta \subset M\}_{\beta\in I_j}$ of A_j such that each of the open sets $\mathcal{O}^j_\beta$ is contained in $\mathcal{U}_\alpha$ for some $\alpha \in I$ and is also contained in $\mathcal{V}_{j+1}\backslash\mathcal{V}_{j-2}$. The union of these finite collections for $j = 1, 2, 3, \ldots$ forms an open cover of M that refines $\{\mathcal{U}_\alpha\}_{\alpha\in I}$ and is also locally finite. $\square$

Exercise 15.14. Show that without loss of generality, one can assume that all of the open sets in the locally finite refinement given by Theorem 15.13 are diffeomorphic to open balls in Euclidean space.
Remark: This fact is frequently used in proofs that smooth manifolds admit partitions of unity, see for example [**Lee13a**, §II.3]. *It is not strictly necessary, however, and we will not use it. The proof given below is conceived to be as close as possible in spirit to proofs of similar results on more general topological spaces, which need not look locally like Euclidean space.*

15.4. Existence of partitions of unity. Now that we know that locally finite refinements can always be found, we need two further ingredients in order to construct partitions of unity. The first is purely topological.

A topological space X is called **normal** if every pair of disjoint closed subsets $A, B \subset X$ have neighborhoods in X that are also disjoint from each other.

Exercise 15.15. Show that all metric spaces are normal.

Lemma 15.16 (the "shrinking lemma"). *Given a locally finite open cover $\{\mathcal{U}_\alpha\}_{\alpha\in I}$ of a normal topological space X, there exists another open cover $\{\mathcal{V}_\alpha\}_{\alpha\in I}$ such that $\bar{\mathcal{V}}_\alpha \subset \mathcal{U}_\alpha$ for every $\alpha \in I$.*

Proof. We shall give a proof under the extra assumption that the set I is at most countable, which is always true on manifolds due to Exercise 15.12. A proof without this assumption is possible using Zorn's lemma, see e.g. [**nLa**].

Since I is at most countable, we can relable the open cover as $\{\mathcal{U}_i\}_{i=1}^N$ where $N \in \mathbb{N} \cup \{\infty\}$. The sets $A_1 := X\backslash\bigcup_{i=2}^\infty \mathcal{U}_2$ and $X\backslash\mathcal{U}_1$ are closed and disjoint, so we can choose $\mathcal{V}_1 \subset X$ to be any open neighborhood of A_1 that is also disjoint from some neighborhood of $X\backslash\mathcal{U}_1$, implying $\bar{\mathcal{V}}_1 \subset \mathcal{U}_1$. Since $X = \mathcal{V}_1 \cup \bigcup_{i=2}^N \mathcal{U}_i$, we can next take the latter as another open cover on X, and perform the same trick on $\mathcal{U}_2$, producing an open set $\mathcal{V}_2 \subset \bar{\mathcal{V}}_2 \subset \mathcal{U}_2$ such that $X = \mathcal{V}_1 \cup \mathcal{V}_2 \cup \bigcup_{i=3}^N \mathcal{U}_i$. Now repeat this procedure for $i = 3, 4, \ldots, N$, producing a sequence of shrunken open sets $\mathcal{V}_1, \mathcal{V}_2, \mathcal{V}_3, \ldots \subset X$ such that for each $m \in \mathbb{N}$,

$$X = \bigcup_{i=1}^{m} \mathcal{V}_i \cup \bigcup_{i=m+1}^{N} \mathcal{U}_i. \tag{15.3}$$

If $N < \infty$ then we are done. If $N = \infty$, we now appeal to local finiteness and observe that for every $p \in M$, there exists a largest $m \in \mathbb{N}$ for which $p \in \mathcal{U}_m$, hence (15.3) implies $p \in \bigcup_{i=1}^{m} \mathcal{V}_i$ and thus $X = \bigcup_{i=1}^{\infty} \mathcal{V}_i$. ☐

LEMMA 15.17 (the smooth Urysohn lemma). *Given a smooth manifold M with subsets $A \subset \mathcal{U} \subset M$ such that A is closed and $\mathcal{U}$ is open, there exists a smooth function $f : M \to [0, 1]$ with support in $\mathcal{U}$ such that $f|_A \equiv 1$.*

PROOF, PART 1. For this first of two steps in the proof, we add the assumption that $A \subset M$ is compact. Since the open sets $\mathcal{U}$ and $M \backslash A$ form a finite open cover of M, the compact case of our existence result for partitions of unity (Lemma 11.1) provides a pair of smooth functions $\varphi, \psi : M \to [0, 1]$ that have compact support in $\mathcal{U}$ and $M \backslash A$ respectively such that $\varphi + \psi \equiv 1$ on A. Since $\psi|_A \equiv 0$, the function we were looking for is φ. ☐

Before finishing the proof of Lemma 15.17, it will be convenient to forge ahead and show how these results imply the existence of partitions of unity.

PROOF OF THEOREM 15.3, WITH A CAVEAT. Starting from an arbitrary open cover $\{\mathcal{U}_\alpha\}_{\alpha \in I}$ of M, we can first replace $\{\mathcal{U}_\alpha\}_{\alpha \in I}$ by a locally finite refinement $\{\mathcal{O}_\beta\}_{\beta \in J}$. The latter has the property that for every $\beta \in J$, we can choose some $\alpha(\beta) \in I$ satisfying

$$\mathcal{O}_\beta \subset \mathcal{U}_{\alpha(\beta)}.$$

Next apply the shrinking lemma to find another open cover $\{\mathcal{V}_\beta\}_{\beta \in J}$ such that $\bar{\mathcal{V}}_\beta \subset \mathcal{O}_\beta$ for each $\beta \in J$. By Lemma 15.17, we can choose for each $\beta \in J$ a smooth function $f_\beta : M \to [0, 1]$ with support in $\mathcal{O}_\beta$ such that $f_\beta|_{\bar{\mathcal{V}}_\beta} \equiv 1$. Local finiteness implies that the sum $\sum_{\beta \in J} f_\beta$ is a well-defined smooth function on M, and since every point is contained in at least one of the sets $\mathcal{V}_\beta$, this sum is everywhere positive. Now for each $\alpha \in I$, define $\psi_\alpha : M \to \mathbb{R}$ by

$$\psi_\alpha := \sum_{\{\beta \in J \mid \alpha(\beta) = \alpha\}} f_\beta.$$

Local finiteness implies that these are also smooth functions on M and satisfy

$$\sum_{\alpha \in I} \psi_\alpha = \sum_{\beta \in J} f_\beta > 0,$$

and moreover, since each f_β in the sum for $\alpha(\beta) = \alpha$ has support in $\mathcal{O}_\beta \subset \mathcal{U}_\alpha$, ψ_α itself has support in $\mathcal{U}_\alpha$ (see Exercise 15.1). The desired functions φ_α can now be defined by

$$\varphi_\alpha := \frac{\psi_\alpha}{\sum_{\beta \in I} \psi_\beta}.$$

☐

Since we did not yet finish the proof of Lemma 15.17, let's pause now to consider what actually has been proved. Lemma 15.17 was used in the above proof to choose the functions f_β with support in $\mathcal{O}_\beta$ that equal 1 on $\bar{\mathcal{V}}_\beta \subset \mathcal{O}_\beta$. If we add to the hypotheses of Theorem 15.3 that each of the open sets $\mathcal{U}_\alpha \subset M$ has compact closure, then it guarantees that the sets $\bar{\mathcal{V}}_\beta$ are also compact, so that we only need to use the case of Lemma 15.17 that has already been proved. In summary, Theorem 15.3 has now been established under the extra hypothesis that each set $\bar{\mathcal{U}}_\alpha \subset M$ is compact. We can use this observation to complete the proof of Lemma 15.17 and thus establish Theorem 15.3 in full generality.

PROOF OF LEMMA 15.17, PART 2. Choose open coverings $\{\mathcal{U}_\alpha \subset M\}_{\alpha\in I}$ of A and $\{\mathcal{O}_\beta \subset M\}_{\beta\in J}$ of $M\backslash A$ such that all of the sets $\mathcal{U}_\alpha, \mathcal{O}_\beta$ have compact closure and

$$\mathcal{U}_\alpha \subset \mathcal{U} \text{ for all } \alpha \in I, \qquad \mathcal{O}_\beta \subset M\backslash A \text{ for all } \beta \in J.$$

Then $M = \bigcup_{\alpha\in I} \mathcal{U}_\alpha \cup \bigcup_{\beta\in J} \mathcal{O}_\beta$, and we can apply the case of Theorem 15.3 that has been proved already to find a locally finite partition of unity subordinate to this cover: it consists of smooth functions $\{\varphi_\alpha\}_{\alpha\in I}$ and $\{\psi_\beta\}_{\beta\in J}$ such that $\operatorname{supp}(\varphi_\alpha) \subset \mathcal{U}_\alpha$ and $\operatorname{supp}(\psi_\beta) \subset \mathcal{O}_\beta$ for all $(\alpha, \beta) \in I \times J$, while $\sum_{\alpha\in I} \varphi_\alpha + \sum_{\beta\in J} \psi_\beta \equiv 1$. Since every $\mathcal{O}_\beta$ is disjoint from A, it follows that $f := \sum_{\alpha\in I} \varphi_\alpha \equiv 1$ on A, and by Exercise 15.1, $\operatorname{supp}(f) \subset \mathcal{U}$. □

The proof of Theorem 15.3 is now complete.

REMARK 15.18. We made use of separability at one step in this lecture—namely in Lemma 15.11 on σ-compactness—because doing so was more convenient than the alternative, but it was not strictly necessary. As mentioned in the proof of Lemma 15.11, the lemma also holds for arbitrary connected and locally compact metric spaces, so if one works on only one connected component at a time, one obtains a proof of paracompactness for "manifolds" that are assumed metrizable but not necessarily separable. Some authors prefer in fact to define a manifold in a slightly more general way than we have, requiring them to be Hausdorff and paracompact but not necessarily separable or second countable—this shows you how highly the existence of partitions of unity is valued by differential geometers. The only difference this makes in reality is that by the more general definition, manifolds can have uncountably many connected components; in the connected case there is no difference. In any case, I have never seen an example of a non-separable "manifold" that I cared about, not even in infinite dimensions.

REMARK 15.19. On a topological space X, there is generally no well-defined notion of smooth functions, but one can still speak of partitions of unity in which the functions $\varphi_\alpha : X \to [0, 1]$ are only required to be continuous. Such constructions are similarly useful in topology for proving existence results, e.g. the fact that every finite-dimensional *topological* manifold admits a proper topological embedding into $\mathbb{R}^N$ for N sufficiently large (see [**Lee11**, Chapter 4]). To prove that partitions of unity exist on a given space X, one obviously needs to know that X is paracompact, and the other major ingredients are the shrinking lemma (Lemma 15.16) and the continuous variant of Urysohn's lemma (Lemma 15.17), both of which hold whenever X is normal. It turns out that paracompact Hausdorff spaces are automatically normal, thus they admit continuous partitions of unity—in fact for Hausdorff spaces in general, the existence of partitions of unity is *equivalent* to paracompactness.

In nonlinear functional analysis, one sometimes also works with infinite-dimensional smooth manifolds that are locally modelled on Banach spaces. These are not locally compact, so our proof of paracompactness via σ-compactness does not adapt well to the infinite-dimensional setting, but one can nonetheless appeal to the fact that metric spaces are *always* paracompact. The simplest (or at least the shortest) proof of this is due to Mary Ellen Rudin [**Rud69**]. If one considers *arbitrary* metric spaces, then the proof makes slightly mysterious use of the axiom of choice, in the form of the well-ordering theorem: in particular, it uses the fact that for any open cover $\{\mathcal{U}_\alpha\}_{\alpha\in I}$, the index set I can be endowed with a total order for which every subset has a smallest element. This is less mysterious however in the case of *separable* metric spaces, because every open cover in the separable case admits a finite subcover (exercise!), so one is free without loss of generality to assume the index set is $\mathbb{N}$. As a consequence, infinite-dimensional Banach manifolds are also paracompact, so long as we still agree that anything carrying the name "manifold" should be metrizable and separable. That is the convention that I adopt when I use these objects in my research, and it is not the only possible convention that people might consider reasonable, but it is relatively uncontroversial.

The existence of smooth partitions of unity in the infinite-dimensional setting is nonetheless a subtle question, because smooth compactly-supported "bump" functions do not always exist on Banach spaces—the basic problem here is that on a Banach space E with norm $\|\cdot\|$, the function $E \to \mathbb{R} : x \mapsto \|x\|^p$ is not generally differentiable at $0 \in E$ for any power $p > 0$, even for $p = 2$. As a result, the smooth Urysohn lemma is not true in this context, so smooth partitions of unity do not exist, and many popular constructions from differential geometry are simply not available on infinite-dimensional Banach manifolds. The exception is the case of Hilbert manifolds, which are locally modelled on Hilbert spaces—the inner product on a Hilbert space $\mathcal{H}$ has the convenient property that $\mathcal{H} \to \mathbb{R} : x \mapsto \|x\|^2 := \langle x, x \rangle$ is a smooth function, thus making smooth bump functions and smooth partitions of unity possible.

Exercise 15.20. Given a smooth manifold M, use an open cover and subordinate partition of unity on M to construct a Riemannian metric on the tangent bundle TM. Do *not* assume that Theorem 15.3 holds for TM.
Remark: This exercise ties up a loose end from early in the course: in Corollary 3.12, we defined a smooth structure on the tangent bundle TM of any smooth manifold M, but we never proved that the topology on TM induced by its maximal smooth atlas is metrizable. The existence of a Riemannian metric implies this, and if you follow the instructions in the exercise, its construction does not need to assume that TM is metrizable—it assumes only that M is.

Exercise 15.21. Here is an example of something that satisfies all of the conditions for being a connected smooth 2-manifold except metrizability. It is a variation due to Calabi and Rosenlicht [CR53] on a construction known as the **Prüfer surface**, and can be visualized as a an uncountable collection of planes that have been glued together along their open upper and lower halves, but not along the x-axis, so that the result is a single plane in which the x-axis has been replaced by uncountably many copies of itself. Here is a precise definition: denote the open upper and lower half-planes by $\mathbb{H}_\pm := \{(x, y) \in \mathbb{R}^2 \mid \pm y > 0\}$, and associate to each $a \in \mathbb{R}$ a copy of the full plane $X_a := \mathbb{R}^2$. As a set, the Prüfer surface is

$$\Sigma := \mathbb{H}_+ \cup \mathbb{H}_- \sqcup \left(\coprod_{a \in \mathbb{R}} X_a \right) \Big/ \sim$$

where the equivalence relation identifies each point $(x, y) \in X_a$ for $y \neq 0$ with the point $(a + yx, y) \in \mathbb{H}_+ \cup \mathbb{H}_-$. Notice that $\mathbb{H}_\pm$ and X_a for each $a \in \mathbb{R}$ can be regarded naturally as subsets of Σ. Let us denote points $(x, y) \in X_a \subset \Sigma$ by

$$(x, y)_a \in \Sigma,$$

so by definition, $(x, y)_a = (x', y')_b$ whenever $y = y' \neq 0$ and $xy + a = x'y' + b$, but $(x, 0)_a$ and $(x', 0)_b$ are never equal when $a \neq b$. Prove:

(a) Σ admits a unique smooth structure for which the natural inclusions $\mathbb{H}_\pm \hookrightarrow \Sigma$ and $X_a \hookrightarrow \Sigma$ for each $a \in \mathbb{R}$ are diffeomorphisms onto their images. *Assume for the remaining parts of this exercise that Σ is equipped with the topology uniquely determined by this smooth structure (cf. Prop. 2.12).*
(b) For any two points $p, q \in \Sigma$, there exist neighborhoods $p \in \mathcal{U} \subset \Sigma$ and $q \in \mathcal{V} \subset \Sigma$ such that $\mathcal{U} \cap \mathcal{V} = \emptyset$. (In topological terminology, Σ is Hausdorff.)
Hint: The only case where it is not so obvious is when p and q are both of the form $(x, 0)_a$ and $(x', 0)_b$. Try drawing pictures of the intersections of neighborhoods of those points with $\mathbb{H}_+ \cup \mathbb{H}_-$.

(c) Σ is connected.
(d) Σ is separable.
Hint: Show that any dense subset of $\mathbb{H}_+ \cup \mathbb{H}_- \subset \Sigma$ *is also dense in* Σ.
(e) Here's where things get weird: the subset $\{(0,0)_a \in \Sigma \mid a \in \mathbb{R}\} \subset \Sigma$ is discrete, i.e. each of its points has a neighborhood that contains none of the others. In particular, all subsets of this set are closed.
(f) Σ is not σ-compact (no pun intended).
Hint: According to part (e), it contains an uncountable discrete subset.

We can now deduce that Σ is not metrizable, as we would otherwise have a contradiction to the proof of Lemma 15.11. Here is an even stranger indication: recall from Exercise 15.15 that all metric spaces are normal.

(g) Suppose we have associated to each $a \in \mathbb{R}$ a "wedge-shaped" region in $\mathbb{H}_+$ of the form

$$W_a := \{(r\cos\theta, r\sin\theta) \in \mathbb{H}_+ \mid r \in (0, r(a)) \text{ and } \theta \in (\pi/2 - \epsilon(a), \pi/2 + \epsilon(a))\}$$

for constants $r(a) > 0$ and $\epsilon(a) > 0$ that are allowed to vary arbitrarily with $a \in \mathbb{R}$. Show that there exists some $a_\infty \in \mathbb{Q}$ and a sequence $a_j \in \mathbb{R}\backslash\mathbb{Q}$ that converges to a_∞ such that $r(a_j)$ and $\epsilon(a_j)$ are both bounded from below.
Big hint: $\mathbb{R} = \mathbb{Q} \cup \bigcup_{N\in\mathbb{N}} A_N$ *where*

$$A_N := \{a \in \mathbb{R}\backslash\mathbb{Q} \mid r(a) \geqslant 1/N \text{ and } \epsilon(a) \geqslant 1/N\}.$$

According to the Baire category theorem, a nonempty complete metric space can never be the countable union of subsets that are nowhere dense, meaning sets whose closures have empty interior. Deduce from this that at least one of the sets A_N *contains an open interval in its closure.*
(h) Deduce that the disjoint subsets

$$Q := \{(0,0)_a \in \Sigma \mid a \in \mathbb{Q}\} \subset \Sigma \qquad \text{and} \qquad I := \{(0,0)_a \in \Sigma \mid a \in \mathbb{R}\backslash\mathbb{Q}\} \subset \Sigma$$

are both closed but do not admit disjoint neighborhoods, i.e. Σ is not normal.
(i) Show that the open cover $\{X_a \subset \Sigma\}_{a\in\mathbb{R}}$ of Σ has no locally finite refinement.
Hint: In any refinement of $\{X_a\}_{a\in\mathbb{R}}$*, points of the form* $(0,0)_a$ *and* $(0,0)_b$ *for* $a \neq b$ *must always belong to different sets in the open cover. Show that for the point* $a_\infty \in \mathbb{R}$ *in part (g), every neighborhood of* $(0,0)_{a_\infty}$ *intersects infinitely many such sets.*

The original Prüfer surface is slightly different from the variation by Calabi and Rosenlicht described above, and can be defined as

$$\Sigma' := \mathbb{H}_+ \sqcup \left(\coprod_{a\in\mathbb{R}} X_a\right) \Big/ \sim,$$

where the equivalence relation identifies points $(x, y) \in X_a$ with $(a + yx, y) \in \mathbb{H}_+$ only for $y > 0$. We can visualize Σ' as an uncountable collection of planes that have been glued together along their upper halves, leaving the lower halves separate.

(j) Show that Σ' has all the same properties we proved above for Σ, except that Σ' is not separable.

16. Vector bundles

We have already seen several examples in this course of sets of the form

$$E = \bigcup_{p\in M} E_p,$$

where M is a manifold and E_p is a vector space associated to each point $p \in M$. The obvious example is the tangent bundle TM, but we have also considered the cotangent bundle T^*M, which is the union of the dual spaces to the tangent spaces, and further examples arise in natural ways by thinking of tensor fields $S \in \Gamma(T^k_\ell M)$ as objects that associate to each point $p \in M$ an element S_p of a certain vector space of multilinear maps. For all of these examples, one can regard the vector spaces E_p as "varying smoothly" with respect to p, but this is an intuitive notion that we have not yet made precise except in the special case of TM, on which we defined a smooth structure so that the natural projection $\pi : TM \to M$ sending T_pM to p is a smooth map.

We will now start defining such notions in greater generality.

16.1. Main Definition. We begin with a few more observations about the motivating example of a vector bundle, namely the tangent bundle TM of a smooth n-manifold M. Recall that each chart $(\mathcal{U}, x)$ on M determines a family of vector space isomorphisms

$$d_px : T_pM \to \mathbb{R}^n, \qquad p \in \mathcal{U}.$$

This information can be repackaged as a bijective map

$$\Phi_x : T\mathcal{U} \to \mathcal{U} \times \mathbb{R}^n$$

whose restriction to each of the individual vector spaces $T_pM \subset T\mathcal{U}$ for $p \in \mathcal{U}$ is $X \mapsto (p, d_px(X)) \in \mathcal{U} \times \mathbb{R}^n$, and the smooth chart $(T\mathcal{U}, Tx)$ for TM can be derived from this by writing

$$Tx(X) = (x(p), d_px(X)) = (x \times \mathbb{1}) \circ \Phi_x(X) \in \mathbb{R}^n \times \mathbb{R}^n \qquad \text{for } X \in T_pM,\ p \in \mathcal{U}.$$

Since $x \times \mathbb{1} : \mathcal{U} \times \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^n$ is clearly a smooth map, the way that we defined the smooth structure on TM makes Φ_x not just a bijection, but also a diffeomorphism. Now, if $(\mathcal{V}, y)$ is another chart with $\mathcal{U} \cap \mathcal{V} \neq \emptyset$, there is a similar diffeomorphism

$$\Phi_y : T\mathcal{V} \to \mathcal{V} \times \mathbb{R}^n,$$

and both Φ_x and Φ_y restrict to diffeomorphisms $T(\mathcal{U} \cap \mathcal{V}) \to (\mathcal{U} \cap \mathcal{V}) \times \mathbb{R}^n$, giving rise to a map

$$\begin{aligned} \Phi_y \circ \Phi_x^{-1} : (\mathcal{U} \cap \mathcal{V}) \times \mathbb{R}^n &\to (\mathcal{U} \cap \mathcal{V}) \times \mathbb{R}^n \\ (p, v) &\mapsto (p, g(p)v), \end{aligned}$$

where

$$g(p) := d_py \circ (d_px)^{-1} = D(y \circ x^{-1})(x(p)) \in \mathrm{GL}(n, \mathbb{R}) \subset \mathbb{R}^{n \times n}.$$

The smooth compatibility of x and y implies that $g : \mathcal{U} \cap \mathcal{V} \to \mathrm{GL}(n, \mathbb{R})$ is also a smooth function. The existence of maps such as Φ_x and Φ_y is one way of making precise the notion that the tangent spaces T_pM vary smoothly with $p \in M$. We take this as motivation for the definition below.

Notation. In everything that follows, we choose a field

$$\mathbb{F} = \text{either } \mathbb{R} \text{ or } \mathbb{C},$$

and assume unless otherwise noted that all vector spaces and linear maps are $\mathbb{F}$-linear. In this way the real and complex cases can be handled simultaneously.

Definition 16.1. Assume M is a smooth n-manifold, E_p is an m-dimensional vector space over $\mathbb{F}$ associated to each point $p \in M$, and define the set

$$E := \bigcup_{p \in M} E_p,$$

where E_p and E_q are regarded as disjoint sets for $p \neq q$.[51] For any subset $\mathcal{U} \subset M$, denote

$$E|_{\mathcal{U}} := \bigcup_{p \in \mathcal{U}} E_p \subset E.$$

A **local trivialization** (*lokale Trivialisierung*) of E is a pair $(\mathcal{U}_\alpha, \Phi_\alpha)$ consisting of an open subset $\mathcal{U}_\alpha \subset M$ and a bijection

$$E|_{\mathcal{U}_\alpha} \xrightarrow{\Phi_\alpha} \mathcal{U}_\alpha \times \mathbb{F}^m$$

such that for each $p \in \mathcal{U}_\alpha$, Φ_α restricts to E_p as a map of the form $v \mapsto (p, \phi_p v)$ for some vector space isomorphism $\phi_p : E_p \to \mathbb{F}^m$.

Any two local trivializations $(\mathcal{U}_\alpha, \Phi_\alpha)$ and $(\mathcal{U}_\beta, \Phi_\beta)$ determine **transition functions** (*Übergangsfunktionen*) $g_{\beta\alpha}, g_{\alpha\beta} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{GL}(m, \mathbb{F})$ such that the map $\Phi_\beta \circ \Phi_\alpha^{-1} : (\mathcal{U}_\alpha \cap \mathcal{U}_\beta) \times \mathbb{F}^m \to (\mathcal{U}_\alpha \cap \mathcal{U}_\beta) \times \mathbb{F}^m$ and its inverse take the form

$$\begin{aligned} \Phi_\beta \circ \Phi_\alpha^{-1}(p, v) &= (p, g_{\beta\alpha}(p)v), \\ \Phi_\alpha \circ \Phi_\beta^{-1}(p, v) &= (p, g_{\alpha\beta}(p)v). \end{aligned} \tag{16.1}$$

We say that $(\mathcal{U}_\alpha, \Phi_\alpha)$ and $(\mathcal{U}_\beta, \Phi_\beta)$ are C^k-compatible for $k \in \mathbb{N} \cup \{0, \infty\}$ (or **smoothly compatible** in the case $k = \infty$) if the transition functions $g_{\beta\alpha}$ and $g_{\alpha\beta}$ are of class C^k.

Exercise 16.2. Show that the two transition functions $g_{\alpha\beta}, g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{GL}(m, \mathbb{F})$ in Definition 16.1 are related to each other by $g_{\beta\alpha}(p) = [g_{\alpha\beta}(p)]^{-1} \in \mathrm{GL}(m, \mathbb{F})$ for all $p \in \mathcal{U}_\alpha \cap \mathcal{U}_\beta$, and conclude that $g_{\alpha\beta}$ is of class C^k if and only if $g_{\beta\alpha}$ is.

Remark 16.3. The notion of C^k-compatibility for transition functions is based on the premise that we know what it means to say that a real or complex matrix-valued function on a smooth manifold is of class C^k. This is fine because $\mathbb{R}^{n \times n}$ and $\mathbb{C}^{n \times n}$ can both be regarded as finite-dimensional real vector spaces (every complex vector space is also a real vector space), and the notion of smoothness for functions $f : M \to V$ is well defined whenever M is a smooth manifold and V is a real vector space. The notion of smoothness would be much less clear if we replaced $\mathbb{F}$ with a different field such as $\mathbb{Z}_2$ or $\mathbb{Q}$; there is no theory of differential calculus for functions on open subsets of $\mathbb{R}^n$ with values only in $\mathbb{Z}_2$ or $\mathbb{Q}$. That is one of a few reasons why we will never consider such generalizations in this course.

Definition 16.4. Assume M is a manifold. A **vector bundle of class C^k with rank m over M** (*ein Vektorbündel von der Klasse C^k mit Rang m über M*) is a collection of m-dimensional vector spaces $E = \bigcup_{p \in M} E_p$ as in Definition 16.1, equipped with a maximal collection of C^k-compatible local trivializations $\{(\mathcal{U}_\alpha, \Phi_\alpha)\}_{\alpha \in I}$ such that $M = \bigcup_{\alpha \in I} \mathcal{U}_\alpha$. The vector spaces E_p for $p \in M$ are called the **fibers** (*Fasern*) of the vector bundle E, M is called the **base** (*Basis*) of E, and the set E itself is called the **total space** (*Totalraum*). The surjective map

$$\pi : E \to M$$

sending each fiber $E_p \subset E$ to the point $p \in M$ is sometimes called the **bundle projection**. We will denote the rank of E by

$$\mathrm{rank}_{\mathbb{F}}(E) := m \geqslant 0,$$

or simply $\mathrm{rank}(E)$ whenever the field $\mathbb{F}$ is clear from context.

[51] In set-theoretic terms, this means we are defining E as the disjoint union of all the sets E_p, so we could also have written $E = \coprod_{p \in M} E_p$. We prefer however to avoid the use of the symbol "$\coprod$" here, because we will soon define a topology on E, and it will not be the disjoint union topology.

EXERCISE 16.5. By identifying $\mathbb{C}^m$ with $\mathbb{R}^{2m}$, show that every complex vector bundle E of class C^k can also be regarded as a real vector vector bundle of class C^k with

$$\operatorname{rank}_{\mathbb{R}}(E) = 2\operatorname{rank}_{\mathbb{C}}(E).$$

REMARK 16.6. A vector bundle of rank m is also sometimes called an **m-plane bundle** or an "m-dimensional" vector bundle, and in the case $m = 1$, a **line bundle** (*Geradenbündel*). The latter terminology is quite intuitive when $\mathbb{F} = \mathbb{R}$, but one must keep in mind that in the complex case, the fibers should be visualized as *planes* rather than lines.

NOTATION. We will often refer to the vector bundle in Definition 16.4 simply as E, but doing so ignores quite a lot of important information, such as the base manifold M, fibers E_p, their vector space structures and the local trivializations. It is common in the literature to abbreviate all this data in terms of the projection map and thus refer to $\pi : E \to M$ or (E, π) as a vector bundle, sometimes also omitting the symbol π and writing

$$E \to M.$$

This is an imperfect convention, but we will sometimes also follow it: the projection map has the advantage that it determines the fibers

$$E_p = \pi^{-1}(p),$$

even though it does not determine their vector space structures or the local trivializations.

Observe that if M is a manifold of class C^ℓ for some finite ℓ, then vector bundles of class C^k make sense for every $k \leqslant \ell$, but cannot be defined for $k > \ell$. As usual, we will mostly only consider the case $k = \ell = \infty$, and then refer to E as a **smooth vector bundle**. We also call E a *real* vector bundle if $\mathbb{F} = \mathbb{R}$, and a *complex* vector bundle if $\mathbb{F} = \mathbb{C}$.

REMARK 16.7. The maximal collection of local trivializations $\{(\mathcal{U}_\alpha, \Phi_\alpha)\}_{\alpha\in I}$ in Definition 16.4 plays a similar role to the maximal atlas on a smooth manifold; maximality serves as a bookkeeping device to make sure in this setting that whenever $\{(\mathcal{U}_\alpha, \Phi_\alpha)\}_{\alpha\in I}$ and $\{(\mathcal{V}_\beta, \Psi_\beta)\}_{\beta\in J}$ are two coverings of E by smoothly compatible local trivializations such that every $(\mathcal{U}_\alpha, \Phi_\alpha)$ is smoothly compatible with every $(\mathcal{V}_\beta, \Psi_\beta)$, both can be understood as defining *the same* smooth vector bundle. As with manifolds, one never actually needs to specify a maximal collection of local trivializations, as a maximal collection is uniquely determined by *any* collection $\{(\mathcal{U}_\alpha, \Phi_\alpha)\}$ for which the sets $\mathcal{U}_\alpha$ cover M. When E is a smooth vector bundle, a local trivialization will be called **smooth** whenever it belongs to the associated maximal collection.

REMARK 16.8. Vector bundles of class C^0, also known as *topological* vector bundles, can be defined without assuming the base M is a manifold—the definition makes sense with an arbitrary topological space in place of M, and one can show that E then admits a natural topology such that the bundle projection $\pi : E \to M$ is continuous and the local trivializations are homeomorphisms. (The definition that appears in topology books usually assumes that E is *given* with a topology such that $\pi : E \to M$ is continuous and the fibers $E_p = \pi^{-1}(p)$ are vector spaces; one then calls $\pi : E \to M$ a vector bundle if and only if every $p \in M$ admits a neighborhood $\mathcal{U}$ for which there exists a homeomorphism $\Phi : \pi^{-1}(\mathcal{U}) \to \mathcal{U} \times \mathbb{F}^m$ that is a local trivialization.) For many applications, it is also advisable to assume that M is a paracompact Hausdorff space, so that partitions of unity can be used for various constructions, e.g. one can endow the fibers E_p with inner products that depend continuously on p, analogous to a Riemannian metric.

REMARK 16.9. The notion of C^k-compatibility between two local trivializations $(\mathcal{U}_\alpha, \Phi_\alpha)$ and $(\mathcal{U}_\beta, \Phi_\beta)$ could have been defined without mentioning the transition functions $g_{\beta\alpha}, g_{\alpha\beta} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{GL}(m, \mathbb{F})$, as it would be equivalent to require that the maps $\Phi_\beta \circ \Phi_\alpha^{-1}$ and $\Phi_\alpha \circ \Phi_\beta^{-1}$ are of class C^k

on $(\mathcal{U}_\alpha \cap \mathcal{U}_\beta) \times \mathbb{F}^m$. In more general contexts, in particular when we talk next semester about *fiber bundles* whose fibers are smooth manifolds rather than vector spaces, it will be necessary to reformulate the notion of smooth compatibility without the transition functions $g_{\alpha\beta}$ and $g_{\beta\alpha}$, as these naturally take values in the diffeomorphism group $\mathrm{Diff}(F)$ of some manifold F, and defining "smoothness" for maps with values in $\mathrm{Diff}(F)$ is something of a technical minefield. We do not have this problem with vector bundles, due to the fact that $\mathrm{GL}(m, \mathbb{F})$ is naturally a smooth finite-dimensional manifold, and (16.1) shows moreover that the transition functions encode all of the essential information in this setting. It will be especially useful to focus on them when we start talking about vector bundles with extra geometric structure such as bundle metrics or volume forms. In reality, this is also true for most fiber bundles that are of interest, because instead of considering $g_{\alpha\beta}$ and $g_{\beta\alpha}$ with values in $\mathrm{Diff}(F)$, one can often take them to have values in some finite-dimensional smooth Lie group G that acts smoothly on the manifold F. We will see many examples of this next semester.

Here is a generalization of the fact that tangent bundles are smooth manifolds.

Proposition 16.10. *For any smooth vector bundle $\pi : E \to M$ over a smooth manifold M, the total space E naturally has the structure of a smooth manifold of dimension*

$$\dim E = \begin{cases} \dim M + \operatorname{rank}(E) & \text{if } \mathbb{F} = \mathbb{R}, \\ \dim M + 2\operatorname{rank}(E) & \text{if } \mathbb{F} = \mathbb{C}, \end{cases}$$

such that the projection map π and the inclusions $E_p \hookrightarrow E$ for $p \in M$ and

$$i : M \hookrightarrow E : p \mapsto 0 \in E_p$$

are all smooth maps.

Proof. The proof is analogous to that of Corollary 3.12, which was the case $E = TM$. The key point is that M can be covered by open sets $\mathcal{U}_\alpha \subset M$ which are domains of charts $x_\alpha : \mathcal{U}_\alpha \to \mathbb{R}^n$ and also appear in local trivializations $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^m$. The map

$$\phi_\alpha := (x_\alpha \times \mathbb{1}) \circ \Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathbb{R}^n \times \mathbb{F}^m \tag{16.2}$$

is then an $(n+m)$-dimensional chart for E on the domain $E|_{\mathcal{U}_\alpha} \subset E$ if $\mathbb{F} = \mathbb{R}$, or if $\mathbb{F} = \mathbb{C}$, an $(n+2m)$-dimensional chart after identifying $\mathbb{C}^m$ with $\mathbb{R}^{2m}$. The smooth compatibility of the charts $(\mathcal{U}_\alpha, x_\alpha)$ and local trivializations $(\mathcal{U}_\alpha, \Phi_\alpha)$ implies (exercise!) that all charts of this form on E are likewise smoothly compatible. The topology defined on E via these charts is metrizable and separable for the same reasons as in the case $E = TM$; in particular, one can use a partition of unity on M to construct a Riemannian metric on the total space E as in Exercise 15.20, proving that E is metrizable. □

Definition 16.11. A **section** (*Schnitt*) of a vector bundle $\pi : E \to M$ is a map $s : M \to E$ such that $\pi \circ s = \mathrm{Id}_M$. In other words, s assigns to each point $p \in M$ a vector in the corresponding fiber $s(p) \in E_p$. We say s is a section **of class** C^k if it is a C^k-map $M \to E$ with respect to the smooth structure on E defined in Proposition 16.10. The vector space of smooth sections is denoted by

$$\Gamma(E) := \left\{ s \in C^\infty(M, E) \;\middle|\; \pi \circ s = \mathrm{Id}_M \right\},$$

with addition and scalar multiplication in $\Gamma(E)$ defined pointwise, e.g. $s + t \in \Gamma(E)$ is defined for $s, t \in \Gamma(E)$ by $(s+t)(p) = s(p) + t(p) \in E_p$.

You might find it unsurprising but not completely obvious that $s + t$ is always a *smooth* section whenever s and t are. To make this obvious, we need to reformulate slightly the meaning of smoothness for a section $s : M \to E$. We observe first that for any local trivialization $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^m$, every section $s : M \to E$ uniquely determines a vector-valued function

$$s_\alpha : \mathcal{U}_\alpha \to \mathbb{F}^m$$

such that

$$\Phi_\alpha(s(p)) = (p, s_\alpha(p)) \qquad \text{for all } p \in \mathcal{U}_\alpha.$$

We will call this the **local representation** of s with respect to the trivialization $(\mathcal{U}_\alpha, \Phi_\alpha)$. After shrinking $\mathcal{U}_\alpha$ if necessary to a smaller neighborhood of any given point in $\mathcal{U}_\alpha$, we are free to assume that it is also the domain of a chart $x_\alpha : \mathcal{U}_\alpha \to \mathbb{R}^n$, in which case (16.2) defines a corresponding chart $\phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathbb{R}^n \times \mathbb{F}^m$ for E with the convenient property that its domain contains $s(p)$ for every $p \in \mathcal{U}_\alpha$. Using the charts x_α on M and ϕ_α on E, we obtain a local coordiate representation for the map $s : M \to E$, in the form

$$\phi_\alpha \circ s \circ x_\alpha^{-1} : x(\mathcal{U}_\alpha) \to x(\mathcal{U}_\alpha) \times \mathbb{F}^m : q \mapsto (q, s_\alpha \circ x_\alpha^{-1}(q)).$$

By definition, $s : M \to E$ is a smooth map if and only if this local coordinate representation is smooth for every choice of smooth chart $(\mathcal{U}_\alpha, x_\alpha)$ on M and smooth local trivialization $(\mathcal{U}_\alpha, \Phi_\alpha)$ of E. The latter is clearly true if and only if s_α is a smooth function, so we've proved:

PROPOSITION 16.12. *A section $s : M \to E$ is smooth if and only if its local coordinate representations $s_\alpha : \mathcal{U}_\alpha \to \mathbb{F}^m$ with respect to arbitrary smooth local trivializations $(\mathcal{U}_\alpha, \Phi_\alpha)$ are all smooth.* □

Since $C^\infty(\mathcal{U}_\alpha, \mathbb{F}^m)$ is a vector space for every open set $\mathcal{U}_\alpha$, Proposition 16.12 implies that $\Gamma(E)$ is also a vector space.

EXERCISE 16.13. Show that if $(\mathcal{U}_\alpha, \Phi_\alpha)$ and $(\mathcal{U}_\beta, \Phi_\beta)$ are two local trivializations of E and $s : M \to E$ is a section, then the local representations $s_\alpha : \mathcal{U}_\alpha \to \mathbb{F}^m$ and $s_\beta : \mathcal{U}_\beta \to \mathbb{F}^m$ are related to each other on $\mathcal{U}_\alpha \cap \mathcal{U}_\beta$ in terms of the transition function $g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{GL}(m, \mathbb{F})$ by

$$s_\beta(p) = g_{\beta\alpha}(p) s_\alpha(p) \qquad \text{for } p \in \mathcal{U}_\alpha \cap \mathcal{U}_\beta.$$

Remark: Since the transition functions on a smooth vector bundle are all smooth, this exercise implies that the condition in Proposition 16.12 does not need to be checked for all possible smooth local trivializations—it suffices to check it for a family of trivializations that cover M.

DEFINITION 16.14. Assume $E \to M$ and $F \to M$ are two smooth vector bundles over the same manifold M. A smooth map $\Psi : E \to F$ is called a **smooth linear bundle map** if for every $p \in M$, the restriction $\Psi|_{E_p}$ is a linear map

$$\Psi_p : E_p \to F_p.$$

We call Ψ **fiberwise injective** / **surjective** if Ψ_p is injective / surjective for every $p \in M$, and Ψ is a **bundle isomorphism** if Ψ_p is a vector space isomorphism for every $p \in M$. The bundles E and F are called **isomorphic** if and only if there exists a bundle isomorphism $E \to F$.

REMARK 16.15. Definition 16.14 presumes that E and F are both bundles over the same field $\mathbb{F}$. If one is a real vector bundle and the other is complex, then one can always regard the complex bundle as a real bundle with twice the rank (see Exercise 16.5) and thus interpret $\Psi : E \to F$ as a smooth *real*-linear bundle map.

EXERCISE 16.16. Suppose $E, F \to M$ are smooth vector bundles and $\Psi : E \to F$ is a map whose restriction to E_p for each p is a linear map $\Psi_p : E_p \to F_p$.

(a) Show that for every pair of smooth local trivializations $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^m$ and $\Psi_\beta : E|_{\mathcal{U}_\beta} \to \mathcal{U}_\beta \times \mathbb{F}^k$, there exists a unique function

$$\Psi_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathbb{F}^{k\times m}$$

such that

$$\Phi_\beta \circ \Psi \circ \Phi_\alpha^{-1} : (\mathcal{U}_\alpha \cap \mathcal{U}_\beta) \times \mathbb{F}^m \to (\mathcal{U}_\alpha \cap \mathcal{U}_\beta) \times \mathbb{F}^k : (p, v) \mapsto (p, \Psi_{\beta\alpha}(p)v).$$

(b) Show that Ψ is a smooth linear bundle map if and only if for all choices of the two smooth local trivializations in part (a), the function $\Psi_{\beta\alpha}$ is smooth.

Definition 16.17. Given a smooth vector bundle $E \to M$, a **smooth subbundle** (*Unterbündel*) of E is a vector bundle $F \to M$ such that for each $p \in M$, F_p is a linear subspace of E_p, and the inclusion $F \hookrightarrow E$ is a smooth linear bundle map.

16.2. Some basic examples. We now relate the definitions above to various examples that have already appeared in this course. For several of them, there is still some work to be done in showing that they naturally admit coverings by families of smoothly compatible local trivializations, and this work will be postponed until the next lecture.

Example 16.18 (**tangent bundle**). If M is an n-manifold, its tangent bundle TM is a smooth real vector bundle of rank n, where each smooth chart $(\mathcal{U}, x)$ determines a local trivialization $\Phi : TM|_{\mathcal{U}} = T\mathcal{U} \to \mathcal{U} \times \mathbb{R}^n$ by $\Phi(X) = (p, d_p x(X))$ for $X \in T_pM$. A smooth section of TM is nothing other than a smooth vector field on M,

$$\Gamma(TM) = \mathfrak{X}(M).$$

Example 16.19 (**cotangent bundle**). The cotangent bundle T^*M of a smooth n-manifold M has fibers $T_p^*M = \mathrm{Hom}(T_pM, \mathbb{R})$ for $p \in M$. We will construct smoothly compatible local trivializations for T^*M in the next lecture—it is a special case of the fact that every smooth vector bundle has a *dual bundle* which is also a smooth vector bundle in a natural way. The smooth sections of T^*M will then be the smooth 1-forms on M,

$$\Gamma(T^*M) = \Omega^1(M).$$

Example 16.20 (**tensor and exterior bundles**). For each $k, \ell \geqslant 0$, there is a natural smooth real vector bundle $T_\ell^k M \to M$ of rank $n^{k+\ell}$ whose fiber at a point p is the vector space $(T_pM)_\ell^k$ of $(k+\ell)$-fold multilinear maps $T_p^*M \times \ldots \times T_p^*M \times T_pM \times \ldots \times T_pM \to \mathbb{R}$. The smooth local trivializations on $T_\ell^k M$ will also arise from more general constructions to be discussed in the next lecture. Consistently with our previous notation, the space of smooth sections $\Gamma(T_\ell^k M)$ will then be precisely the space of smooth tensor fields of type (k, ℓ).

For each $k \geqslant 0$, there is an important subbundle

$$\Lambda^k T^*M \subset T_k^0 M$$

of rank $\binom{n}{k}$ whose fiber over $p \in M$ is the vector space of alternating k-forms $\Lambda^k T_p^* M \subset (T_pM)_k^0$. The sections of $\Lambda^k T^*M$ will of course be the smooth differential k-forms,

$$\Gamma(\Lambda^k T^*M) = \Omega^k(M).$$

Note that by definition,

$$T_1^0 M = T^*M = \Lambda^1 T^*M,$$

and since $(T_pM)_0^0$ is defined simply as $\mathbb{R}$ for every p, $T_0^0 M = \Lambda^0 T^*M$ is simply the *trivial* line bundle $M \times \mathbb{R}$ (cf. Example 16.21 below).

EXAMPLE 16.21 (**trivial bundle**). For any manifold M, the **trivial m-plane bundle** over M is the product $E = M \times \mathbb{F}^m$, with fibers

$$E_p := \{p\} \times \mathbb{F}^m,$$

understood in the obvious way as m-dimensional vector spaces. This is a smooth vector bundle because (M, Id) is a local trivialization that covers the entirety of M, so the associated maximal collection of local trivializations consists of all that are smoothly compatible with this one. Smooth sections $s : M \to M \times \mathbb{F}^m$ are smooth maps of the form $p \mapsto (p, f(p))$ and are thus equivalent to smooth functions $f : M \to \mathbb{F}^m$.

DEFINITION 16.22. A vector bundle $\pi : E \to M$ of rank m is (globally) **trivial**[52] if it admits a bundle isomorphism to the trivial m-plane bundle over M.

A local trivialization $\Phi : E|_{\mathcal{U}} \to \mathcal{U} \times \mathbb{F}^m$ of a vector bundle E can be understood as a bundle isomorphism between the restriction $E|_{\mathcal{U}} \to \mathcal{U}$ and the trivial m-plane bundle over $\mathcal{U}$. By definition, every vector bundle is therefore *locally* trivial, meaning that its restriction to any sufficiently small open subset must be trivial. The next example shows that globally, this need not be true.

EXAMPLE 16.23 (**a nontrivial real line bundle**). Identify S^1 with the unit circle in $\mathbb{C}$, and define $\ell \subset S^1 \times \mathbb{R}^2$ as the union of the sets $\{e^{i\theta}\} \times \ell_{e^{i\theta}} \subset S^1 \times \mathbb{R}^2$ for all $\theta \in \mathbb{R}$, where $\ell_{e^{i\theta}} \subset \mathbb{R}^2$ is the 1-dimensional subspace

$$\ell_{e^{i\theta}} = \mathbb{R}\begin{pmatrix} \cos(\theta/2) \\ \sin(\theta/2) \end{pmatrix} \subset \mathbb{R}^2.$$

Exercise 16.24 below shows that ℓ can be regarded as a smooth line bundle over S^1 with fibers $\ell_{e^{i\theta}}$ for $e^{i\theta} \in S^1$. Observe that if we consider the subset

$$\{(e^{i\theta}, v) \in \ell \mid \theta \in \mathbb{R},\ |v| \leqslant 1\}$$

consisting only of vectors of length at most 1, we obtain a Möbius strip. Local trivializations of $\ell \to S^1$ can be constructed as follows: for any $\theta_0 \in \mathbb{R}$, set $p := e^{i\theta_0} \in S^1$, and define

$$\Phi : \ell|_{S^1 \backslash \{p\}} \to (S^1 \backslash \{p\}) \times \mathbb{R} : \left(e^{i\theta}, c \begin{pmatrix} \cos(\theta/2) \\ \sin(\theta/2) \end{pmatrix} \right) \mapsto (e^{i\theta}, c), \tag{16.3}$$

with θ assumed to vary in the interval $(\theta_0, \theta_0 + 2\pi)$.

EXERCISE 16.24. For the line bundle $\ell \to S^1$ in Example 16.23, prove:

(a) Any two local trivializations defined as in (16.3) with different choices of $\theta_0 \in \mathbb{R}$ are smoothly compatible.
(b) ℓ is a smooth subbundle of the trivial bundle $S^1 \times \mathbb{R}^2$.
(c) There exists no continuous section of ℓ that is nowhere zero.
(d) ℓ is not globally trivial.

17. Constructions of vector bundles

17.1. Local frames and components. Local trivializations of a vector bundle are generally not very easy to visualize, which makes them tricky in practice to construct. We now introduce

[52]If we were being more pedantic, we would say **globally trivializable** in Definition 16.22 instead of "trivial", and reserve the latter for any vector bundle that is literally presented as a product $M \times \mathbb{F}^m$ with the identity map as a smooth trivialization, rather than just being isomorphic to one. But the looser use of the word "trivial" to mean "isomorphic to a trivial bundle" is widespread, so you should get used to it.

a notion that is equivalent, but arguably easier to work with. Recall that if $E \to M$ is a smooth vector bundle and $\mathcal{U} \subset M$ is a subset, we denote the union of all the fibers over points in $\mathcal{U}$ by

$$E|_{\mathcal{U}} := \bigcup_{p \in \mathcal{U}} E_p,$$

and call this the **restriction** of E to $\mathcal{U}$. It should be clear that if $\mathcal{U} \subset M$ is an *open* subset, then $E|_{\mathcal{U}}$ is a smooth vector bundle over $\mathcal{U}$ in a natural way. (We will generalize this below to the case where $\mathcal{U} \subset M$ is an arbitrary submanifold.) The space $\Gamma(E|_{\mathcal{U}})$ of smooth sections of $E|_{\mathcal{U}}$ thus consists of all smooth maps $\mathcal{U} \to E$ that send each point $p \in \mathcal{U}$ to an element of the corresponding fiber E_p. It often happens with bundles that a section $s \in \Gamma(E)$ with certain desirable properties cannot be assumed to exist *globally*, but does exist *locally*, meaning that for any sufficiently small open subset $\mathcal{U} \subset M$, a section of $E|_{\mathcal{U}}$ with those properties can be found. We sometimes refer to sections of the restricted bundle $E|_{\mathcal{U}}$ as **local sections** of E over the subset $\mathcal{U} \subset M$.

DEFINITION 17.1. For a vector bundle $E \to M$ and open set $\mathcal{U} \subset M$, a **frame** for E over $\mathcal{U}$ is a tuple of local sections $e_1, \dots, e_m : \mathcal{U} \to E$ of E over $\mathcal{U}$ such that for every $p \in \mathcal{U}$, the vectors $e_1(p), \dots, e_m(p)$ form a basis of E_p. We call $e_1, \dots, e_m$ a **smooth frame** if the sections are smooth.

Having a basis $e_1(p), \dots, e_m(p)$ for each fiber E_p means that in the region where the frame is defined, we can talk about **components**: every $v \in E_p$ for $p \in \mathcal{U}$ is of the form

$$v = v^j e_j \tag{17.1}$$

for unique real or complex numbers $v^1, \dots, v^m \in \mathbb{F}$. Note that the Einstein summation convention is in effect in (17.1), and we will continue using it in similar expressions wherever possible: since the possible values of j on the right hand side are $1, \dots, m$, it means in this case that there is an implied summation $\sum_{j=1}^m$ but the summation symbol has been omitted. Any section $s : M \to E$ is now uniquely expressible over $\mathcal{U}$ in terms of component functions $s^1, \dots, s^m : \mathcal{U} \to \mathbb{F}$, namely as

$$s(p) = s^j(p) e_j(p).$$

The proof of the following statement is more-or-less immediate:

PROPOSITION 17.2. *Over any open set $\mathcal{U} \subset M$, there is a natural bijective correspondence between frames $e_1, \dots, e_m : \mathcal{U} \to E$ and local trivializations $\Phi : E|_{\mathcal{U}} \to \mathcal{U} \times \mathbb{F}^m$, such that Φ is defined in terms of $e_1, \dots, e_m$ by*

$$\Phi(v^i e_i(p)) = (p, (v^1, \dots, v^m)).$$

Conversely, Φ determines $e_1, \dots, e_m$ by

$$e_i(p) = \Phi^{-1}(p, \mathbf{e}_i),$$

where $\mathbf{e}_1, \dots, \mathbf{e}_m$ denotes the standard basis of $\mathbb{F}^m$. □

EXAMPLE 17.3. On the tangent bundle $TM \to M$, the local trivialization determined by a chart $(\mathcal{U}, x)$ on M corresponds to the frame over $\mathcal{U}$ defined via the coordinate vector fields $\frac{\partial}{\partial x^1}, \dots, \frac{\partial}{\partial x^n} \in \Gamma(TM|_{\mathcal{U}})$.

Recall from the previous lecture that every local trivialization $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^m$ associates to each section $s : M \to E$ a function $s_\alpha : \mathcal{U}_\alpha \to \mathbb{F}^m$ such that $\Phi_\alpha(s(p)) = (p, s_\alpha(p))$. If $e_1^\alpha, \dots, e_m^\alpha$ denotes the local frame corresponding to Φ_α, then we can also write $s(p) = s^i(p) e_i^\alpha(p)$ for unique component functions $s^i : \mathcal{U}_\alpha \to \mathbb{F}$, and the correspondence in Proposition 17.2 gives

$$\Phi_\alpha(s(p)) = \Phi_\alpha(s^i(p) e_i^\alpha(p)) = (p, (s^1(p), \dots, s^m(p))) = (p, s_\alpha(p)).$$

This shows that the vector-valued local representation $s_\alpha : \mathcal{U}_\alpha \to \mathbb{F}^m$ is made up of the component functions $s^1, \dots, s^m$ with respect to the frame:

$$s_\alpha(p) = (s^1(p), \dots, s^m(p)) \in \mathbb{F}^m.$$

If E is a smooth vector bundle, we conclude from this and Proposition 16.12 that a section $s : M \to E$ is smooth if and only if for every smooth local trivialization $(\mathcal{U}, \Phi)$, the component functions $s^1, \dots, s^m : \mathcal{U} \to \mathbb{F}$ with respect to the corresponding local frame are smooth.

Now let's think about smooth compatibility: suppose $(\mathcal{U}_\alpha, \Phi_\alpha)$ and $(\mathcal{U}_\beta, \Phi_\beta)$ are two local trivializations related by the transition functions $g_{\beta\alpha}, g_{\alpha\beta} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{GL}(m, \mathbb{F})$, and denote the corresponding local frames by $e_1^\alpha, \dots, e_m^\alpha : \mathcal{U}_\alpha \to E$ and $e_1^\beta, \dots, e_m^\beta : \mathcal{U}_\beta \to E$. On $\mathcal{U}_\alpha \cap \mathcal{U}_\beta$, each of the sections e_i^α has uniquely-defined components with respect to the other frame $e_1^\beta, \dots, e_m^\beta$, giving functions $h_i^{\ j} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathbb{F}$ such that

$$e_i^\alpha = h_i^{\ j} e_j^\beta \qquad \text{on } \mathcal{U}_\alpha \cap \mathcal{U}_\beta.$$

For $p \in \mathcal{U}_\alpha \cap \mathcal{U}_\beta$, let us denote by $\mathbf{h}(p) \in \mathbb{F}^{m \times m}$ the matrix whose ith row and jth column is $h_i^{\ j}(p)$. For any $\mathbf{v} = (v^1, \dots, v^m) \in \mathbb{F}^m$, the correspondence in Proposition 17.2 then gives

$$\Phi_\beta \circ \Phi_\alpha^{-1}(p, \mathbf{v}) = \Phi_\beta(v^i e_i^\alpha(p)) = \Phi_\beta(v^i h_i^{\ j}(p) e_j^\beta(p)) = \Phi_\beta(v^j h_j^{\ i}(p) e_i^\beta(p)) = (p, \mathbf{h}(p)^T \mathbf{v}),$$

implying that the transition function $g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{GL}(m, \mathbb{F})$ is given by

$$g_{\beta\alpha}(p) = \mathbf{h}(p)^T.$$

In particular, the matrix-valued function $g_{\beta\alpha}$ is exactly as smooth as the least smooth among the scalar-valued functions $h_i^{\ j}$, which are simply the components of the sections $e_1^\alpha, \dots, e_m^\alpha$ with respect to the other frame. We've proved:

Proposition 17.4. *Two local frames correspond to smoothly compatible local trivializations if and only if the component functions for the sections in each frame with respect to the other frame are all smooth.* □

Corollary 17.5. *On a smooth vector bundle $E \to M$, a local trivialization $\Phi : E|_{\mathcal{U}} \to \mathcal{U} \times \mathbb{F}^m$ is smooth if and only if the sections forming the corresponding local frame $e_1, \dots, e_m : \mathcal{U} \to E$ are smooth.*

Proof. If Φ is a smooth local trivialization, then the local representations of the sections $e_1, \dots, e_m$ with respect to Φ are constant, and thus smooth, implying via Proposition 16.12 that the sections are smooth. Conversely, smoothness of the sections $e_1, \dots, e_m$ means that their components with respect to the frame corresponding to any smooth local trivialization are all smooth, which implies via Proposition 17.4 that Φ is smoothly compatible with any smooth local trivialization, and is therefore also smooth. □

17.2. Pullbacks and restrictions. Suppose $f : M \to N$ is a smooth map and $E \to N$ is a smooth vector bundle. The **pullback** of $E \to N$ via f, also known as the **induced bundle**, is a smooth vector bundle

$$f^*E \to M$$

whose fiber over the point $p \in M$ is

$$(f^*E)_p := E_{f(p)}.$$

To see that this is naturally a smooth vector bundle, suppose $\{\mathcal{U}_\alpha\}_{\alpha \in I}$ is an open covering of N with smoothly compatible local trivializations $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^m$, and write

$$\Phi_\alpha(v) = (p, \Phi_{\alpha,p} v) \qquad \text{for } p \in \mathcal{U}_\alpha,\ v \in E_p,$$

defining vector space isomorphisms $\Phi_{\alpha,p} : E_p \to \mathbb{F}^m$. The sets $\{f^{-1}(\mathcal{U}_\alpha) \subset M\}_{\alpha \in I}$ then form an open covering of M, and for each $\alpha \in I$, we can define a local trivialization $f^*\Phi_\alpha$ of f^*E by

$$f^*\Phi_\alpha : (f^*E)|_{f^{-1}(\mathcal{U}_\alpha)} \to f^{-1}(\mathcal{U}_\alpha) \times \mathbb{F}^m,$$
$$v \mapsto (p, \Phi_{\alpha,f(p)}v) \qquad \text{for } p \in \mathcal{U}_\alpha,\ v \in (f^*E)_p = E_{f(p)}.$$

The transition function $g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{GL}(m, \mathbb{F})$ relating Φ_α and Φ_β takes the form $g_{\beta\alpha}(p) = \Phi_{\beta,p}\Phi_{\alpha,p}^{-1}$ for $p \in \mathcal{U}_\alpha \cap \mathcal{U}_\beta$, thus the map $(f^*\Phi_\beta) \circ (f^*\Phi_\alpha)^{-1}$ from $(f^{-1}(\mathcal{U}_\alpha) \cap f^{-1}(\mathcal{U}_\beta)) \times \mathbb{F}^m$ to itself is given by

$$(f^*\Phi_\beta) \circ (f^*\Phi_\alpha)^{-1}(p, v) = (p, \Phi_{\beta,f(p)}\Phi_{\alpha,f(p)}^{-1}) = (p, g_{\beta\alpha}(f(p))), \qquad p \in f^{-1}(\mathcal{U}_\alpha) \cap f^{-1}(\mathcal{U}_\beta),$$

and the resulting transition function $f^{-1}(\mathcal{U}_\alpha) \cap f^{-1}(\mathcal{U}_\beta) \to \mathrm{GL}(m, \mathbb{F})$ is therefore $g_{\beta\alpha} \circ f$. This is smooth, so f^*E is a smooth bundle.

Remark 17.6. The above argument shows more generally that if $E \to N$ is a vector bundle of class C^k and $f : M \to N$ a map of class C^ℓ, then $f^*E \to M$ is a bundle of class $C^{\min\{k,\ell\}}$.

Exercise 17.7. In the situation above, show that the canonical map $f^*E \to E$ that sends $(f^*E)_p$ to $E_{f(p)}$ as the identity map for each $p \in M$ is smooth.

The map $f^*E \to E$ is a "fiberwise isomorphism" in the sense that it maps each fiber of f^*E isomorphically to a fiber of E, but it is not a *bundle map* in the sense defined in the previous lecture since f^*E and E are bundles over different manifolds. It is instead an example of the following more general notion:

Definition 17.8. Assume $E \to M$ and $F \to N$ are two smooth vector bundles and $\psi : M \to N$ is a smooth map. A smooth map $\Psi : E \to F$ that sends each fiber E_p linearly to the fiber $F_{\psi(p)}$ is called **smooth linear bundle map covering** ψ.

Our previous notion of smooth linear bundle maps was the special case of Definition 17.8 in which $M = N$ and $\psi : M \to N$ is the identity map. For a bundle $E \to N$ and map $f : M \to N$, we can now understand the canonical map $f^*E \to E$ as a smooth linear bundle map covering f.

Remark 17.9. Actually, a smooth linear bundle map $\Phi : E \to F$ covering a map $\psi : M \to N$ is more-or-less equivalent to a smooth linear bundle map from E to ψ^*F; the former is just the latter composed with the canonical map $\psi^*F \to F$.

Example 17.10. For a smooth map $f : M \to N$, the fiber of the pullback bundle f^*TN over a point $p \in M$ is the tangent space $T_{f(p)}N$, and a section $X \in \Gamma(f^*TN)$ therefore associates to each $p \in M$ a tangent vector $X(p) \in T_{f(p)}N$. Sections of this type are called **vector fields along** f; they generalize the usual notion of a vector field on M, which is the special case where $M = N$ and f is the identity map. These objects arise naturally in the following context: suppose $\{f_t : M \to N\}_{t \in (-\epsilon,\epsilon)}$ is a smooth 1-parameter family of maps with $f := f_0$, where "smooth family" in this situation means that the map $(-\epsilon, \epsilon) \times M \to N : (t, p) \mapsto f_t(p)$ is smooth. Then

$$X(p) := \partial_t f_t(p)|_{t=0} \in T_{f(p)}N$$

defines a vector field along f. Informally, if one thinks of $C^\infty(M, N)$ as an infinite-dimensional manifold, this means that its tangent space at $f \in C^\infty(M, N)$ is $\Gamma(f^*TM)$. (With minor modifications, this statement can be made precise in the language of smooth Banach manifolds.)

Example 17.11. If $N \subset M$ is a smooth submanifold and $i : N \hookrightarrow M$ denotes the inclusion map, then any smooth vector bundle $E \to M$ admits a **restriction** to N,

$$E|_N = i^*E \to N,$$

which is also a smooth vector bundle. (Its transition functions are just the restrictions of the transition functions of E to the submanifold.)

17.3. Subbundles, quotients, and normal bundles. The following result puts subbundles on a similar footing with sub*manifolds* by constructing the analogue of slice charts for local trivializations:

PROPOSITION 17.12. *Suppose $E \to M$ is a smooth vector bundle of rank m, $F \subset E$ is a subset, and denote for a point $p \in M$ or subset $\mathcal{U} \subset M$*

$$F_p := E_p \cap F, \qquad F|_{\mathcal{U}} := E|_{\mathcal{U}} \cap F.$$

The following statements are equivalent:

(1) *F is a smooth subbundle of rank k in the sense of Definition 16.17, i.e. it admits the structure of a smooth vector bundle of rank k such that the inclusion $F \hookrightarrow E$ is a smooth linear bundle map.*
(2) *For every $p \in M$, there exists a neighborhood $\mathcal{U} \subset M$ of p and a smooth local trivialization $\Phi : E|_{\mathcal{U}} \to \mathcal{U} \times \mathbb{F}^m$ of E such that*

$$\Phi(F|_{\mathcal{U}}) = \mathcal{U} \times (\mathbb{F}^k \times \{0\}) \subset \mathcal{U} \times \mathbb{F}^m.$$

PROOF. Suppose first that F is a smooth subbundle of rank k in the sense of Definition 16.17. Given $p \in M$, choose a smooth local trivialization $F|_{\mathcal{U}} \to \mathcal{U} \times \mathbb{F}^k$ of F with $p \in \mathcal{U}$ and let $e_1, \dots, e_k \in \Gamma(F|_{\mathcal{U}})$ denote the corresponding smooth local frame. Since the inclusion $F \hookrightarrow E$ is a smooth linear bundle map, the $e_1, \dots, e_k$ can equally well be regarded as smooth sections of $E|_{\mathcal{U}}$, and they are linearly independent at every point. After shrinking $\mathcal{U}$ if necessary, we can then use a local trivialization of E over $\mathcal{U}$ to find additional smooth sections $e_{k+1}, \dots, e_m \in \Gamma(E|_{\mathcal{U}})$ for which $e_1, \dots, e_m$ remain linearly independent and therefore form a basis of the fiber of E at every point in $\mathcal{U}$; the idea here is that in a local trivialization of E over $\mathcal{U}$, each of the sections $e_1, \dots, e_k$ is identified with a smooth function $\mathcal{U} \to \mathbb{F}^m$ that can be assumed *nearly* constant after shrinking $\mathcal{U}$, so that it is easy to find $m - k$ constant functions that complete the basis at every point. With this understood, we now have a smooth local frame $e_1, \dots, e_m \in \Gamma(E|_{\mathcal{U}})$ such that the sections $e_1, \dots, e_k$ span the fiber of F over every point in $\mathcal{U}$. The corresponding local trivialization then has the desired property.

Conversely, if local trivializations of E with this property always exist, then it is clear that the sets $F_p \subset E_p$ are linear subspaces and the trivializations of E determine smoothly compatible local trivializations of F by restriction. It is easy to check that the inclusion $F \hookrightarrow E$ is then a smooth map. (This step is analogous to the way that slice charts for a smooth submanifold $N \subset M$ are used to define a smooth structure on N so that the inclusion $N \hookrightarrow M$ is smooth.) □

REMARK 17.13. It will be useful in the following to allow a mild generalization of our previous notion of a local trivialization $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^m$. Specifically, nothing important changes if we replace the "standard" vector space $\mathbb{F}^m$ with any other m-dimensional vector space V and thus consider bijections of the form

$$\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times V$$

that map each fiber E_p isomorphically to $\{p\} \times V$. The transition functions relating two trivializations of this form take values in the group $\mathrm{GL}(V)$ of invertible $\mathbb{F}$-linear maps $V \to V$, which is an open subset of the (real or complex) vector space $\mathrm{End}(V)$. To reduce this to our previous notion, one only has to choose an isomorphism $\Psi : V \to \mathbb{F}^m$ and use it consistently, so that transition functions $g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{GL}(V)$ become

$$\widetilde{g}_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{GL}(m, \mathbb{F}) : p \mapsto \Psi g_{\beta\alpha}(p) \Psi^{-1};$$

clearly $g_{\beta\alpha}$ is smooth if and only if $\widetilde{g}_{\beta\alpha}$ is smooth.

Given a smooth vector bundle $E \to M$ of rank m and smooth subbundle $F \subset E$ of rank k, the **quotient bundle**

$$E/F \to M$$

is a smooth vector bundle of rank $m - k$ whose fiber over a point $p \in M$ is the quotient vector space

$$(E/F)_p := E_p/F_p.$$

One defines suitable local trivializations on E/F as follows: according to Proposition 17.12, we can find an open cover $\{\mathcal{U}_\alpha\}_{\alpha \in I}$ of M and local trivializations $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^m$ such that $\Phi_\alpha(F|_{\mathcal{U}_\alpha}) = \mathcal{U}_\alpha \times \mathbb{F}^k$, where $\mathbb{F}^k$ is identified with the linear subspace

$$\mathbb{F}^k := \mathbb{F}^k \times \{0\} \subset \mathbb{F}^m.$$

Writing $\Phi_\alpha(v) = (p, \Phi_{\alpha,p} v)$ for $p \in \mathcal{U}_\alpha$ and $v \in E_p$, it follows that the vector space isomorphism $\Phi_{\alpha,p} : E_p \to \mathbb{F}^m$ identifies the subspaces $F_p \subset E_p$ and $\mathbb{F}^k \subset \mathbb{F}^m$, thus it descends to an isomorphism of the quotient spaces,

$$\Phi_{\alpha,p} : E_p/F_p \to \mathbb{F}^m/\mathbb{F}^k : [v] \mapsto [\Phi_{\alpha,p} v].$$

A local trivialization

$$(E/F)|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times (\mathbb{F}^m/\mathbb{F}^k)$$

in the generalized sense of Remark 17.13 can thus be defined by sending $[v] \in E_p/F_p$ for $p \in \mathcal{U}_\alpha$ to $(p, [\Phi_{\alpha,p} v])$. Covering E/F with local trivializations defined in this way, the resulting transition functions are derived from the transition functions $g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{GL}(m, \mathbb{F})$ of E by observing that since we chose the Φ_α to respect the subbundle $F \subset E$ as in Proposition 17.12, the linear map $g_{\beta\alpha}(p) : \mathbb{F}^m \to \mathbb{F}^m$ preserves the subspace $\mathbb{F}^k \subset \mathbb{F}^m$ for each $p \in \mathcal{U}_\alpha \cap \mathcal{U}_\beta$ and thus descends to an isomorphism on the quotient $\mathbb{F}^m/\mathbb{F}^k$, determining a smooth function

$$\mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{GL}(\mathbb{F}^m/\mathbb{F}^k).$$

This is the transition function for the local trivializations we defined above on E/F from Φ_α and Φ_β.

Exercise 17.14. Show that for a smooth subbundle $F \subset E$ of $E \to M$, the natural surjective map $E \to E/F$ that restricts to the fiber over each point $p \in M$ as the quotient projection $E_p \to E_p/F_p$ is a smooth linear bundle map.

Example 17.15. Suppose $N \subset M$ is a smooth k-dimensional submanifold in an m-manifold M. Any slice chart for N determines a local trivialization of TM that also has the property in Proposition 17.12 for the subset $TN \subset TM|_N$, thus producing a smooth subbundle

$$TN \subset TM|_N.$$

The quotient

$$\nu N := (TM|_N)/TN \to N$$

is called the **normal bundle** of the submanifold $N \subset M$.

One can gain a better intuitive picture of the normal bundle of a submanifold $N \subset M$ by choosing a Riemannian metric g on M and looking at the orthogonal complements

$$(T_pN)^\perp := \left\{ X \in T_pM \;\middle|\; g(X, \cdot)|_{T_pN} = 0 \right\}$$

at points $p \in N$.

Exercise 17.16. Given a smooth submanifold N in a Riemannian manifold (M, g), prove:

(a) $TN^\perp := \bigcup_{p \in N} (T_pN)^\perp$ is a smooth subbundle of $TM|_N$.
Hint: Construct smooth local frames $X_1, \ldots, X_n$ for $TM|_N$ such that $X_1, \ldots, X_k$ are tangent to N and $X_{k+1}, \ldots, X_n$ lie in $(TN)^\perp$.

(b) The composition of the inclusion $TN^\perp \hookrightarrow TM|_N$ with the fiberwise quotient projection $TM|_N \to TM|_N/TN$ from Exercise 17.14 defines a bundle isomorphism $TN^\perp \to \nu N$.

The "normal vector fields" along hypersurfaces $N \subset M$ we considered in Lectures 11 and 12 can now be understood as smooth sections of the bundle $(TN)^\perp$, which according to Exercise 17.16, is equivalent to the normal bundle of N.

17.4. Algebraic operations. Several natural operations that produce new vector spaces from old ones can now be generalized to the setting of vector bundles. In the following list, the smoothness of the bundles we construct can be verified easily by constructing local frames; we will leave the details as exercises.

17.4.1. *Direct sums.* The direct sum of two vector spaces V and W is the same thing as their Cartesian product,

$$V \oplus W := V \times W,$$

in which V and W can be identified naturally with the subspaces $V \times \{0\}$ and $\{0\} \times W$ respectively. When extending this notion to vector bundles, it becomes especially useful to distinguish between the symbols "$\oplus$" and "$\times$": in particular, the direct sum of two smooth vector bundles $E, F \to M$ of ranks m and k respectively is a bundle $E \oplus F \to M$ of rank $m + k$ with fibers

$$(E \oplus F)_p := E_p \oplus F_p = E_p \times F_p.$$

Notice that at the level of sets, the total space $E \oplus F = \bigcup_{p \in M}(E_p \times F_p)$ is not *at all* the same thing as the product $E \times F$. Any local trivializations of E and F over the same region can be combined in a natural way to produce a local trivialization of $E \oplus F$ over that region, and if one covers $E \oplus F$ with local trivializations constructed in this way with respect to an open covering $\{\mathcal{U}_\alpha\}_{\alpha \in I}$, one finds that the resulting transition functions $g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{GL}(m + k, \mathbb{F})$ take the form of block matrices,

$$g_{\beta\alpha}(p) = \begin{pmatrix} g^E_{\beta\alpha}(p) & 0 \\ 0 & g^F_{\beta\alpha}(p) \end{pmatrix},$$

where $g^E_{\beta\alpha}(p) \in \mathrm{GL}(m, \mathbb{F})$ and $g^F_{\beta\alpha}(p) \in \mathrm{GL}(k, \mathbb{F})$ are the transition functions for E and F respectively. Clearly $g_{\beta\alpha}$ is smooth if $g^E_{\beta\alpha}$ and $g^F_{\beta\alpha}$ are.

Remark 17.17. One *can* define a "product" bundle $E \times F$ whose total space is the Cartesian product of E and F, but it is naturally a bundle over $M \times M$ rather than M. More generally, two bundles $E \to M$ and $F \to N$ over potentially different manifolds have a product which is a bundle over $M \times N$

17.4.2. *The dual bundle.* Any smooth vector bundle $E \to M$ has a **dual bundle**

$$E^* \to M$$

whose fiber over a point $p \in M$ is the dual space $E^*_p = \mathrm{Hom}(E_p, \mathbb{F})$. Any local frame $e_1, \dots, e_m$ for E over an open subset $\mathcal{U} \subset M$ then determines a dual frame $e^1_*, \dots, e^m_*$ for E^* via the usual notion of a dual basis, i.e. for each $p \in \mathcal{U}$,

$$e^i_*(p)\,(e_j(p)) = \delta^i_j.$$

It is a straightforward algebraic exercise to verify that whenever two frames for E on overlapping regions are smoothly compatible in the sense of Proposition 17.4, their dual frames are also smoothly compatible, so in this way one can cover E^* with smoothly compatible local trivializations, making it a smooth vector bundle. This establishes in particular that for any smooth n-manifold M, the **cotangent bundle**

$$T^*M \to M$$

is a smooth real vector bundle of rank n.

17.4.3. *The complex conjugate.* [53]

One can associate to any complex vector space V another complex vector space $\bar{V}$, which is defined as the same set with the same notion of vector addition but a different notion of scalar multiplication, defined as follows. Since V and $\bar{V}$ are identical sets, the identity map defines a canonical map between them, which we shall denote by

$$V \to \bar{V} : v \mapsto \bar{v}. \tag{17.2}$$

In other words, $\bar{v}$ is our notation for the vector $v \in V$ when it is regarded as an element of $\bar{V}$. With this understood, multiplication of a scalar $\lambda \in \mathbb{C}$ by a vector $\bar{v} \in \bar{V}$ is defined by

$$\lambda\bar{v} := \overline{\bar{\lambda}v},$$

where for $\lambda = a + ib$ with $a, b \in \mathbb{R}$, we denote the complex conjugate by $\bar{\lambda} := a - ib$. Another way to say this is that multiplication of a *real* scalar by a vector in $\bar{V}$ is defined exactly the same as in V, so that V and $\bar{V}$ are identical as real vector spaces, but multiplication by i is defined in $\bar{V}$ with a sign change, i.e. $i\bar{v} = -\overline{iv}$. This makes the bijection in (17.2) an isomorphism of real vector spaces, but *not* an isomorphism of complex vector spaces, as it is not even a complex-linear map; it is instead complex *antilinear*.

The **conjugate** of a complex vector bundle $E \to M$ of rank m is now defined as another complex vector bundle

$$\bar{E} \to M$$

of rank m whose fiber over a point $p \in M$ is $\bar{E}_p$. Strictly speaking, E and $\bar{E}$ are the same set, and the identity map thus defines a canonical bijection between them which we will again denote by

$$E \to \bar{E} : v \mapsto \bar{v}.$$

They are different complex vector bundles because one cannot use the same local trivializations for both—any local trivialization $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{C}^m$ of E defines a complex vector space isomorphism $\Phi_{\alpha,p} : E_p \to \mathbb{C}^m$ for every $p \in \mathcal{U}_\alpha$, but this map is not complex-linear when regarded as a bijection $\bar{E}_p \to \mathbb{C}^m$, it is antilinear. The solution is to compose it with a complex-antilinear isomorphism $\mathbb{C}^m \to \mathbb{C}^m$ such as the complex conjugation map $z \mapsto \bar{z}$, and this produces a local trivialization $\bar{\Phi}_\alpha$ of $\bar{E}$ over the same set $\mathcal{U}_\alpha$, namely

$$\bar{E}|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{C}^m : \bar{v} \mapsto (p, \overline{\Phi_{\alpha,p} v})$$

for $p \in \mathcal{U}_\alpha$ and $\bar{v} \in \bar{E}_p$. The next exercise shows that the collection of all local trivializations of $\bar{E}$ constructed in this way makes $\bar{E} \to M$ a smooth vector bundle.

EXERCISE 17.18. Show that if $(\mathcal{U}_\alpha, \Phi_\alpha)$ and $(\mathcal{U}_\beta, \Phi_\beta)$ are two local trivializations of a complex vector bundle $E \to M$ related by a transition function $g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{GL}(m, \mathbb{C})$, then the transition function relating the corresponding local trivializations $(\mathcal{U}_\alpha, \bar{\Phi}_\alpha)$ and $(\mathcal{U}_\beta, \bar{\Phi}_\beta)$ of $\bar{E}$ is given by

$$\mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{GL}(m, \mathbb{C}) : p \mapsto \overline{g_{\beta\alpha}(p)},$$

where the bar on the right hand side means the usual notion of complex conjugation for m-by-m complex matrices. In particular, this transition function is smooth if and only if $g_{\beta\alpha}$ is smooth.

A finite-dimensional complex vector space V is always isomorphic to its conjugate space $\bar{V}$ since the two spaces have the same dimension, but on the other hand, there is no *canonical* choice of isomorphism (remember that the map $v \mapsto \bar{v}$ does not count because it is not complex linear). Vector bundles provide a means for measuring in some precise way the non-existence of a canonical choice: we will see that in general, a complex vector bundle $E \to M$ and its conjugate bundle $\bar{E} \to M$ need not be isomorphic.

[53] This section is inessential and was skipped in the lecture, but is provided here for your information.

REMARK 17.19. In the setting of a complex manifold M, whose transition maps are holomorphic maps between open subsets of $\mathbb{C}^n$ so that the notion of holomorphic complex-valued functions on open subsets of M can be defined, one can also consider so-called *holomorphic vector bundles*, which are required to admit coverings by local trivializations such that all transition functions are holomorphic, and the notion of holomorphic sections can therefore be defined. Exercise 17.18 shows that if $E \to M$ is a holomorphic vector bundle, then its conjugate $\bar{E} \to M$ is naturally a smooth complex vector bundle but is not a *holomorphic* vector bundle in any natural way, as its transition functions are not holomorphic, they are antiholomorphic (i.e. they are complex conjugates of holomorphic functions).

17.4.4. *Tensor products.* We will not discuss in this course the abstract definition of the tensor product of two vector spaces V and W, but if both spaces are finite dimensional, one obtains an easy equivalent definition using the canonical identification of V and W with the duals of their dual spaces. For our purposes, if V and W have dual spaces V^* and W^*, then $V^* \otimes W^*$ can be defined as the vector space

$$V^* \otimes W^* := \{\text{bilinear maps } V \times W \to \mathbb{F}\},$$

with the tensor product $\lambda \otimes \mu \in V^* \otimes W^*$ of two elements $\lambda^* \in V^*$ and $\mu^* \in W^*$ defined by

$$(\lambda \otimes \mu)(v, w) := \lambda(v)\mu(w) \qquad \text{for } v \in V,\ w \in W.$$

Our definition of $V \otimes W$ is then actually a definition of $V^{**} \otimes W^{**}$, i.e.

$$V \otimes W := \{\text{bilinear maps } V^* \times W^* \to \mathbb{F}\},$$

and $v \otimes w \in V \otimes W$ for $v \in V$ and $w \in W$ will be the bilinear map $V^* \times W^* \to \mathbb{F}$ given by

$$(v \otimes w)(\lambda, \mu) := v(\lambda)w(\mu) := \lambda(v)\mu(w) \qquad \text{for } \lambda \in V^*,\ \mu \in W^*.$$

Given bases $v_1, \ldots, v_m \in V$ and $w_1, \ldots, w_k \in W$, one can easily check via evaluation on the corresponding dual bases of V^* and W^* that the mk distinct tensor products $v_i \otimes w_j$ form a basis of $V \otimes W$.

While it is a bit tedious from this perspective, one can also check that the tensor product is an associative operation, i.e. for any three finite-dimensional vector spaces V, W, X, there is a natural isomorphism

$$(V \otimes W) \otimes X \cong V \otimes (W \otimes X)$$

that identifies $(v \otimes w) \otimes x$ with $v \otimes (w \otimes x)$ for every $v \in V$, $w \in W$ and $x \in X$. For this reason we will usually not write the parentheses in such expressions, and arbitrary tensor products of finitely many finite-dimensional vector spaces $V_1, \ldots, V_k$ can also be defined without parentheses; in fact, there is a natural isomorphism of $V_1 \otimes \ldots \otimes V_k$ with the space of multilinear maps $V_1^* \times \ldots \times V_k^* \to \mathbb{F}$.

All of this extends to the context of smooth vector bundles E over a manifold M, after observing that the canonical isomorphisms $E_p \to E_p^{**}$ give rise to canonical bundle isomorphisms $E \to E^{**} := (E^*)^*$. For two smooth vector bundles $E, F \to M$ of ranks m and k respectively, the tensor product $E \otimes F \to M$ is thus a bundle of rank mk with fibers

$$(E \otimes F)_p := E_p \otimes F_p.$$

Given local frames $e_1, \ldots, e_m$ for E and $f_1, \ldots, f_k$ for F over $\mathcal{U}$, a local frame for $E \otimes F$ over $\mathcal{U}$ is given by the sections

$$e_i \otimes f_j : \mathcal{U} \to (E \otimes F)|_{\mathcal{U}}, \qquad i = 1, \ldots, m,\ j = 1, \ldots, k,$$

where the tensor product of sections is defined pointwise, meaning $(e_i \otimes f_j)(p) := e_i(p) \otimes f_j(p) \in E_p \otimes F_p$. It is again a straightforward exercise to check that for any smoothly compatible choices of frames for E and F on overlapping regions, the resulting frames for $E \otimes F$ are also smoothly compatible.

This discussion extends in an obvious way to arbitrary finite tensor products of vector bundles. In particular, we can now generalize the bundles $T^k_\ell M$ mentioned in Example 16.20 to

$$E^k_\ell := \underbrace{E \otimes \ldots \otimes E}_{k} \otimes \underbrace{E^* \otimes \ldots \otimes E^*}_{\ell},$$

with the convention that for $k = \ell = 0$, the fibers are just $(E_p)^0_0 := \mathbb{R}$ and E^0_0 is thus the trivial real line bundle $M \times \mathbb{R}$ over M. We also have $E^1_0 = E$ and $E^0_1 = E^*$.

For each $k \geqslant 0$, there is an important subbundle

$$\Lambda^k E \subset \underbrace{E \otimes \ldots \otimes E}_{k},$$

whose fibers are the spaces $\Lambda^k E_p$ of antisymmetric k-fold multilinear maps $E^*_p \times \ldots \times E^*_p \to \mathbb{R}$, i.e. in terms of our notation from Lecture 9, we are defining $\Lambda^k E_p := \Lambda^k V^*$ for $V := E^*_p$ after identifying E_p with its double dual. That $\Lambda^k E \subset E^{\otimes k}$ is a smooth subbundle follows mainly from the observation that any local frame $e_1, \ldots, e_m$ for E gives rise to a local frame for $\Lambda^k E$ on the same region, consisting of the k-fold wedge products

$$e_{i_1} \wedge \ldots \wedge e_{i_k}, \qquad i_1 < \ldots < i_k.$$

In particular, this makes $\Lambda^k T^* M$ into a smooth vector bundle with $\Gamma(\Lambda^k T^* M) = \Omega^k(M)$.

17.4.5. *Bundles of linear maps.* [54]

It is often useful to notice that for two vector spaces V, W, the space of linear maps $\mathrm{Hom}(V, W)$ is naturally isomorphic to the tensor product $V^* \otimes W$. Indeed:

Exercise 17.20. Show that for finite-dimensional vector spaces V and W, the identifying $\lambda \otimes w \in V^* \otimes W$ for each $\lambda \in V^*$ and $w \in W$ with the linear map $V \to W$ given by

$$(\lambda \otimes w)(v) := \lambda(v) w$$

uniquely determines an isomorphism $V^* \otimes W \to \mathrm{Hom}(V, W)$.

This gives us the quickest way to see that for any two smooth vector bundles $E, F \to M$ with rank m and k respectively, there exists a smooth vector bundle

$$\mathrm{Hom}(E, F) \to M$$

with rank km, having fibers $\mathrm{Hom}(E, F)_p := \mathrm{Hom}(E_p, F_p)$. In fact, $\mathrm{Hom}(E, F)$ is canonically isomorphic to the tensor product bundle $E^* \otimes F$, but without worrying about this, one can also just take Exercise 17.20 as a hint on how to define local frames for $\mathrm{Hom}(E, F)$: given frames $e_1, \ldots, e_m$ for E and $f_1, \ldots, f_k$ for F over a region $\mathcal{U} \subset M$, one takes the dual frame $e^1_*, \ldots, e^m_*$ for E^* and defines a frame for $\mathrm{Hom}(E, F)$ over $\mathcal{U}$ consisting of the products $e^i_* \otimes f_j$, each interpreted at any point $p \in \mathcal{U}$ as the linear map $E_p \to F_p : v \mapsto e^i_*(p)(v) f_j(p)$. It is another straightforward exercise to show that any two local frames for $\mathrm{Hom}(E, F)$ constructed from smooth frames on E and F will be smoothly compatible.

We can now state a much more succinct version of one of the definitions in the previous lecture: given two smooth vector bundles $E, F \to M$, a *smooth linear bundle map* $E \to F$ is a smooth section of the bundle $\mathrm{Hom}(E, F)$.

In the case $\mathbb{F} = \mathbb{C}$, it is sometimes also useful to include complex *anti*-linear maps in the discussion, where a map $A : V \to W$ between two complex vector spaces is called **antilinear** if it satisfies

$$A(v + w) = Av + Aw, \qquad A(\lambda v) = \bar{\lambda} A v$$

[54]This section is provided for your information and will occasionally be referred to later, but it was not covered in the lecture due to lack of time.

for all $v, w \in V$ and $\lambda \in \mathbb{C}$. The space

$$\overline{\mathrm{Hom}}(V, W) := \{A : V \to W \mid A \text{ complex antilinear}\}$$

is a complex vector space in a natural way, and the following exercise yields a useful alternative perspective on it in terms of the conjugate vector space (see §17.4.3).

EXERCISE 17.21. Assume V, W are finite-dimensional complex vector spaces with dual space V^*, W^* and conjugates $\bar{V}, \bar{W}$. Find natural isomorphisms between the following pairs of complex vector spaces.

(a) $\mathrm{Hom}(\bar{V}, W)$ and $\overline{\mathrm{Hom}}(V, W)$.
(b) $(\bar{V})^*$ and $\overline{V^*}$.
(c) $\bar{V}^* \otimes W^*$ and the space of real-bilinear maps $V \times W \to \mathbb{C}$ that are complex antilinear in the first factor and complex linear in the second factor.

It follows from Exercise 17.21 that for smooth complex vector bundles $E, F \to M$, one can also define a smooth bundle

$$\overline{\mathrm{Hom}}(E, F) \to M$$

whose fiber at a point $p \in M$ is the space of complex-antilinear maps $E_p \to F_p$; this bundle is canonically isomorphic to $\bar{E}^* \otimes F$.

18. Vector bundles with extra structure

In this lecture we discuss various types of geometric structure that can be added to the fibers of a vector bundle, such as orientations and inner products. There is a useful way to incorporate all possible types of structures under a single umbrella in terms of the so-called *structure group* of a bundle, and this discussion requires an initial digression on the topic of Lie groups.

18.1. Some basic Lie groups. Roughly speaking, a Lie group is a group that is also a smooth manifold. I intend to discuss this subject in earnest in the followup to this course next semester, but for now, we need to become acquainted with a few of the basic examples and their properties. The first one is $\mathrm{GL}(m, \mathbb{F})$, which is naturally a manifold because it is an open subset of the (real or complex) vector space $\mathbb{F}^{m \times m}$.

DEFINITION 18.1. A **Lie subgroup of** $\mathrm{GL}(m, \mathbb{F})$ is a subgroup $G \subset \mathrm{GL}(m, \mathbb{F})$ that is also a smooth submanifold. Its associated **Lie algebra** is the tangent space

$$\mathfrak{g} := T_{\mathbb{1}} G \subset T_{\mathbb{1}} \mathrm{GL}(m, \mathbb{F}) = \mathbb{F}^{m \times m}.$$

The discussion of why $\mathfrak{g} = T_{\mathbb{1}} G$ is called a "Lie algebra" will have to wait for next semester; for our immediate purposes, it will be enough to notice that $\mathfrak{g}$ is a linear subspace of $\mathbb{F}^{m \times m}$.

It will sometimes be useful to observe that the natural maps defined by matrix multiplication

$$\mathrm{GL}(m, \mathbb{F}) \times \mathrm{GL}(m, \mathbb{F}) \to \mathrm{GL}(m, \mathbb{F}) : (\mathbf{A}, \mathbf{B}) \mapsto \mathbf{AB}$$

and inversion

$$\mathrm{GL}(m, \mathbb{F}) \to \mathrm{GL}(m, \mathbb{F}) : \mathbf{A} \mapsto \mathbf{A}^{-1}$$

are both smooth. Indeed, the first is simply a quadratic function of the entries of $\mathbf{A}$ and $\mathbf{B}$, and by Cramer's rule, the second is $1/\det(\mathbf{A})$ times a polynomial function of the entries, where $\det(\mathbf{A})$ is itself a polynomial function of the entries and is nonzero as long as we restrict to the open subset $\mathrm{GL}(m, \mathbb{F}) \subset \mathbb{F}^{m \times m}$. Since restrictions of smooth maps to smooth submanifolds are also smooth, it follows that for every Lie subgroup $G \subset \mathrm{GL}(m, \mathbb{F})$, the maps

$$G \times G \to G : (\mathbf{A}, \mathbf{B}) \mapsto \mathbf{AB}, \qquad G \to G : \mathbf{A} \mapsto \mathbf{A}^{-1}$$

are both smooth.

EXAMPLE 18.2. The **orthogonal** group $\mathrm{O}(m) \subset \mathrm{GL}(m, \mathbb{R})$ consists of all linear transformations $\mathbb{R}^n \to \mathbb{R}^n$ that preserve the standard Euclidean inner product $\langle \mathbf{v}, \mathbf{w} \rangle_{\mathbb{R}^m} := \mathbf{v}^T \mathbf{w} = \sum_{j=1}^m v^j w^j$. It is a Lie subgroup according to Exercise 4.22, and its Lie algebra is the space of real antisymmetric matrices

$$\mathfrak{o}(m) := \left\{ \mathbf{A} \in \mathbb{R}^{m \times m} \;\middle|\; \mathbf{A}^T = -\mathbf{A} \right\}.$$

EXAMPLE 18.3. The **unitary** group $\mathrm{U}(m) \subset \mathrm{GL}(m, \mathbb{C})$ consists of all linear transformations $\mathbb{C}^n \to \mathbb{C}^n$ that preserve the standard Hermitian inner product $\langle \mathbf{v}, \mathbf{w} \rangle_{\mathbb{C}^m} := \mathbf{v}^\dagger \mathbf{w} = \sum_{j=1}^m \bar{v}^j w^j$. It is a Lie subgroup according to Exercise 4.23, and its Lie algebra is the space of complex anti-Hermitian matrices

$$\mathfrak{u}(m) := \left\{ \mathbf{A} \in \mathbb{C}^{m \times m} \;\middle|\; \mathbf{A}^\dagger = -\mathbf{A} \right\}.$$

EXAMPLE 18.4. The group of orientation-preserving linear transformations on $\mathbb{R}^n$ is

$$\mathrm{GL}_+(m, \mathbb{R}) := \left\{ \mathbf{A} \in \mathrm{GL}(m, \mathbb{R}) \;\middle|\; \det(\mathbf{A}) > 0 \right\},$$

which is both a subgroup and an open subset of $\mathrm{GL}(m, \mathbb{R})$, and therefore a Lie subgroup. Since it is also an open subset of $\mathbb{R}^{m \times m}$, its Lie algebra is

$$\mathfrak{gl}_+(m, \mathbb{R}) = \mathfrak{gl}(m, \mathbb{R}) := \mathbb{R}^{m \times m}.$$

EXAMPLE 18.5. Exercise 4.24 shows that for $\mathbb{F}$ equal to either $\mathbb{R}$ or $\mathbb{C}$, the **special linear** group $\mathrm{SL}(m, \mathbb{F}) := \left\{ \mathbf{A} \in \mathrm{GL}(m, \mathbb{F}) \;\middle|\; \det(\mathbf{A}) = 1 \right\}$ is a Lie subgroup whose Lie algebra consists of the traceless matrices,

$$\mathfrak{sl}(m, \mathbb{F}) := \left\{ \mathbf{A} \in \mathbb{F}^{m \times m} \;\middle|\; \mathrm{tr}(\mathbf{A}) = 0 \right\}.$$

The special linear group consists of all linear transformations $\mathbb{F}^m \to \mathbb{F}^m$ that preserve the "standard" alternating m-form

$$\mu(\mathbf{v}_1, \ldots, \mathbf{v}_m) := \det \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_m \end{pmatrix}. \tag{18.1}$$

In the real case, one obtains a useful geometric interpretation by relating $\mathrm{SL}(m, \mathbb{R})$ to the larger group

$$\widehat{\mathrm{SL}}(m, \mathbb{R}) := \left\{ \mathbf{A} \in \mathrm{GL}(m, \mathbb{R}) \;\middle|\; \det(\mathbf{A}) \in \{1, -1\} \right\},$$

which is the group of all volume-preserving linear transformations on $\mathbb{R}^n$. Since

$$\mathrm{SL}(m, \mathbb{R}) = \widehat{\mathrm{SL}}(m, \mathbb{R}) \cap \mathrm{GL}_+(m, \mathbb{R}),$$

$\mathrm{SL}(m, \mathbb{R})$ therefore consists of all linear transformations that preserve both orientation and volume.

EXAMPLE 18.6. The **special orthogonal** group $\mathrm{SO}(m) := \mathrm{O}(m) \cap \mathrm{SL}(m, \mathbb{R})$ is an open subset of $\mathrm{O}(m)$, and thus has the same Lie algebra,

$$\mathfrak{so}(m) = \mathfrak{o}(m),$$

which is contained in $\mathfrak{sl}(m, \mathbb{R})$ since real antisymmetric matrices vanish along the diagonal. Since every $\mathbf{A} \in \mathrm{O}(m)$ has $\det(\mathbf{A}) = \pm 1$, one could equally well write

$$\mathrm{SO}(m) = \mathrm{O}(m) \cap \mathrm{GL}_+(m, \mathbb{R}),$$

and thus interpret $\mathrm{SO}(m)$ as the group of all orientation-preserving orthogonal transformations.

EXAMPLE 18.7. The complex analogue of $\mathrm{SO}(m)$ is the **special unitary** group $\mathrm{SU}(m) := \mathrm{U}(m) \cap \mathrm{SL}(m, \mathbb{C})$, but there is a qualitative difference from the real case: according to Exercise 4.25, $\mathrm{SU}(m)$ is also a Lie subgroup, but its dimension is one less than that of $\mathrm{U}(m)$, and its Lie algebra

$$\mathfrak{su}(m) := \mathfrak{u}(m) \cap \mathfrak{sl}(m, \mathbb{C})$$

is the space of matricies that are both anti-Hermitian *and* traceless, which is not identical to $\mathfrak{u}(m)$ since anti-Hermitian matrices can have arbitrary imaginary entries on the diagonal. One can interpret $\mathrm{SU}(m)$ as the group of linear transformations on $\mathbb{C}^m$ that preserve both the standard Hermitian inner product and the alternating m-form μ in (18.1).

EXAMPLE 18.8. The following generalization of the orthogonal group is important in physics: given integers $k, \ell \geqslant 0$ with $k + \ell = m$, the **indefinite orthogonal group**

$$\mathrm{O}(k, \ell) \subset \mathrm{GL}(m, \mathbb{R})$$

consists of all linear transformations $\mathbf{A} : \mathbb{R}^m \to \mathbb{R}^m$ that satisfy $\langle \mathbf{A}\mathbf{v}, \mathbf{A}\mathbf{w}\rangle_{k,\ell} = \langle \mathbf{v}, \mathbf{w}\rangle_{k,\ell}$ for all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$, where

$$\langle \mathbf{v}, \mathbf{w}\rangle_{k,\ell} := \sum_{j=1}^{k} v^j w^j - \sum_{j=k+1}^{m} v^j w^j.$$

The case $\mathrm{O}(1, 3)$ is known as the **Lorentz group** and plays a fundamental role in relativity, where the sign difference in $\langle\ ,\ \rangle_{1,3}$ between the first and the other three coordinates gives a qualitative distinction between the three dimensions of physical space and a fourth dimension, interpreted as time. There is also a complex analogue, the **indefinite unitary group** $\mathrm{U}(k, \ell) \subset \mathrm{GL}(m, \mathbb{C})$.

EXERCISE 18.9. For integers $k, \ell \geqslant 0$ with $k+\ell = m$, define the block matrix $\boldsymbol{\eta} := \begin{pmatrix} \mathbb{1}_k & 0 \\ 0 & -\mathbb{1}_\ell \end{pmatrix} \in \mathrm{GL}(m, \mathbb{R})$, where for any $q \geqslant 0$ we write $\mathbb{1}_q$ for the q-by-q identity matrix. Prove:

(a) A matrix $\mathbf{A} \in \mathbb{R}^{m\times m}$ belongs to $\mathrm{O}(k, \ell)$ if and only if $\mathbf{A}\boldsymbol{\eta}\mathbf{A}^T\boldsymbol{\eta} = \mathbb{1}$.
(b) Every $\mathbf{A} \in \mathrm{O}(k, \ell)$ satisfies $\det(\mathbf{A}) = \pm 1$.
(c) $\mathrm{O}(k, \ell)$ is a smooth submanifold and thus a Lie subgroup of $\mathrm{GL}(m, \mathbb{R})$, with Lie algebra

$$\mathfrak{o}(k, \ell) := \left\{ \mathbf{A} \in \mathbb{R}^{m\times m} \;\middle|\; \mathbf{A}^* = -\mathbf{A} \right\},$$

where $\mathbf{A}^* := \boldsymbol{\eta}\mathbf{A}^T\boldsymbol{\eta}$.
Hint: For every $\mathbf{A} \in \mathrm{GL}(m, \mathbb{R})$, $\mathbf{A}\boldsymbol{\eta}\mathbf{A}^T\boldsymbol{\eta}$ belongs to the vector space $\left\{ \mathbf{H} \in \mathbb{R}^{n\times n} \;\middle|\; \mathbf{H}^* = \mathbf{H} \right\}$.

EXAMPLE 18.10. [55] There is a natural way of regarding $\mathrm{GL}(m, \mathbb{C})$ as a Lie subgroup of $\mathrm{GL}(2m, \mathbb{R})$. The idea is to identify $\mathbb{C}^m$ with $\mathbb{R}^{2m}$ via the real-linear isomorphism

$$\mathbb{C}^m \to \mathbb{R}^{2m} = \mathbb{R}^m \times \mathbb{R}^m : \mathbf{x} + i\mathbf{y} \mapsto (\mathbf{x}, \mathbf{y}),$$

so that scalar multiplication by i becomes the linear transformation $\mathbb{R}^{2m} \to \mathbb{R}^{2m}$ defined by the matrix

$$\mathbf{J}_0 := \begin{pmatrix} 0 & -\mathbb{1}_m \\ \mathbb{1}_m & 0 \end{pmatrix}. \tag{18.2}$$

A linear transformation $\mathbf{A} : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ then represents a complex-linear transformation on $\mathbb{C}^n$ if and only if it commutes with the matrix $\mathbf{J}_0$, giving an identification of $\mathbb{C}^{m\times m}$ with the linear subspace

$$\mathrm{End}_{\mathbb{C}}(\mathbb{R}^{2m}) := \left\{ \mathbf{A} \in \mathbb{R}^{2m\times 2m} \;\middle|\; \mathbf{A}\mathbf{J}_0 = \mathbf{J}_0\mathbf{A} \right\} \subset \mathbb{R}^{2m\times 2m}. \tag{18.3}$$

In this way, the group $\mathrm{GL}(m, \mathbb{C})$ gets identified with the open subset of $\mathrm{End}_{\mathbb{C}}(\mathbb{R}^{2m})$ consisting of invertible transformations, making it a smooth submanifold and thus a Lie subgroup of $\mathrm{GL}(2m, \mathbb{R})$, with Lie algebra $\mathfrak{gl}(m, \mathbb{C}) = \mathrm{End}_{\mathbb{C}}(\mathbb{R}^{2m})$.

EXERCISE 18.11. Show that under the identification of $\mathrm{GL}(m, \mathbb{C})$ with a subgroup of $\mathrm{GL}(2m, \mathbb{R})$ explained in Example 18.10, $\mathrm{O}(2m) \cap \mathrm{GL}(m, \mathbb{C}) = \mathrm{U}(m) \subset \mathrm{GL}(2m, \mathbb{R})$.
Hint: Using the identification $\mathbb{C}^m = \mathbb{R}^{2m}$, write down a formula for the Hermitian inner product of $\mathbb{C}^m$ in terms of the Euclidean inner product of $\mathbb{R}^{2m}$ and the matrix $\mathbf{J}_0$ in (18.2).

[55] This example was not mentioned in the lecture but is provided here for your information. The same applies to one or two other things in Lecture 18 regarding the relationship between real and complex bundles.

18.2. The structure group of a vector bundle. Assume in the following that $G \subset \mathrm{GL}(m, \mathbb{F})$ is a Lie subgroup.

DEFINITION 18.12. A G**-structure** on a smooth vector bundle $E \to M$ of rank m is a maximal collection of smoothly compatible local trivializations $\{(\mathcal{U}_\alpha, \Phi_\alpha)\}_{\alpha \in I}$ of E with the property that $M = \bigcup_{\alpha \in I} \mathcal{U}_\alpha$ and the associated transition functions $g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{GL}(m, \mathbb{F})$ take their values in G. When a G-structure on $E \to M$ has been given, we call G the **structure group** of the bundle, and the local trivializations that belong to the G-structure will be called G**-compatible trivializations**.

For a given group G and bundle $E \to M$, a G-structure may or may not exist, and it will typically not be unique. *Every* vector bundle of rank m has structure group $\mathrm{GL}(m, \mathbb{F})$ by default, but for a given subgroup $G \subset \mathrm{GL}(m, \mathbb{F})$, it may or may not be possible to *reduce* the structure group to G by deleting a subcollection of its smooth local trivializations, so that those which remain are related to each other by G-valued transition functions. Note also that if G is a subgroup of some larger Lie subgroup $H \subset \mathrm{GL}(m, \mathbb{R})$, then a G-structure on $E \to M$ determines an H-structure, obtained by including all local trivializations that are related by H-valued transition functions to the G-compatible trivializations. A G-structure should be thought of as a *preferred* class of local trivializations that cover M, or equivalently, a preferred class of local frames, which we will also refer to in the following as G**-compatible frames**. Our first definition of *orientations* in §10.2 was somewhat analogous to this: choosing an orientation on a manifold M means selecting a preferred class of charts to be called "oriented" charts, and deleting those which are not compatible with them via orientation-preserving transition maps. A G-structure on a bundle E is also sometimes called a **reduction of the structure group** of E to G.

There is almost always a useful alternative way to interpret G-structures without mentioning transition functions, but the alternative interpretation varies depending on the specific group G. We will look next at several examples.

18.3. Global trivializations: $G = \{\mathbb{1}\}$. The trivial group $G := \{\mathbb{1}\} \subset \mathrm{GL}(m, \mathbb{F})$ is a 0-dimensional Lie subgroup of $\mathrm{GL}(m, \mathbb{F})$, and a G-structure on a bundle $E \to M$ then consists of a covering of M by a collection of local trivializations $\{(\mathcal{U}_\alpha, \Phi_\alpha)\}_{\alpha \in I}$ that are all *identical* wherever they overlap. If such a collection exists, then all of them are restrictions to the subsets $\mathcal{U}_\alpha \subset M$ of some global trivialization $\Phi : E \to M \times \mathbb{F}^m$, meaning a bundle isomorphism to the trivial m-plane bundle, so E is globally trivial. Conversely, any global trivialization $\Phi : E \to M \times \mathbb{F}^m$ determines a G-structure for $G := \{\mathbb{1}\}$ consisting of the restrictions of Φ to all possible open subsets $\mathcal{U}_\alpha \subset M$.

18.4. Orientations: $G = \mathrm{GL}_+(m, \mathbb{R})$. An **orientation** of a real vector bundle $E \to M$ is a choice of orientations for the fibers $\{E_p\}_{p \in M}$ that depend continuously on p, meaning that any collection of continuous sections $s_1, \dots, s_m : \mathcal{U} \to E$ on a neighborhood $\mathcal{U} \subset M$ of p that form a positively-oriented basis of E_p also form positively oriented bases of E_q for all q near p. Note that an orientation of a general vector bundle $E \to M$ need not have anything to do with an orientation of the base M, which may or may not be orientable—according to Proposition 10.25, an orientation of M is equivalent to an orientation of the specific bundle $TM \to M$.

An orientation of $E \to M$ determines a preferred class of local frames for E, namely those which are positively oriented at every point. Equivalently, the preferred class of local trivializations consists of those which define orientation-preserving isomorphisms between $\mathbb{R}^m$ and the fibers E_p. The transition functions that relate two such trivializations to each other must therefore take values in the group of orientation-preserving transformations of $\mathbb{R}^m$, that is, $\mathrm{GL}_+(m, \mathbb{R})$. An orientation of $E \to M$ thus determines a $\mathrm{GL}_+(m, \mathbb{R})$-structure. Conversely, any $\mathrm{GL}_+(m, \mathbb{R})$-structure on $E \to M$ determines an orientation of the fibers via the condition that an ordered basis of a fiber E_p is positively oriented if and only if some $\mathrm{GL}_+(m, \mathbb{R})$-compatible local trivialization identifies it with a positively-oriented basis of $\mathbb{R}^m$. If this holds for one of the preferred trivializations defined at p, then it holds for all the others as well, because the transition functions that relate them act by orientation-preserving transformations on $\mathbb{R}^m$.

We've proved:

PROPOSITION 18.13. *On a real vector bundle $E \to M$, there is a natural bijective correspondence between orientations and $\mathrm{GL}_+(m,\mathbb{R})$-structures.* □

Though you might already find it obvious that every trivial real vector bundle is orientable, we can now give a quick new proof of this fact in the language of structure groups: if $E \to M$ is trivial, then it admits a G-structure for $G := \{\mathbb{1}\}$, which is a subgroup of $\mathrm{GL}_+(m,\mathbb{R})$, so $E \to M$ therefore also admits a $\mathrm{GL}_+(m,\mathbb{R})$-structure, meaning an orientation. In fact, this argument shows that any global trivialization of a vector bundle determines a G-structure for every Lie subgroup $G \subset \mathrm{GL}(m,\mathbb{F})$.

EXERCISE 18.14. Prove that the line bundle $\ell \to S^1$ in Example 16.23 is not orientable.

18.5. Bundle metrics: $G = \mathrm{O}(m)$, $\mathrm{U}(m)$, $\mathrm{O}(k,\ell)$. The following definition generalizes the notion of a Riemannian metric in two respects: it is defined on an arbitrary vector bundle instead of a tangent bundle $TM \to M$, and it requires the weaker condition of *nondegeneracy* in place of positive-definiteness.

DEFINITION 18.15. A **bundle metric** on a real vector bundle $E \to M$ is a smooth function

$$\langle\ ,\ \rangle : E \oplus E \to \mathbb{R}$$

whose restriction to the fiber $E_p \times E_p$ for each $p \in M$ is all of the following:

(i) (bilinear) $E_p \times E_p \to \mathbb{R} : (v,w) \mapsto \langle v,w\rangle$ is a bilinear map
(ii) (symmetric) $\langle v,w\rangle = \langle w,v\rangle$ for all $v,w \in E_p$
(iii) (nondegenerate) The map $E_p \to E_p^* : v \mapsto \langle v,\cdot\rangle$ is injective.

We will say additionally that $\langle\ ,\ \rangle$ is **positive** if the third condition is strengthened to:

(iii) (positive) $\langle v,v\rangle > 0$ for all nonzero $v \in E_p$.

For a complex vector bundle $E \to M$, we modify the above definition as follows: $\langle\ ,\ \rangle$ is a smooth function

$$\langle\ ,\ \rangle : E \oplus E \to \mathbb{C}$$

whose restriction to $E_p \times E_p$ is:

(i) (sesquilinear) $E_p \times E_p \to \mathbb{C} : (v,w) \mapsto \langle v,w\rangle$ is linear in the second factor and antilinear in the first
(ii) (Hermitian) $\langle v,w\rangle = \overline{\langle w,v\rangle}$ for all $v,w \in E_p$
(iii) (nondegenerate) Same as in the real case.

The positivity condition in the complex case is also the same as in the real case.

REMARK 18.16. A bundle metric is called **indefinite** if it is nondegenerate but not positive. In most of the literature, bundle metrics are assumed to be positive by default, and it is generally wise to assume this unless the word "indefinite" is included. A large portion of what we have to say about them will however be valid without assuming positivity, so in these notes, we will use "bundle metric" as a general term that includes the indefinite case. Bundle metrics are also often referred to as "Euclidean" bundle metrics in the real case and "Hermitian" bundle metrics in the complex case.

Example 18.17. A Riemannian metric on a manifold M is a positive bundle metric on its tangent bundle $TM \to M$. If g is instead an indefinite bundle metric on $TM \to M$, it is called a **pseudo-Riemannian** (or also **semi-Riemannian**) metric on M, and the pair (M, g) is called a **pseudo-Riemannian manifold**.

In the real case, a bundle metric on $E \to M$ can also be regarded as a smooth section of the vector bundle $E^* \otimes E^* \to M$, whose fiber at $p \in M$ is the space of bilinear maps $E_p \times E_p \to \mathbb{R}$. Sesquilinearity modifies this statement in the complex case and replaces $E^* \otimes E^*$ with $\bar{E}^* \otimes E^*$, whose fiber at $p \in M$ is (according to Exercise 17.21) naturally isomorphic to the space of maps $E_p \times E_p \to \mathbb{C}$ that are antilinear in the first and linear in the second factor.

A positive bundle metric assigns to each fiber what is conventionally called an inner product, and as we observed in §15.2, the set of positive-definite inner products on any vector space is convex. Our previous existence result for Riemannian metrics therefore generalizes in a straightforward way:

Theorem 18.18. *Every vector bundle admits a positive bundle metric.*

Proof. Trivial bundles obviously admit positive bundle metrics since one can simply choose the standard Euclidean inner product of $\mathbb{R}^m$ or (for the case $\mathbb{F} = \mathbb{C}$) the standard Hermitian inner product of $\mathbb{C}^m$ on every fiber. It follows that on any vector bundle $E \to M$ with a collection of local trivializations $\{\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^m\}_{\alpha \in I}$ covering M, one can choose bundle metrics on each $E|_{\mathcal{U}_\alpha}$, and then piece these together using a partition of unity on M subordinate to the cover $\{\mathcal{U}_\alpha\}_{\alpha \in I}$. $\square$

It is interesting to note that if we'd been allowed to assume in Theorem 18.18 that the bundle $E \to M$ has a G-structure for $G = \mathrm{O}(m)$ or (in the case $\mathbb{F} = \mathbb{C}$) $G = \mathrm{U}(m)$, then the proof would not have required a partition of unity. Indeed, if one defines $\langle\ ,\ \rangle$ over regions $\mathcal{U}_\alpha \subset M$ so that it matches the standard inner product of $\mathbb{F}^m$ in some choice of *G-compatible* local trivialization over $\mathcal{U}_\alpha$, then this definition is independent of that choice: having transition functions valued in $G \in \{\mathrm{O}(m), \mathrm{U}(m)\}$ means that they preserve the standard inner product on $\mathbb{F}^m$, so any other G-compatible local trivialization on an overlapping region produces *the same* inner product on the fibers. This means that if an $\mathrm{O}(m)$- or $\mathrm{U}(m)$-structure is given, then it determines a unique positive bundle metric on $E \to M$ that looks like the standard inner product of $\mathbb{F}^m$ in any compatible local trivialization. There is also a converse to this: if a positive bundle metric $\langle\ ,\ \rangle$ is given, then every smooth local frame on a region $\mathcal{U}_\alpha \subset M$ can be modified via the Gram-Schmidt algorithm to produce one that is orthonormal at every point $p \in \mathcal{U}_\alpha$, so that the corresponding local trivialization identifies $\langle\ ,\ \rangle$ with the standard inner product of $\mathbb{F}^m$. Any two trivializations produced in this way will then be related by a transition function whose values preserve this inner product, meaning they are in $\mathrm{O}(m)$ or $\mathrm{U}(m)$. We've proved:

Proposition 18.19. *On a vector bundle $E \to M$ of rank m, there is a natural bijective correspondence between positive bundle metrics and $\mathrm{O}(m)$-structures if E is real, or $\mathrm{U}(m)$-structures if E is complex.* $\square$

Remark 18.20. In light of Proposition 18.19, an $\mathrm{O}(m)$-structure on a vector bundle is also sometimes called a **Euclidean structure**, and a $\mathrm{U}(m)$-structure is called a **Hermitian structure**.

Extending this discussion to the case of an indefinite bundle metric $\langle\ ,\ \rangle$ on $E \to M$ requires a suitable generalization of the notion of orthonormal frames. In the following, we confine our attention to *real* vector bundles since that is the case that arises most often in applications, but there are no substantial differences in the complex case. We will say that a local frame $e_1, \ldots, e_m$ for E on some region $\mathcal{U} \subset M$ is **orthonormal** with respect to $\langle\ ,\ \rangle$ if for some $k \in \{0, \ldots, m\}$, it satisfies

$$\begin{aligned} &\langle e_j, e_j\rangle = 1 \quad \text{for } j = 1, \dots, k \\ &\langle e_j, e_j\rangle = -1 \quad \text{for } j = k+1, \dots, m \\ &\langle e_i, e_j\rangle = 0 \quad \text{for } i \neq j. \end{aligned} \tag{18.4}$$

The integers k and $\ell := m - k$ are determined by the bundle metric, and can be characterized as the dimensions of the largest subspace of any fiber on which $\langle\ ,\ \rangle$ is positive-definite or negative-definite respectively. In general, these numbers need not be the same everywhere on M, though it should be clear that they are *locally* constant and thus constant on each connected component. As a rule, the only interesting examples are those in which k and ℓ are constant everywhere; in this case, the pair (k, ℓ) is called the **signature** of the bundle metric $\langle\ ,\ \rangle$. The spacetime manifolds of general relativity are 4-manifolds with pseudo-Riemannian metrics of signature $(1, 3)$; these are known as **Lorentzian manifolds**.

Lemma 18.21. *For any real vector bundle $E \to M$ with an indefinite bundle metric $\langle\ ,\ \rangle$, every point $p \in M$ has a neighborhood $\mathcal{U} \subset M$ on which E admits an orthonormal frame.*

Since Lemma 18.21 is a local statement and all vector bundles are locally trivial, it suffices to prove it for the special case of a trivial m-plane bundle

$$E := \mathcal{U} \times \mathbb{R}^m$$

over some open subset $\mathcal{U} \subset M$ of a manifold. The restriction of $\langle\ ,\ \rangle$ to the fiber over a point $p \in \mathcal{U}$ is in this case a bilinear form on $E_p = \mathbb{R}^m$ that can be written as

$$\langle \mathbf{v}, \mathbf{w}\rangle_p = \langle \mathbf{v}, \mathbf{H}(p)\mathbf{w}\rangle_{\mathbb{R}^m} \qquad \text{for } \mathbf{v}, \mathbf{w} \in \mathbb{R}^m,$$

where $\langle\ ,\ \rangle_{\mathbb{R}^m}$ denotes the standard Euclidean inner product on $\mathbb{R}^m$ and $\mathbf{H}(p) \in \mathbb{R}^{m\times m}$ is a uniquely determined matrix that depends smoothly on $p \in \mathcal{U}$. Symmetry and nondegeneracy imply moreover that $\mathbf{H}(p)$ is always both symmetric and invertible respectively. It follows then from the spectral theorem that at every point $p \in \mathcal{U}$, $\mathbb{R}^m$ splits uniquely

$$\mathbb{R}^m = E_p^+ \oplus E_p^-$$

into the subspaces $E_p^+, E_p^- \subset \mathbb{R}^m$ spanned by the positive and negative eigenvalues respectively of $\mathbf{H}(p)$; on these two subspaces, $\langle\ ,\ \rangle_p$ is positive- or negative-definite respectively. Notice that E_p^+ and E_p^- are orthogonal to each other with respect to both the Euclidean inner product and the given bundle metric $\langle\ ,\ \rangle$. We will see below that these subspaces vary smoothly with p, but since that fact is not so obvious, let us first give a proof of Lemma 18.21 that does not require it.

First proof of Lemma 18.21. For a given point $p \in \mathcal{U}$, let $k := \dim E_p^+$ and $\ell := \dim E_p^-$, choose orthonormal bases of E_p^+ and E_p^- and choose a smooth frame $\hat{e}_1, \dots, \hat{e}_m$ for E on a neighborhood of p such that at the point p itself, $\hat{e}_1, \dots, \hat{e}_k$ matches the chosen orthonormal basis of E_p^+ while $\hat{e}_{k+1}, \dots, \hat{e}_m$ matches the chosen orthonormal basis of E_p^-. Since E_p^+ and E_p^- are orthogonal with respect to $\langle\ ,\ \rangle$, it follows that $\hat{e}_1, \dots, \hat{e}_m$ satisfy the orthonormality condition (18.4) at p, and we will now use a minor variation on the Gram-Schmidt algorithm to produce from this an orthonormal frame $e_1, \dots, e_m$ that is defined on a neighborhood of p and matches $\hat{e}_1, \dots, \hat{e}_m$ at p. The key observation making this possible is that since $\langle\ ,\ \rangle$ is positive on E_p^+ and negative on E_p^-, it is also positive / negative on the subbundles spanned by $\hat{e}_1, \dots, \hat{e}_k$ and $\hat{e}_{k+1}, \dots, \hat{e}_m$ respectively. Now, define $e_1, \dots, e_k$ simply by applying the usual Gram-Schmidt procedure to $\hat{e}_1, \dots, \hat{e}_k$. Since $\langle \hat{e}_{k+1}, \hat{e}_{k+1}\rangle < 0$, the correct definition of e_{k+1} is slightly different: we set

$$e_{k+1} := f_1 \cdot \left(\hat{e}_{k+1} - \sum_{j=1}^{k} \langle \hat{e}_{k+1}, e_j\rangle e_j \right),$$

with a positive function f_1 chosen to ensure that $\langle \hat{e}_{k+1}, \hat{e}_{k+1} \rangle \equiv -1$ on a neighborhood of p, which is possible because the expression in parentheses matches $\hat{e}_{k+1}$ at p, so that its product with itself is negative. Continuing in this way inductively, the new section e_{k+i} is defined out of $e_1, \dots, e_{k+i-1}$ for each $i = 1, \dots, \ell$ by

$$e_{k+i} := f_i \cdot \left(\hat{e}_{k+i} - \sum_{j=1}^{k} \langle \hat{e}_{k+i}, e_j \rangle e_j + \sum_{j=1}^{i-1} \langle \hat{e}_{k+i}, e_{k+j} \rangle e_{k+j} \right),$$

with the positive function f_i again chosen to achieve the normalization $\langle e_{k+i}, e_{k+i} \rangle \equiv -1$. □

I mentioned above that the subspaces $E_p^\pm \subset \mathbb{R}^m$ vary smoothly with p, which will give rise to a slightly simpler proof of Lemma 18.21. These subspaces are defined as direct sums of certain eigenspaces of the matrix $\mathbf{H}(p)$, but we have to be a bit careful here, because in general, individual eigenspaces cannot be assumed to depend smoothly on the matrix—one can show that they do whenever the corresponding eigenvalue is simple, but in our situation, eigenvalues with multiplicity may occur and there is no general way to avoid them. What we are interested in however is not an individual eigenspace, but direct sums of several eigenspaces corresponding to eigenvalues in fixed open subsets of $\mathbb{R}$, namely $(-\infty, 0)$ and $(0, \infty)$. In this situation, Cauchy's integration theory from complex analysis provides a useful trick:

LEMMA 18.22. *Suppose $\mathbf{A} \in \mathbb{C}^{m \times m}$ is a diagonalizable matrix,*

$$\sigma(\mathbf{A}) = \sigma_0 \sqcup \sigma_1 \subset \mathbb{C}$$

is a decomposition of its spectrum $\sigma(\mathbf{A})$ into two disjoint subsets, and write

$$\mathbb{C}^m = V_0 \oplus V_1, \qquad \textit{where} \qquad V_j := \bigoplus_{\lambda \in \sigma_j} \ker(\lambda \mathbb{1} - \mathbf{A}), \quad j = 0, 1$$

for the corresponding splitting of $\mathbb{C}^m$ into direct sums of eigenspaces. Then for any smoothly embedded oriented circle $\gamma \subset \mathbb{C}$ that does not intersect $\sigma(\mathbf{A})$ and has winding number j around each eigenvalue in σ_j for $j = 0, 1$, the matrix-valued Cauchy integral

$$\mathbf{P} := \frac{1}{2\pi i} \int_\gamma (z\mathbb{1} - \mathbf{A})^{-1} \, dz \in \mathbb{C}^{m \times m}$$

defines the linear projection to V_1 along V_0.

PROOF. The function $\mathbb{C} \backslash \sigma(\mathbf{A}) \to \mathbb{C}^{m \times m} : z \mapsto (z\mathbb{1} - \mathbf{A})^{-1}$ is holomorphic since $z\mathbb{1} - \mathbf{A}$ is an affine function of z and, for arbitrary invertible matrices $\mathbf{B} \in \mathrm{GL}(m, \mathbb{C})$, the entries of $\mathbf{B}^{-1}$ are rational functions of the entries in $\mathbf{B}$. Cauchy's theorem thus implies that the integral will not change if γ is replaced by a disjoint union of small circles around the specific eigenvalues in σ_1, and it suffices therefore to consider the case where σ_1 consists of only one eigenvalue $\lambda_1 \in \mathbb{C}$ and γ is parametrized by the boundary of the ϵ-disk around λ_1 for $\epsilon > 0$ small. Since $\mathbf{A}$ is diagonalizable we can also assume after a change of basis on $\mathbb{C}^m$ that $\mathbf{A}$ is diagonal; let us write its diagonal entries as $\Lambda_1, \dots, \Lambda_m \in \mathbb{C}$, keeping in mind that these are all elements of $\sigma(\mathbf{A})$ and some of them may be repeated. The values of the function $(z\mathbb{1} - \mathbf{A})^{-1}$ are then also diagonal matricies, whose diagonal entries are the complex-valued functions $\frac{1}{z - \Lambda_j}$ for $j = 1 \dots, m$. For any j such that $\Lambda_j \neq \lambda_1$, we can assume $\frac{1}{z - \Lambda_j}$ is a holomorphic function on the disk enclosed by γ, so its integral is 0. On the other hand, whenever $\Lambda_j = \lambda_1$, integration makes the corresponding diagonal element into

$$\frac{1}{2\pi i} \int_\gamma \frac{dz}{z - \lambda_1} = 1.$$

We conclude that in our chosen basis of eigenvectors for $\mathbf{A}$, $\mathbf{P}$ is a diagonal matrix whose entries are all 1 or 0, with 1 appearing only in the places where the corresponding entry of $\mathbf{A}$ is λ_1. In other words, $\mathbf{P}$ acts as the identity on the eigenspace of λ_1 and as 0 on all the other eigenspaces. $\square$

COROLLARY 18.23. *The subspaces $E_p^+, E_p^- \subset \mathbb{R}^m$ defined as the direct sums of the positive and negative eigenspaces respectively of $\mathbf{H}(p)$ vary smoothly with the point $p \in \mathcal{U}$.*

PROOF. Given $p \in \mathcal{U}$, choose an embedded oriented circle $\gamma \subset \mathbb{C}$ that surrounds the positive eigenvalues of $\mathbf{H}(p)$ but stays in the right half-plane, so its winding around every negative eigenvalue is 0. Then according to Lemma 18.22, the matrix $\frac{1}{2\pi i}\int_\gamma (z\mathbb{1} - \mathbf{H}(p))^{-1}\,dz$ defines the orthogonal projection to E_p^+ along E_p^-, and this remains true if p is moved within a small enough region so that the eigenvalues of $\mathbf{H}(p)$ never touch γ. This matrix-valued integral clearly depends smoothly on p, and therefore so does the complementary projection to E_p^-. $\square$

SECOND PROOF OF LEMMA 18.21. Corollary 18.23 implies that the subspaces $E_p^+, E_p^- \subset \mathbb{R}^m$ form the fibers of smooth subbundles $E^\pm \subset E$, giving a splitting

$$\mathcal{U} \times \mathbb{R}^m = E = E^+ \oplus E^- \tag{18.5}$$

such that $\pm\langle\ ,\ \rangle$ restricts to a positive bundle metric on $E^\pm$, and moreover, the fibers of E^+ and E^- are mutually orthogonal with respect to $\langle\ ,\ \rangle$. An orthonormal frame for E is then constructed by combining orthonormal frames of E^+ and E^-, and this can be done on a sufficiently small neighborhood of any given point. $\square$

For a bundle metric of signature (k, ℓ), the local trivialization corresponding to an orthonormal frame identifies $\langle\ ,\ \rangle$ on each fiber with the "standard" indefinite inner product

$$\langle \mathbf{v}, \mathbf{w} \rangle_{k,\ell} = \sum_{j=1}^{k} v^j w^j - \sum_{j=k+1}^{m} v^j w^j,$$

thus any two local trivializations constructed in this way are related by a transition function with values in the group $\mathrm{O}(k, \ell)$ from Example 18.8. We summarize:

PROPOSITION 18.24. *On a real vector bundle $E \to M$ of rank m with integers $k, \ell \geqslant 0$ satisfying $k + \ell = m$, there is a natural bijective correspondence between bundle metrics of signature (k, ℓ) and $\mathrm{O}(k, \ell)$-structures.* $\square$

EXERCISE 18.25. Show that for any real vector bundle $E \to M$ with a bundle metric $\langle\ ,\ \rangle$ of signature (k, ℓ) there exist smooth subbundles $E^+ \subset E$ and $E^- \subset E$ of ranks k and ℓ respectively and a bundle of isomorphism $E \cong E^+ \oplus E^-$.
Caution: This does not follow immediately from the splitting in (18.5)*, because that splitting was defined specifically for a trivial bundle; it can always be done locally since all bundles are locally trivial, but the result will depend on the choice of local trivialization. Obtaining such a splitting globally will require another choice, but it is a choice that can always be made.*

REMARK 18.26. When $k, \ell \geqslant 1$, the existence of the splitting $E = E^+ \oplus E^-$ in Exercise 18.25 is a nontrivial condition that is not satisfied for all bundles, thus unlike the positive case, bundle metrics of arbitrary signature do not always exist. We will later see for instance that S^2 does not admit any pseudo-Riemannian metric of signature $(1, 1)$.

18.6. Volume forms: $G = \mathrm{SL}(m, \mathbb{F})$. A **volume form** on a vector bundle $E \to M$ of rank m is a section $\mu \in \Gamma(\Lambda^m E^*)$ that satisfies $\mu(p) \neq 0$ for all $p \in M$. In other words, for every $p \in M$, $\mu(p)$ is an alternating m-fold multilinear form $E_p \times \ldots \times E_p \to \mathbb{F}$ that evaluates to something nonzero on some (and therefore any) basis $v_1, \ldots, v_m \in E_p$. The terminology has a geometric motivation in the case $\mathbb{F} = \mathbb{R}$, as one can then use μ to define the notion of volume in every fiber by saying that $|\mu(p)(v_1, \ldots, v_m)|$ is the volume of the parallelepiped in E_p spanned by $v_1, \ldots, v_m$. No such geometric interpretation is available in the complex case, but the definition makes sense algebraically.

Given a volume form $\mu \in \Gamma(\Lambda^m E^*)$ and a local frame $e_1, \ldots, e_m$ for E over an open set $\mathcal{U}_\alpha \subset M$, one can always modify e_1 by multiplication with a scalar-valued function to arrange

$$\mu(e_1, \ldots, e_m) \equiv 1 \quad \text{on } \mathcal{U}_\alpha. \tag{18.6}$$

The corresponding local trivialization then identifies μ over $\mathcal{U}_\alpha$ with the "standard" volume form on $\mathbb{F}^m$, given by

$$\mu_{\mathrm{std}}(\mathbf{v}_1, \ldots, \mathbf{v}_m) := \det \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_m \end{pmatrix} \in \mathbb{F}$$

for $\mathbf{v}_1, \ldots, \mathbf{v}_m \in \mathbb{F}^m$. The group of linear transformations $\mathbb{F}^m \to \mathbb{F}^m$ that preserve μ_{std} is the special linear group, $\mathrm{SL}(m, \mathbb{F})$, thus covering M with local frames that satisfy (18.6) determines an $\mathrm{SL}(m, \mathbb{F})$-structure on E. Conversely, if an $\mathrm{SL}(m, \mathbb{F})$-structure is given, then there is a unique volume form $\mu \in \Gamma(\Lambda^m E^*)$ that looks like μ_{std} in every $\mathrm{SL}(m, \mathbb{F})$-compatible local trivialization, proving:

PROPOSITION 18.27. *On any vector bundle $E \to M$ of rank m over the field $\mathbb{F}$, there is a natural bijective correspondence between volume forms and $\mathrm{SL}(m, \mathbb{F})$-structures.* □

Several fundamental facts about volume forms on a manifold can now be generalized and proved as easy corollaries of basic observations about specific subgroups of $\mathrm{GL}(m, \mathbb{R})$:

PROPOSITION 18.28. *Every real vector bundle admitting a volume form is orientable.*

PROOF. If $E \to M$ has an $\mathrm{SL}(m, \mathbb{R})$-structure, then this determines a $\mathrm{GL}_+(m, \mathbb{R})$-structure since $\mathrm{SL}(m, \mathbb{R}) \subset \mathrm{GL}_+(m, \mathbb{R})$. □

PROPOSITION 18.29. *On any oriented real vector bundle $E \to M$, any bundle metric determines a unique volume form μ such that $\mu(v_1, \ldots, v_m) = 1$ for every positively-oriented orthonormal basis of a fiber.*

PROOF. As an oriented bundle, E has structure group $\mathrm{GL}_+(m, \mathbb{R})$, and introducing a bundle metric of signature (k, ℓ) reduces its structure group further to

$$\mathrm{SO}(k, \ell) := \mathrm{O}(k, \ell) \cap \mathrm{SL}(m, \mathbb{R}) = \mathrm{O}(k, \ell) \cap \mathrm{GL}_+(m, \mathbb{R}),$$

where the equality of these two intersections results from the fact that every $\mathbf{A} \in \mathrm{O}(k, \ell)$ has determinant ± 1. Since $\mathrm{SO}(k, \ell) \subset \mathrm{SL}(m, \mathbb{R})$, we therefore also have an $\mathrm{SL}(m, \mathbb{R})$-structure and thus a volume form, which evaluates to 1 on the standard basis whenever it is viewed in an $\mathrm{SO}(k, \ell)$-compatible trivialization; in particular, this means bases that are positively oriented and orthonormal. □

EXERCISE 18.30. Suppose $E \to M$ is an oriented real vector bundle of rank m with a bundle metric $\langle \ , \ \rangle$.

(a) Reprove Proposition 18.29 by an argument analogous to Corollary 11.10 on the Riemannian volume form $d\mathrm{vol} \in \Gamma(\Lambda^n T^* M)$ for an oriented Riemannian manifold (M, g), i.e. show that the volume form $\mu \in \Gamma(\Lambda^m E^*)$ determined by the orientation and bundle metric on E can be written locally in the form $e^1_* \wedge \ldots \wedge e^m_*$ using the dual frame to any positively-oriented orthonormal local frame $e_1, \ldots, e_m$.

(b) Generalize the local coordinate formula for the Riemannian volume form in Exercise 11.12 as follows. Assume $e_1, \dots, e_m$ is a positively-oriented but not necessarily orthonormal local frame over some open set $\mathcal{U} \subset M$, and write $g_{ij} := \langle e_i, e_j \rangle : \mathcal{U} \to \mathbb{R}$ for the resulting component functions of the bundle metric. Show
$$\mu = \sqrt{\pm \det \mathbf{g}}\, e_*^1 \wedge \dots \wedge e_*^m \qquad \text{on } \mathcal{U},$$
where $e_*^1, \dots, e_*^m$ is the dual frame to $e_1, \dots, e_m$, $\mathbf{g} : \mathcal{U} \to \mathbb{R}^{m \times m}$ is the matrix-valued function whose entries are g_{ij}, and the sign $\pm$ is chosen to make the expression under the square root positive (this will depend on the signature of the bundle metric).

18.7. Complex structures: $G = \mathrm{GL}(m, \mathbb{C}) \subset \mathrm{GL}(2m, \mathbb{R})$. [56]

Along the lines of Example 18.10, identifying $\mathbb{C}^m$ with $\mathbb{R}^{2m}$ makes any complex vector bundle $E \to M$ of rank m into a real vector bundle of rank $2m$ that is endowed with a G-structure for $G \cong \mathrm{GL}(m, \mathbb{C})$ defined as the subgroup of $\mathrm{GL}(2m, \mathbb{R})$ consisting of all invertible linear transformations $\mathbb{R}^{2m} \to \mathbb{R}^{2m}$ that commute with the matrix
$$\mathbf{J}_0 := \begin{pmatrix} 0 & -\mathbb{1}_m \\ \mathbb{1}_m & 0 \end{pmatrix}.$$
Recall from §7.1.4 that on any even-dimensional vector space V, a linear map $J : V \to V$ satisfying $J^2 = -\mathbb{1}$ is called a **complex structure**, thus $\mathbf{J}_0$ is an example of a complex structure on $\mathbb{R}^{2m}$. Any complex structure $J : V \to V$ makes V into a complex vector space by defining complex scalar multiplication to mean
$$(a + ib)v := av + bJv, \qquad a, b \in \mathbb{R}, \;\; v \in V.$$
If $v_1, \dots, v_m \in V$ is any complex basis of this vector space, then $v_1, \dots, v_m, Jv_1, \dots, Jv_m$ is a real basis in which the matrix representing the transformation J is $\mathbf{J}_0$; this proves in particular that *every* complex structure on $\mathbb{R}^{2m}$ is equivalent to $\mathbf{J}_0$ via a change of basis.

More generally, a **complex structure** on a real vector bundle $E \to M$ of rank $2m$ is a smooth section J of the bundle
$$\mathrm{End}(E) := \mathrm{Hom}(E, E)$$
such that $J(p) : E_p \to E_p$ is a complex structure on E_p for every $p \in M$. Choosing a complex structure on E makes every fiber into a complex vector space of dimension m, and on a sufficiently small neighborhood $\mathcal{U} \subset M$ of any point p, one can choose a complex basis $v_1, \dots, v_m$ of E_p and find a tuple of smooth sections $e_1, \dots, e_m : \mathcal{U} \to E$ such that $e_j(p) = v_j$ for every $j = 1, \dots, m$; after shrinking the neighborhood $\mathcal{U}$, we can then assume without loss of generality that the vectors $e_1, \dots, e_m$ remain complex-linearly independent and thus form a basis of every fiber over points in $\mathcal{U}$. It follows that $e_1, \dots, e_m, Je_1, \dots, Je_m$ then forms a smooth frame for E over $\mathcal{U}$, and it defines a local trivialization $E|_{\mathcal{U}} \to \mathcal{U} \times \mathbb{R}^{2m}$ that identifies J on each fiber over points in $\mathcal{U}$ with the standard complex structure $\mathbf{J}_0 : \mathbb{R}^{2m} \to \mathbb{R}^{2m}$. The transition functions relating any two local trivializations constructed in this way must then take values in the subgroup $\mathrm{GL}(m, \mathbb{C}) \subset \mathrm{GL}(2m, \mathbb{R})$, so we have constructed a $\mathrm{GL}(m, \mathbb{C})$-structure on E, and if we replace $\mathbb{R}^{2m}$ by $\mathbb{C}^m$, $E \to M$ can now be understood as a complex vector bundle of rank m. Conversely, any $\mathrm{GL}(m, \mathbb{C})$-structure on a real bundle $E \to M$ of rank $2m$ determines a complex structure $J \in \Gamma(\mathrm{End}(E))$ that is identified with $\mathbf{J}_0 : \mathbb{R}^{2m} \to \mathbb{R}^{2m}$ by any $\mathrm{GL}(m, \mathbb{C})$-compatible local trivialization. This proves:

PROPOSITION 18.31. *There is a natural bijective correspondence between complex structures $J \in \Gamma(\mathrm{End}(E))$ on a real vector bundle $E \to M$ of rank $2m$ and $\mathrm{GL}(m, \mathbb{C})$-structures on E, where $\mathrm{GL}(m, \mathbb{C})$ is identified with a subgroup of $\mathrm{GL}(2m, \mathbb{R})$ as in Example 18.10. Moreover, any smooth real vector bundle E of rank $2m$ with complex structure J can be regarded naturally as a smooth*

[56] Like Example 18.10, this section was not covered in the lecture but is provided here for your information.

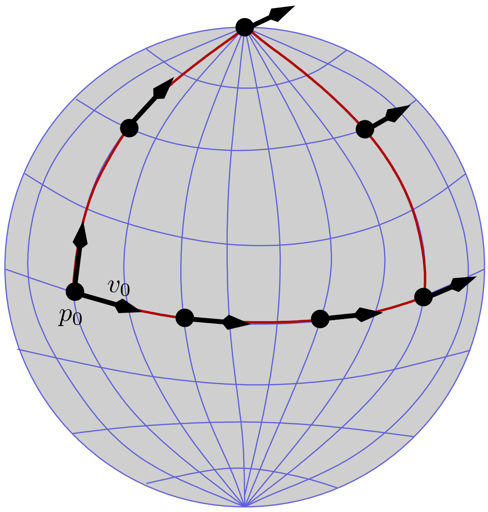

FIGURE 8. Parallel transport of a tangent vector around a closed path in S^2.

complex vector bundle of rank m whose fibers over points $p \in M$ are the vector spaces E_p with complex scalar multiplication defined by $(a + ib)v := av + bJ(p)v$. □

19. Connections on vector bundles

By way of motivation for what we will do in the next few lectures, I'd like to take a second look at a thought-experiment that was mentioned in Lecture 1. Figure 8 shows a closed path on S^2 that is made up of three smooth paths intersecting at right angles: one moving along the equator, and two that connect the equator to the north pole via longitudes. In this scenario, we pick a starting point p_0 for this path and a tangent vector $v_0 \in T_{p_0}S^2$, and then ask: if we move v_0 in a "parallel" manner along the path, keeping it tangent to the sphere as we go, will it come back to the same starting vector when the path returns to p_0? The question is imprecisely stated, because I have not said what "parallel" in this situation should mean, and that is a detail we will need to discuss. Nonetheless, the scenario in Figure 8 looks as if v_0 is being moved along the path in the most natural way possible, and the answer is clearly no: the vector it comes back to at the end of the closed loop is different.

This is not something that happens in *Euclidean* geometry. If our manifold were $\mathbb{R}^2$ instead of S^2, then there would be an obvious way to define what moving a tangent vector in a "parallel" manner along a path should mean: it means that the vector is constant, and it will therefore always return to itself when the path comes back to its starting point. On S^2, on the other hand, there is no obvious way to define what it should mean for a vector field to be *constant*, due to the fact that the tangent spaces T_pS^2 themselves are not constant as the point p moves. We will see nonetheless that if we endow the tangent spaces T_pS^2 with the Euclidean inner product and thus regard S^2 as a Riemannian manifold, then there is a natural way to define what it means for a vector field along a path to be *parallel*—that is what we will call the natural generalization of the word "constant" in this context—but this notion will have some counterintuitive properties, e.g. that no vector field can ever be parallel on an entire open subset, no matter how small. Such properties are symptoms of the fact that S^2 has nontrivial *curvature*, while $\mathbb{R}^2$ with its Euclidean inner product does not.

In order to clarify what this means, we will first consider a general vector bundle $E \to M$ and ask what it might mean to say that a section $s \in \Gamma(E)$ is "constant" along a path. Such a notion can be defined, but the definition is not canonical: it depends on an extra piece of geometric data that must be chosen, and that data is called a *connection*. Several distinct definitions of the term "connection" can be found in various textbooks, and all of them are equivalent but look cosmetically quite different. Our first task is thus to understand why these particular definitions are the ones we need, and why they are equivalent.

19.1. Parallel transport and horizontal lifts. We assume for the rest of this lecture that

$$\pi : E \to M$$

is a smooth real or complex ($\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$) vector bundle of rank m over an n-manifold M. For a section $s \in \Gamma(E)$, we have defined what it means for s to be *differentiable*, but we have not yet talked about actually differentiating it. If one wants to define, say, the derivative of s at a point $p \in M$ in the direction $X \in T_pM$, one quickly encounters a problem that we have seen before when talking about vector and tensor fields: choosing a path $\gamma : (-\epsilon, \epsilon) \to M$ with $\gamma(0) = p$ and $\dot{\gamma}(0) = X$, one cannot simply define[57]

$$ds(X) \overset{?}{:=} \left.\frac{d}{dt}s(\gamma(t))\right|_{t=0} = \lim_{t\to 0}\frac{s(\gamma(t)) - s(p)}{t} \tag{19.1}$$

since $s(\gamma(t))$ and $s(p)$ belong to different vector spaces $E_{\gamma(t)} \neq E_p$. Before one can make sense of such an expression, one needs a way of identifying these vector spaces so that $s(\gamma(t)) - s(p)$ can be defined, e.g. one needs a smooth family of vector space isomorphisms

$$P^t_\gamma : E_{\gamma(0)} \xrightarrow{\cong} E_{\gamma(t)}, \qquad \text{such that} \qquad P^0_\gamma = \mathbb{1}. \tag{19.2}$$

Under suitable conditions to be clarified below, we will refer to families of isomorphisms of this form as **parallel transport** (*Parallelverschiebung*) (or also **parallel translation**) maps along the path γ. If such a family is given, then one can use it to turn (19.1) into a sensible definition, namely

$$\nabla_X s := \nabla_t s(\gamma(t))|_{t=0} := \left.\frac{d}{dt}(P^t_\gamma)^{-1}(s(\gamma(t)))\right|_{t=0} = \lim_{t\to 0}\frac{(P^t_\gamma)^{-1}(s(\gamma(t))) - s(p)}{t} \in E_p. \tag{19.3}$$

This is called the **covariant derivative** (*kovariante Ableitung*) of s at p **in the direction** X, and also the **covariant derivative** of s **along the path** γ at $t = 0$.

Once parallel transport and covariant derivatives have been defined, one can also say what it means for s to be "constant" along the path γ: it means simply that

$$s(\gamma(t)) = P^t_\gamma(s(p))$$

for all t, or in terms of the covariant derivative, $\nabla_t s(\gamma(t)) \equiv 0$. Since "constant" is not really an appropriate term when the vector spaces $E_{\gamma(t)}$ vary with t, a section with this property is said to be **parallel** (or also **covariantly constant**) along the path γ. This notion clearly depends on the parallel transport isomorphisms P^t_γ, i.e. if one chose these isomorphisms differently, then a section that is parallel for one choice might not be parallel for another.

So, how does one actually go about defining parallel transport isomorphisms as in (19.2)? In the special case $E = TM$, we found one conceivable answer to this question in §6.3: one can assume that γ is a flow line of a vector field X and obtain a family of isomorphisms from the linearized flow,

$$T_p\varphi^t_X : T_pM \xrightarrow{\cong} T_{\gamma(t)}M.$$

[57]The question mark over the equal sign in (19.1) is meant to convey a sense of confusion—because the definition does not really make sense.

This approach gave us the definition of the Lie derivative $\mathcal{L}_X Y$ of a vector field $Y \in \Gamma(TM) = \mathfrak{X}(M)$. The first obvious problem is that this approach only makes sense on tangent bundles, though one can perhaps imagine generalizing it to the various tensor bundles that are defined in terms of tangent bundles, leading to the Lie derivatives of tensor fields that were defined in Lecture 8. But there is a more basic problem here: a vector field $X \in \mathfrak{X}(M)$ does not define isomorphisms as in (19.2) along *arbitrary* paths γ, it only defines them along flow lines, and the derivative $\mathcal{L}_X Y$ that one ends up defining in this way is not just a derivative of Y, it also depends on the first derivative of X. (This is apparent from the local coordinate formula for $[X, Y]$ in Exercise 6.2, which matches $\mathcal{L}_X Y$ by Proposition 6.7.) For this reason, $\mathcal{L}_X Y(p)$ cannot accurately be interpreted as a directional derivative of Y at p in the direction $X(p)$.

It turns out that on a general vector bundle $E \to M$, there is no *canonical* way to define parallel transport along arbitrary paths, so instead of looking for a unique "correct" definition, it is more useful to consider what properties a reasonable definition of parallel transport should be required to satisfy. In particular, we would like the covariant derivative to behave in certain respects the way that derivatives are expected to behave, for instance:

(i) $\nabla_X s := \nabla_t s(\gamma(t))|_{t=0}$ should depend on the section $s \in \Gamma(E)$ near $p \in M$ and the tangent vector $X = \dot{\gamma}(0) \in T_pM$, but not otherwise on the path γ;
(ii) The map $T_pM \to E_p : X \mapsto \nabla_X s$ should be linear.[58]

We will see below that these two conditions lead more-or-less inevitably to the correct definition of a connection on a vector bundle.

Recall that sections of E are by definition smooth maps $s : M \to E$ that satisfy $\pi \circ s = \mathrm{Id}_M$. Similarly, if γ is a smooth path in M, then a smooth path $t \mapsto s(t) \in E$ satisfying

$$\pi(s(t)) = \gamma(t)$$

for all t is called a **lift** of γ to E; equivalently, $s(t)$ belongs to the fiber $E_{\gamma(t)}$ for every t and thus defines a section of the pullback bundle γ^*E, also known as a section of E *along* γ. A family of parallel transport maps $P^t_\gamma : E_p \to E_{\gamma(t)}$ associates to every $v \in E_p$ a lift $s(t) := P^t_\gamma(v)$ of γ such that $s(0) = v$. In order to fully understand this perspective on parallel transport, it may be helpful for a while to forget that E is a vector bundle, and think of it merely as a smooth manifold that happens to be presented as a union of a smooth family of disjoint submanifolds $E_p \subset E$, its fibers. (Objects of this kind—in which the fibers are all disjoint but diffeomorphic submanifolds that need not necessarily be vector spaces—are called *fiber bundles*, and we will study them more seriously next semester.) The fibers are, in particular, the level sets of a smooth submersion $\pi : E \to M$, so differentiating π at $v \in E$ and taking its kernel gives the tangent space to the fiber containing v, which we will call the **vertical subspace** of T_vE:

$$V_vE := \ker\left(T_vE \xrightarrow{\pi_*} T_{\pi(v)}M\right) = T_v(E_{\pi(v)}) \subset T_vE.$$

All together, these subspaces define a distinsuished subbundle of $TE \to E$, called the **vertical subbundle**

$$VE = \ker(\pi_*) = \bigcup_{v \in E} V_vE \subset TE.$$

If we now choose to "unforget" the fact that each fiber E_p is also a vector space, then we notice that since tangent spaces to a vector space can be identified with the vector space itself, there is a canonical isomorphism

$$\mathrm{Vert}_v : E_p \xrightarrow{\cong} V_vE \qquad \text{for every } v \in E_p,\ p \in M,$$

[58] If $\mathbb{F} = \mathbb{C}$, then since T_pM is *not* naturally a complex vector space, we ignore the complex structure of E_p in order to talk about linearity of the map $T_pM \to E_p$, i.e. "linear" just means "real-linear".

sending $w \in E_p$ to $\frac{d}{dt}(v + tw)\big|_{t=0} \in V_v E$. A fancier way to say this is that the isomorphisms Vert_v define a canonical vector bundle isomorphism between $VE \to E$ and the pullback $\pi^* E \to E$ of $E \to M$ via its own bundle projection $\pi : E \to M$. In the lemma below, the vector space structure of the fibers E_p will be relevant for this one detail *only*, and for the most part, it will be more helpful to forget that the fibers E_p are vector spaces and think of them merely as the regular level sets of a smooth submersion $\pi : E \to M$.

Lemma 19.1. *Suppose there is a smooth family of vector space isomorphisms*[59] $P^t_\gamma : E_{\gamma(0)} \to E_{\gamma(t)}$ *associated to every smooth path* $\gamma : (-\epsilon, \epsilon) \to M$ *such that the covariant derivative defined in* (19.3) *satisfies properties (i) and (ii) above. Then for each* $p \in M$ *and* $v \in E_p$*, there is a unique linear injection*

$$\mathrm{Hor}_v : T_p M \to T_v E$$

such that $\mathrm{Hor}_v(\dot\gamma(0)) = \frac{d}{dt} P^t_\gamma(v)\big|_{t=0}$ *for all paths* γ *with* $\gamma(0) = p$*. Moreover,* $\pi_* \circ \mathrm{Hor}_v : T_p M \to T_p M$ *is the identity map, and the image*

$$H_v E := \operatorname{im} \mathrm{Hor}_v \subset T_v E$$

is complementary to the vertical subspace $V_v E \subset T_v E$*, so it determines a splitting of* TE *into a direct sum of smooth subbundles,*

$$TE = VE \oplus HE, \qquad \textit{where} \qquad HE := \bigcup_{v \in E} H_v E.$$

Proof. Fix $p \in M$. For any smooth path $\gamma : (-\epsilon, \epsilon) \to M$ with $\gamma(0) = p$ and any $v \in E_p$, we can think of $t \mapsto P^t_\gamma(v)$ as a smooth path in the total space of the pullback bundle $\gamma^* E = \bigcup_{t \in (-\epsilon,\epsilon)} E_{\gamma(t)}$, i.e. we regard $P^t_\gamma(v)$ is living in the fiber $(\gamma^* E)_t$. We can therefore define a smooth vector field Y on the total space of $\gamma^* E$ such that for any $v \in E_p$ and any $t \in (-\epsilon, \epsilon)$,

$$Y(P^t_\gamma(v)) = \frac{d}{dt} P^t_\gamma(v),$$

and the parallel transport maps P^t_γ can now be written in terms of the flow φ^t_Y of Y as

$$P^t_\gamma = \varphi^t_Y|_{E_p} : E_p = (\gamma^* E)_0 \to (\gamma^* E)_t = E_{\gamma(t)}.$$

The inverse of P^t_γ is then given by reversing the flow of Y, so for a section $s : M \to E$ with $s(p) = v$,

$$\mathrm{Vert}_v\left(\nabla_{\dot\gamma(0)} s\right) = \frac{d}{dt} \varphi^{-t}_Y(s(\gamma(t)))\bigg|_{t=0} \in E_p.$$

Note that $t \mapsto \varphi^{-t}_Y(s(\gamma(t)))$ is a path through the point $\varphi^0_Y(s(p)) = s(p) = v$ in the submanifold $E_p \subset E$, so we are regarding its derivative as an element of $V_v E \subset T_v E$, so that the canonical isomorphism $\mathrm{Vert}_v^{-1} : V_v E \to E_p$ identifies it with the covariant derivative as defined in (19.3). To compute it, we write $F(t_1, t_2) = \varphi^{t_1}_Y(s(\gamma(t_2)))$ and apply the chain rule, giving

$$\mathrm{Vert}_v\left(\nabla_{\dot\gamma(0)} s\right) = \frac{d}{dt} F(-t, t)\bigg|_{t=0} = -\frac{\partial F}{\partial t_1}(0,0) + \frac{\partial F}{\partial t_2}(0,0) = -Y(v) + Ts(\dot\gamma(0)) \in T_v E,$$

and thus

$$\mathrm{Hor}_v(\dot\gamma(0)) := \frac{d}{dt} P^t_\gamma(v)\bigg|_{t=0} = Y(v) = Ts(\dot\gamma(0)) - \mathrm{Vert}_v\left(\nabla_{\dot\gamma(0)} s\right) \in T_v E. \tag{19.4}$$

[59]With very minor modifications, Lemma 19.1 is also valid on an arbitrary fiber bundle, assuming only that the maps $P^t_\gamma : E_{\gamma(0)} \to E_{\gamma(t)}$ are diffeomorphisms. In this case there may not be canonical isomorphisms $V_v E \cong E_{\pi(v)}$ since fibers need not be vector spaces, so the covariant derivative $\nabla_X s$ of a section $s \in \Gamma(E)$ in direction $X \in T_p M$ naturally takes its value in $V_{s(p)} E$ instead of E_p.

Properties (i) and (ii) above imply that this expression is a linear function of $\dot{\gamma}(0)$ and does not otherwise depend on the choice of path γ. Writing $X := \dot{\gamma}(0)$, Hor_v also satisfies

$$\pi_* \circ \mathrm{Hor}_v(X) = T\pi \left(\frac{d}{dt} P^t_\gamma(v) \Big|_{t=0} \right) = \frac{d}{dt} \left(\pi \circ P^t_\gamma(v) \right) \Big|_{t=0} = \frac{d}{dt} \gamma(t) \Big|_{t=0} = X,$$

thus $\mathrm{Hor}_v : T_p M \to T_v E$ is injective and its image $H_v E := \operatorname{im} \mathrm{Hor}_v$ has trivial intersection with $\ker \pi_* = V_v E$. Finally, we observe that any non-vertical vector $\xi \in T_v E \backslash V_v E$ can be written as $Ts(\dot{\gamma}(0))$ for some path γ and section s, and we then have $\xi = \mathrm{Hor}_v(\dot{\gamma}(0)) + \mathrm{Vert}_v \left(\nabla_{\dot{\gamma}(0)} s \right) \in H_v E + V_v E$, thus

$$H_v E \oplus V_v E = T_v E.$$

□

Any subbundle

$$HE \subset TE$$

that satisfies $TE = VE \oplus HE$ as in Lemma 19.1 is called a **horizontal subbundle** of TE. Unlike VE, horizontal subbundles are not unique or canonical, but a choice of horizontal subbundle is equivalent via the formula $H_v E = \operatorname{im} \mathrm{Hor}_v$ to a choice of a smoothly varying family of linear **horizontal lift** maps,

$$\mathrm{Hor}_v : T_p M \to T_v E \qquad \text{such that} \qquad \pi_* \circ \mathrm{Hor}_v(X) = X \text{ for all } X \in T_p M.$$

The lemma tells us that any sensible choice of parallel transport maps for E along smooth paths in M determines a horizontal subbundle in a natural way. Conversely, any horizontal subbundle $HE \subset TE$ uniquely determines parallel transport maps by requiring all parallel lifts $s(t) := P^t_\gamma(v) \in E$ of paths $\gamma(t) \in M$ to be tangent to HE, i.e. the derivative $\dot{s}(t) \in T_{s(t)} E$ should always be horizontal, which means it is the horizontal lift of the derivative of γ:

$$\dot{s}(t) = \mathrm{Hor}_{s(t)}(\dot{\gamma}(t)). \tag{19.5}$$

This is a first-order ordinary differential equation, so $s(t)$ is uniquely determined by the initial condition $s(0) = v$. One can also see this by using horizontal lifts to define a vector field on the total space of $\gamma^* E$ as in the proof of Lemma 19.1; the parallel transport maps are then given by the flow of that vector field.

This is as far as we can go without paying attention to the fact that fibers E_p are vector spaces, and you may notice that a problem has arisen from this relaxation of assumptions. Indeed, for an arbitrary choice of horizontal subbundle $HE \subset TE$, there is no guarantee that the ODE in (19.5) with any given initial condition $s(0) = v$ will have solutions beyond an arbitrarily small interval around $t = 0$, and if it does, then the resulting family of maps $P^t_\gamma : E_p \to E_{\gamma(t)}$ will be diffeomorphisms, but they need not be linear. The following useful characterization of linearity will provide an easy remedy for this.

Lemma 19.2. *Let V and W be normed vector spaces over $\mathbb{F}$. Then any map $F : V \to W$ that is differentiable*[60] *at 0 and satisfies $F(\lambda v) = \lambda F(v)$ for all scalars $\lambda \in \mathbb{F}$ and all $v \in V$ is linear.*

Proof. The key is to show that under this assumption, F is actually equal to its derivative at zero, $DF(0) : V \to W$. Clearly $F(0) = 0$, so we can write

$$F(v) = DF(0)v + \|v\| \cdot R(v)$$

[60] If $\mathbb{F} = \mathbb{C}$, then differentiability of $F : V \to W$ can be taken to mean the same thing as in the real case, i.e. we simply regard V and W as real vector spaces, so the derivative $DF(0) : V \to W$ is a real-linear map. It is not necessary to assume that $DF(0)$ is complex linear, which would be a holomorphicity condition on F.

for some function $R : V \to W$ such that $\lim_{v\to 0} R(v) = 0$. Then taking $\lambda > 0$,

$$F(v) = \lim_{\lambda\to 0^+} \frac{1}{\lambda} F(\lambda v) = \lim_{\lambda\to 0^+} \frac{DF(0)\lambda v + \lambda\|v\| \cdot R(\lambda v)}{\lambda} = DF(0)v + \lim_{\lambda\to 0^+} \|v\| \cdot R(\lambda v) = dF(0)v.$$

Since $DF(0) : V \to W$ is real linear, this proves that F also respects vector addition. □

On our vector bundle $E \to M$, each scalar $\lambda \in \mathbb{F}$ defines a smooth map

$$m_\lambda : E \to E : v \mapsto \lambda v,$$

which is a diffeomorphism for $\lambda \neq 0$, and its tangent map $(m_\lambda)_* : TE \to TE$ then defines vector space isomorphisms $(m_\lambda)_* : T_vE \to T_{\lambda v}E$ for every $v \in E$.

Lemma 19.3. *For a given horizontal subbundle $HE \subset TE$, the following conditions are equivalent:*

(i) The parallel transport maps $P^t_\gamma : E_p \to E_{\gamma(t)}$ defined via (19.5) exist for every t in the domain of an arbitrary smooth path γ, and are linear.
(ii) For all $v \in E$ and $\lambda \in \mathbb{F}$, $H_{\lambda v}E = (m_\lambda)_* (H_vE)$.

Proof. For any $p \in M$, $v \in E_p$, $\lambda \in \mathbb{F}$ and a smooth path γ in M through $\gamma(0) = p$, assume first that P^t_γ exists and is linear for every t. Writing $s(t) := P^t_\gamma(v)$, we have $\dot{s}(0) \in H_vE$ by definition. The corresponding lift with initial condition $\lambda v \in E_p$ is then $P^t_\gamma(\lambda v) = \lambda s(t) = m_\lambda (s(t))$, implying

$$\frac{d}{dt} m_\lambda (s(t))\Big|_{t=0} = (m_\lambda)_* \dot{s}(0) \in H_{\lambda v}E,$$

hence $(m_\lambda)_*$ maps H_vE to $H_{\lambda v}E$. Conversely, if this condition on HE holds, then for $\lambda \neq 0$, the same calculation implies that $s(t)$ is a horizontal lift of $\gamma(t)$ if and only if $\lambda s(t)$ is. Here it is convenient to assume $\lambda \neq 0$ so that the map $T_vE \xrightarrow{(m_\lambda)_*} T_{\lambda v}E$ is an isomorphism, but since the fibers of HE vary continuously, one can also take $\lambda \to 0$ and conclude that at points along the zero-section

$$Z := \bigcup_{p\in M} \{0\} \subset E,$$

the horizontal subspaces are uniquely determined, namely $H_vE = T_vZ$ whenever $v \in Z$. It follows that solutions to (19.5) with initial condition $s(0) = 0$ exist for all t and are identically zero. This implies in turn that for any t, solutions also exist with initial condition in some sufficiently small neighborhood of 0, but the ability to find further solutions via multiplication with arbitrary scalars now produces solutions for all t with arbitrary initial conditions. Moreover, the resulting diffeomorphisms $P^t_\gamma : E_p \to E_{\gamma(t)}$ are smooth and respect scalar multiplication, so by Lemma 19.2, they are linear. □

19.2. Two equivalent definitions. The point of the previous section was to motivate the following definition.

Definition 19.4 (Connections, version 1). A **connection** (*Zusammenhang*) on the vector bundle $\pi : E \to M$ is a choice of subbundle

$$HE \subset TE$$

that is complementary to the vertical subbundle $VE \subset TE$ and satisfies $(m_\lambda)_* (HE) = HE$ for every scalar $\lambda \in \mathbb{F}$.

It should not be obvious to you at this stage whether connections always exist—they do, but this is something we will have to prove,, and the proof unsurprisingly requires a partition of unity. If a connection is chosen, then it determines the notions of parallel transport, horizontal lifts and covariant derivatives as we defined them in the previous section.

There are at least two other popular ways to reformulate Definition 19.4. One of them uses the fact that a connection determines a splitting $TE = VE \oplus HE$, and splittings of vector spaces (or bundles) can be characterized in terms of linear projection maps. Indeed, let

$$\widehat{K} : TE \to VE$$

be the unique smooth linear bundle map that restricts to the identity on VE and vanishes on HE, so $HE = \ker \widehat{K}$. Since each vertical subspace $V_v E$ is canonically isomorphic to the fiber $E_{\pi(p)}$, we can compose $\widehat{K}$ with the resulting canonical map $VE \to E$ to produce a map $K : TE \to E$ as in the following definition.

Definition 19.5 (Connections, version 2). A **connection** (*Zusammenhang*) on the vector bundle $\pi : E \to M$ is a smooth map $K : TE \to E$ such that

(1) For each $v \in E$, K defines a real-linear map $T_v E \to E_{\pi(v)}$.[61]
(2) $K(\mathrm{Vert}_v(w)) = w$ for all $v, w \in E_p$, $p \in M$.
(3) For all scalars $\lambda \in \mathbb{F}$, $K \circ (m_\lambda)_* = m_\lambda \circ K$.

To get from Definition 19.5 back to Definition 19.4, one defines a hoirzontal subbundle $HE \subset TE$ from $K : TE \to E$ by

$$HE := \ker(K) \subset TE,$$

meaning $H_v E$ is the kernel of the linear map $T_v E \xrightarrow{K} E_{\pi(v)}$.

Exercise 19.6. Show that under the correspondence described above between horizontal subbundles $HE \subset TE$ and maps $K : TE \to E$, the condition $(m_*)(HE) = HE$ is equivalent to $K \circ (m_\lambda)_* = m_\lambda \circ K$ for all $\lambda \in \mathbb{F}$.

The projection $K : TE \to E$ provides a simpler formula for the covariant derivative of a section $s \in \Gamma(E)$ in the direction of a tangent vector $X \in T_p M$ at a point $p \in M$. Recall from (19.4) the relation

$$Ts(X) = \mathrm{Hor}_{s(p)}(X) + \mathrm{Vert}_{s(p)}(\nabla_X s)\,.$$

Since K annihilates horizontal vectors, applying it to both sides of this relation gives

$$\boxed{\nabla_X s = K \circ Ts(X)}, \tag{19.6}$$

so the covariant derivative is actually just the "vertical part" of the tangent map of $s : M \to E$ in the direction of X, obtained by removing from $Ts(X)$ its horizontal part and then identifying the resulting vertical vector with an element of E_p. Note that although the vertical subbundle is independent of any choices, the notion of a "vertical part" of a vector in TE does depend on the choice of the complementary subbundle $HE \subset TE$ along which to project it. The covariant derivative thus depends on the choice of connection, except in certain special situations such as the following.

Exercise 19.7. Assume $E \to M$ is a vector bundle and $s \in \Gamma(E)$. The **zero set** (*Nullstelle*) of s, sometimes denoted by $s^{-1}(0) \subset M$, is the set of all points $p \in M$ such that $s(p)$ is the zero vector in its respective fiber.

[61] We emphasize that if $\mathbb{F} = \mathbb{C}$, then E must be treated as a real vector space for the purposes of this condition, as the tangent spaces $T_v E$ are not complex in any natural way. This is because M is only a real manifold, not complex.

(a) Show that for any $p \in s^{-1}(0)$ the linear map

$$Ds(p) : T_pM \to E_p : X \mapsto \nabla_X s$$

is independent of the choice of connection (needed for defining $\nabla_X s$). We call this the **linearization** (*Linearisierung*) of s at p.

(b) We say that $s \in \Gamma(E)$ is **transverse to the zero-section** (*transversal zum Nullschnitt*) if the linearization $Ds(p) : T_pM \to E_p$ is surjective for every $p \in s^{-1}(0)$. Show that whenever this holds, $s^{-1}(0)$ is a smooth submanifold of M, with dimension $\dim(M) - \operatorname{rank}(E)$ in the case $\mathbb{F} = \mathbb{R}$, or $\dim(M) - 2\operatorname{rank}(E)$ in the case $\mathbb{F} = \mathbb{C}$.
Hint: Any local trivialization $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^m$ determines a connection on $E|_{\mathcal{U}_\alpha}$ such that covariantly differentiating s over $\mathcal{U}_\alpha$ becomes equivalent to differentiating its local repesentative $s_\alpha : \mathcal{U}_\alpha \to \mathbb{F}$.

20. More on connections

As in the previous lecture, we fix a smooth vector bundle $\pi : E \to M$ of rank m over the field $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, where the base M is a smooth n-manifold.

20.1. The Leibniz rule (a third definition). In our discussion so far, choosing a connection on $\pi : E \to M$ means choosing a horizontal subbundle $HE \subset TE$ that satisfies the conditions of Definition 19.4, or equivalently, a map $K : TE \to E$ satisfying the conditions of Definition 19.5. (We will sometimes refer to K as the **vertical projection** defining the connection.) We have two ways of writing down the covariant derivative operator determined by this connection: for a section $s \in \Gamma(E)$, point $p \in M$ and tangent vector $X \in T_pM$, we can choose a smooth path $\gamma : (-\epsilon, \epsilon) \to M$ with $\gamma(0) = p$ and $\dot\gamma(0) = X$, and write

$$\nabla_X s = \frac{d}{dt} (P^t_\gamma)^{-1} \left(s(\gamma(t)) \right) \Big|_{t=0} \in E_p. \tag{20.1}$$

Alternatively, we saw in (19.6) that $\nabla_X s$ can be written in terms of $K : TE \to E$ and the tangent map $Ts : TM \to TE$ of $s : M \to E$ as

$$\nabla_X s = K(Ts(X)).$$

While this formula looks simpler, (20.1) is often more useful for proving basic properties of the covariant derivative, for instance the Leibniz rule in Exercise 20.3 below.

Example 20.1. On the trivial bundle $E = M \times \mathbb{F}^m$, there is a natural **trivial connection**, defined by viewing the two factors in the obvious splitting $T_{(p,v)}E = T_pM \oplus T_v\mathbb{F}^m = T_pM \oplus \mathbb{F}^m$ as horizontal and vertical subspaces respectively. The vertical projection $K : TE \to E$ is then given by $K(X, w) = w \in \mathbb{F}^m = E_v$ for $(X, w) \in T_pM \oplus \mathbb{F}^m = T_{(p,v)}E$, and the parallel transport maps $P^t_\gamma : E_{\gamma(0)} = \mathbb{F}^m \to \mathbb{F}^m = E_{\gamma(t)}$ are all the identity map on $\mathbb{F}^m$. Under the obvious identification of sections $s \in \Gamma(E)$ with functions $f : M \to \mathbb{F}^m$, the covariant derivative $\nabla_X s$ is then simply the differential $df(X)$.

Since $\nabla_X s$ depends linearly on X, the covariant derivative of $s \in \Gamma(E)$ in all possible directions can be packaged as a section

$$\nabla s \in \Gamma(\operatorname{Hom}(TM, E))$$

defined by $\nabla s(p)(X) := \nabla_X s$. There is a clear analogy here with differentials: a real-valued function $f : M \to \mathbb{R}$ is the same thing as a section of the trivial real line bundle $M \times \mathbb{R} \to M$, and its differential assigns to every point $p \in M$ the linear map $d_pf : T_pM \to \mathbb{R}$. The covariant derivative of $s \in \Gamma(E)$ similarly assigns to each point $p \in M$ a linear map $\nabla s(p) : T_pM \to E_p : X \mapsto \nabla_X s$, defining what is sometimes called a “bundle-valued” 1-form $\nabla s \in \Omega^1(M, E) := \Gamma(\operatorname{Hom}(TM, E))$.

Note that if E is a complex vector bundle, then $\mathrm{Hom}(TM, E)$ means the bundle of *real*-linear maps from TM to E, since TM is not naturally a complex bundle. On the other hand, $\mathrm{Hom}(TM, E)$ does have a natural complex structure if E is complex, in which case $\Gamma(\mathrm{Hom}(TM, E))$ is also a complex vector space and one can therefore speak (as in Exercise 20.2 below) of complex-linear maps $\Gamma(E) \to \Gamma(\mathrm{Hom}(TM, E))$.

The next two exercises are both easy applications of (20.1), and depend crucially on the fact that parallel transport maps $P^t_\gamma : E_{\gamma(0)} \to E_{\gamma(t)}$ are linear.

EXERCISE 20.2. Show that the map $\nabla : \Gamma(E) \to \Gamma(\mathrm{Hom}(TM, E))$ is linear.

EXERCISE 20.3. Show that for any $s \in \Gamma(E)$ and $f \in C^\infty(M, \mathbb{F})$, the covariant derivative of $fs \in \Gamma(E)$ in a direction $X \in T_pM$ at $p \in M$ satisfies

$$\nabla_X(fs) = df(X)s(p) + f(p)\nabla_X s.$$

This Leibniz rule is often abbreviated in the form

$$\nabla(fs) = df \cdot s + f\,\nabla s. \tag{20.2}$$

Exercises 20.2 and 20.3 have a converse of sorts, which leads to yet another equivalent version of the definition of a connection.

DEFINITION 20.4 (Connections, version 3). A **connection** (*Zusammenhang*) on the vector bundle $\pi : E \to M$ is a linear operator

$$\nabla : \Gamma(E) \to \Gamma(\mathrm{Hom}(TM, E))$$

that satisfies the Leibniz rule (20.2) for all $f \in C^\infty(M, \mathbb{F})$ and $s \in \Gamma(E)$.

To see that this is equivalent to our previous two definitions, we need to show that every linear operator $\nabla : \Gamma(E) \to \Gamma(\mathrm{Hom}(TM, E))$ satisfying the Leibniz rule (20.2) is in fact the covariant derivative operator determined by a unique connection in the sense of Definitions 19.4 and 19.5. The uniqueness is an easy consequence of (19.4), since for any $p \in M$, $X \in T_pM$ and $v \in E_p$, one can choose any section $s \in \Gamma(E)$ with $s(p) = v$ and write

$$\mathrm{Hor}_v(X) = Ts(X) - \mathrm{Vert}_v(\nabla s(X)), \tag{20.3}$$

thus using the operator ∇ to determine the horizontal lift maps Hor_v, and in this way the horizontal subbundle $HE \subset TE$. Existence will follow similarly if we can show that the right hand side of this expression does not depend on the choice of section s with $s(p) = v$. The following result will help, and is important for other reasons as well.

PROPOSITION 20.5. *For any two connections $\nabla, \hat{\nabla}$ on $\pi : E \to M$ in the sense of Definition 20.4, there exists a smooth linear bundle map $A : E \to \mathrm{Hom}(TM, E)$ such that $\hat{\nabla}s = \nabla s + As$ for all $s \in \Gamma(E)$.*

PROOF. We can use a minor adaptation of the notion of C^∞-linearity from §8.1. For any two vector bundles E and F, a smooth linear bundle map $A : E \to F$ defines a linear map $\Gamma(E) \to \Gamma(F) : s \mapsto As$ that is also C^∞-linear in the sense that fs is sent to $f \cdot As$ for any $f \in C^\infty(M, \mathbb{F})$. Conversely, any C^∞-linear map $\hat{A} : \Gamma(E) \to \Gamma(F)$ arises in this way from a smooth linear bundle map $A : E \to F$, meaning in particular that for any $s \in \Gamma(E)$, the value at any given point $p \in M$ of the section $\hat{A}s \in \Gamma(F)$ is determined by $s(p) \in E_p$ and is otherwise independent of the section s. The proof of this statement is almost identical to that of Proposition 8.2.

With this understood, we observe that while the term $df \cdot s$ in the Leibniz rule prevents either of ∇ or $\hat{\nabla}$ from being a C^∞-linear map $\Gamma(E) \to \Gamma(\mathrm{Hom}(TM, E))$, this term is identical for both connections and thus cancels when we consider $A := \hat{\nabla} - \nabla : \Gamma(E) \to \Gamma(\mathrm{Hom}(TM, E))$. It follows that the latter is C^∞-linear, and thus arises from a bundle map $E \to \mathrm{Hom}(TM, E)$. $\square$

REMARK 20.6. By Definition 20.4, the set of all connections on $\pi : E \to M$ can be regarded as a subset of the infinite-dimensional vector space $\operatorname{Hom}(\Gamma(E), \Gamma(\operatorname{Hom}(TM, E)))$, but it is not a linear subspace, e.g. it does not contain the zero element of this space. Proposition 20.5 shows however that it is an *affine* space over the vector space $\Gamma(\operatorname{Hom}(E, \operatorname{Hom}(TM, E)))$, which sits naturally inside $\operatorname{Hom}(\Gamma(E), \Gamma(\operatorname{Hom}(TM, E)))$ as the space of all C^∞-linear maps $\Gamma(E) \to \Gamma(\operatorname{Hom}(TM, E))$. This shows in particular that the set of connections is convex.

Returning to (20.3), the right hand side clearly depends on $s \in \Gamma(E)$ only in a neighborhood of p, thus we are free to restrict our attention to a small neighborhood $\mathcal{U} \subset M$ of p on which E is a trivial bundle. Choosing a trivialization $\Phi : E|_{\mathcal{U}} \to \mathcal{U} \times \mathbb{F}^m$ yields a corresponding choice of parallel transport isomorphisms, for which sections are parallel if and only if their representations in the local trivialization are constant—this defines a connection in the sense of our previous two definitions, and it matches the "trivial" connection of Example 20.1 if we use Φ to identify $E|_{\mathcal{U}}$ with the trivial bundle $\mathcal{U} \times \mathbb{F}^m$. Let us denote the horizontal lift and covariant derivative operators for this connection by $\widehat{\operatorname{Hor}}_v$ and $\hat{\nabla}$ respectively. According to Proposition 20.5, $\hat{\nabla} = \nabla + A$ for a bundle map $A : E \to \operatorname{Hom}(TM, E)$, and (20.3) therefore becomes

$$\begin{aligned}\operatorname{Hor}_v(X) &= Ts(X) - \operatorname{Vert}_v(\nabla s(X)) = Ts(X) - \operatorname{Vert}_v\left(\hat{\nabla} s(X) - Av(X)\right) \\ &= Ts(X) - \operatorname{Vert}_v\left(\hat{\nabla} s(X)\right) + \operatorname{Vert}_v\left(Av(X)\right) = \widehat{\operatorname{Hor}}_v(X) + \operatorname{Vert}_v\left(Av(X)\right).\end{aligned}$$

Now it is clear that the right hand side does not depend on the choice of section s satisfying $s(p) = v$, thus proving that any operator ∇ as in Definition 20.4 uniquely determines a horizontal subbundle $HE \subset TE$ whose covariant derivative operator is ∇. That the parallel transport maps arising from HE are linear can then be deduced from the assumption that $\nabla : \Gamma(E) \to \Gamma(\operatorname{Hom}(TM, E))$ is linear: indeed, for any path γ with $\dot{\gamma}(0) = X \neq 0 \in T_pM$, sections $s \in \Gamma(E)$ that are parallel along γ are characterized by the condition

$$\nabla_{\dot{\gamma}(t)} s = 0 \quad \text{for all } t,$$

and the set of solutions to this equation is a vector space. It follows via Lemma 19.3 that HE satisfies the conditions of Definition 19.4, and all three of our definitions of a connection are therefore equivalent.[62]

20.2. Local coordinates and Christoffel symbols. There are two standard ways to present a connection in local coordinates, and both rely mainly on the same two facts: (1) every trivialization determines a corresponding *trivial* connection as in Example 20.1, and (2) by Proposition 20.5, every other connection differs from that one by a bundle map. This bundle map always appears in coordinates as a so-called "zeroth-order" term, meaning that unlike the covariant derivative itself, it is not a *differential* operator.

Fix a local trivialization

$$\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^m$$

over some open subset $\mathcal{U}_\alpha \subset M$, and let D denote the covariant derivative operator for the resulting trivial connection on $E|_{\mathcal{U}_\alpha}$, i.e. the one for which sections are parallel if and only if Φ_α identifies them with constant functions. Given any other connection ∇ on E, $\nabla - D$ then defines a smooth linear bundle map $A : E|_{\mathcal{U}_\alpha} \to \operatorname{Hom}(TM, E)|_{\mathcal{U}_\alpha}$, which we shall write in the form

$$(Av)X = \Gamma_\alpha(X, v) \in E_p \quad \text{for} \quad p \in \mathcal{U}_\alpha, \ X \in T_pM, \ v \in E_p,$$

[62] The argument for why Definition 20.4 implies Definition 19.4 unfortunately got skipped in the lecture, due mainly to absent-mindedness.

thus defining a smooth *bilinear* bundle map

$$\Gamma_\alpha : (TM \oplus E)|_{\mathcal{U}_\alpha} \to E|_{\mathcal{U}_\alpha}.$$

For any section $s \in \Gamma(E)$, ∇s can now be expressed over $\mathcal{U}_\alpha$ in the form

$$\nabla_X s = D_X s + \Gamma_\alpha(X, s(p)) \qquad \text{for } p \in \mathcal{U}_\alpha,\ X \in T_pM. \tag{20.4}$$

Note that Γ_α is real linear in the first factor and $\mathbb{F}$-linear in the second. It must be emphasized that Γ_α is not globally defined, and it depends on the choice of trivialization.

One sees Γ_α expressed more often in local coordinates as a set of locally defined functions with three indices. Assume $\mathcal{U}_\alpha$ admits a coordinate chart $(x^1, \dots, x^n)$; this then determines a frame $(\partial_1, \dots, \partial_n)$ for the tangent bundle $TM|_{\mathcal{U}_\alpha}$. There is similarly a frame $(e_1, \dots, e_m)$ for $E|_{\mathcal{U}_\alpha}$ corresponding to the trivialization Φ_α. Then there are smooth functions

$$\Gamma^a_{ib} : \mathcal{U}_\alpha \to \mathbb{F}, \qquad i \in \{1, \dots, n\},\ a, b \in \{1, \dots, m\}$$

uniquely determined by the condition

$$\Gamma_\alpha(\partial_i, e_b) = \Gamma^a_{ib} e_a.$$

For any $X = X^i \partial_i \in T_pM$ and $v = v^b e_b \in E_p$ at a point $p \in \mathcal{U}_\alpha$, we then have

$$\Gamma_\alpha(X, v) = \Gamma_\alpha(X^i \partial_i, v^b e_b) = X^i v^b \Gamma_\alpha(\partial_i, e_b) = \Gamma^a_{ib} X^i v^b e_a,$$

so the ath component of $\Gamma_\alpha(X, v) \in E_p$ with respect to the frame $e_1, \dots, e_m$ is

$$(\Gamma_\alpha(X, v))^a = \Gamma^a_{ib} X^i v^b.$$

The functions Γ^a_{ib} are called the **Christoffel symbols** determined by the connection.

Recall that any section $s \in \Gamma(E)$ can be expressed over $\mathcal{U}_\alpha$ in terms of its component functions $s^1, \dots, s^m : \mathcal{U}_\alpha \to \mathbb{F}$ as $s = s^a e_a$. Let us write

$$\nabla_i := \nabla_{\partial_i} = \nabla_{\frac{\partial}{\partial x^i}}$$

for the covariant derivative operator in the ith coordinate direction. One now obtains another formula for the Christoffel symbols from (20.4), using the observation that the frame sections $e_1, \dots, e_m$ all satisfy $De_a \equiv 0$ by the definition of the trivial connection. Indeed, this together with (20.4) implies $\nabla_i e_b = \Gamma_\alpha(\partial_i, e_b) = \Gamma^a_{ib} e_a$, and thus

$$\Gamma^a_{ib} = (\nabla_i e_b)^a. \tag{20.5}$$

For a general section $s = s^a e_a$ over $\mathcal{U}_\alpha$, we then apply the Leibniz rule to compute

$$\nabla_i s = \nabla_i (s^b e_b) = (\partial_i s^b) e_b + s^b \nabla_i e_b = \left(\partial_i s^a + \Gamma^a_{ib} s^b\right) e_a,$$

where we've relabelled the summed index in the first term and used (20.5) in the second term, giving rise to the formula

$$(\nabla_i s)^a = \partial_i s^a + \Gamma^a_{ib} s^b. \tag{20.6}$$

This is of practical use for coordinate computations of covariant derivatives.

You can see from (20.6) that the Christoffel symbols Γ^a_{ib} fully determine the covariant derivative operator, and therefore the connection, at least over the region $\mathcal{U}_\alpha$. This observation gives rise to yet another variant of the definition of a connection, one that is not very elegant, but is favored by physicists: a connection is an association to every open set $\mathcal{U}_\alpha \subset M$ with a chart $x^1, \dots, x^n$ and local trivialization $(\mathcal{U}_\alpha, \Phi_\alpha)$ of a set of smooth functions $\Gamma^a_{ib} : \mathcal{U}_\alpha \to \mathbb{F}$, which are then fed into (20.6) to define the covariant derivative. Of course, the functions Γ^a_{ib} cannot be chosen arbitrarily for all possible local trivializations and charts: once they have been chosen for one particular chart and trivialization over a set $\mathcal{U}_\alpha$, the connection over $\mathcal{U}_\alpha$ is fully determined, and any choice on a different region $\mathcal{U}_\beta$ (with different coordinates and trivialization) that overlaps $\mathcal{U}_\alpha$ had better give the same result on $\mathcal{U}_\alpha \cap \mathcal{U}_\beta$. One must therefore derive a suitable transformation formula for Christoffel symbols under changes of coordinates and local trivializations, and make sure that that formula is always satisfied. The unfortunate fact is that the correct transformation formula does not follow from anything we've already done, because Christoffel symbols do *not* define a tensor, i.e. since Γ_α always depends on the choice of trivialization Φ_α and is defined only on $\mathcal{U}_\alpha$, there is generally no globally defined tensor field or section of any vector bundle whose locally-defined components are the functions Γ^a_{ib}. This does not make the situation impossible, it only means there is still some work to be done if you want to use Christoffel symbols as a complete characterization of a connection. We leave the details as an exercise:

Exercise 20.7. Given a bundle $\pi : E \to M$ and a sufficiently small open set $\mathcal{U} \subset M$, let us use a coordinate chart $(x^1, \dots, x^n)$ and a frame $(e_1, \dots, e_m)$ to identify $E|_{\mathcal{U}}$ with the trivial bundle $\mathcal{V} \times \mathbb{F}^m$, where $\mathcal{V}$ is an open subset of $\mathbb{R}^n$. Suppose Γ^a_{ib} are the corresponding Christoffel symbols for some connection ∇ on E. Then another choice of coordinates and frame over the same region can be expressed via smooth functions

$$\begin{aligned}
\mathcal{V} \to \mathbb{R}^n &: (x^1, \dots, x^n) \mapsto (\tilde{x}^1, \dots, \tilde{x}^n) \\
\mathcal{V} \to \mathbb{R}^m &: (x^1, \dots, x^n) \mapsto \tilde{e}_1 = (\tilde{e}^1_1, \dots, \tilde{e}^m_1) \\
&\vdots \\
\mathcal{V} \to \mathbb{R}^m &: (x^1, \dots, x^n) \mapsto \tilde{e}_m = (\tilde{e}^1_m, \dots, \tilde{e}^m_m)
\end{aligned}$$

Let $\widetilde{\Gamma}^a_{ib}$ denote the Christoffel symbols of ∇ with respect to the coordinates $(\tilde{x}^1, \dots, \tilde{x}^n)$ and frame $(\tilde{e}_1, \dots, \tilde{e}_m)$. Derive the transformation formula

$$\widetilde{\Gamma}^a_{ib} = \frac{\partial x^j}{\partial \tilde{x}^i} \tilde{e}^c_b \Gamma^a_{jc} + \frac{\partial x^j}{\partial \tilde{x}^i} \frac{\partial}{\partial x^j} \tilde{e}^a_b.$$

As a special case when $E = TM$, show that this becomes

$$\widetilde{\Gamma}^i_{jk} = \frac{\partial x^p}{\partial \tilde{x}^j} \frac{\partial x^q}{\partial \tilde{x}^k} \Gamma^i_{pq} + \frac{\partial x^p}{\partial \tilde{x}^j} \frac{\partial}{\partial x^p} \left(\frac{\partial x^i}{\partial \tilde{x}^k} \right).$$

Remark: I have to be honest—I don't actually recommend doing this exercise. But a physicist would consider it essential.

20.3. Connection 1-forms and G-structures. As an alternative to the Christoffel symbols, one can express covariant derivatives in local trivializations via matrix-valued 1-forms. Suppose again that $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^m$ is a trivialization over some open subset $\mathcal{U}_\alpha \subset M$, and write

$$\Phi_\alpha(v) = (p, v_\alpha) \qquad \text{for } p \in \mathcal{U}_\alpha,\ v \in E_p,$$

thus defining $v_\alpha \in \mathbb{F}^m$. This is just a pointwise version of our usual "local representation" of sections $s \in \Gamma(E)$ over $\mathcal{U}_\alpha$ as functions $s_\alpha : \mathcal{U}_\alpha \to \mathbb{F}^m$, defined such that

$$\Phi_\alpha \circ s(p) = (p, s_\alpha(p)) \qquad \text{for } p \in \mathcal{U}_\alpha.$$

In terms of local repesentatives, the trivial connection D determined by the trivialization Φ_α becomes (cf. Example 20.1) the standard differential, meaning

$$(D_X s)_\alpha = ds_\alpha(X) \qquad \text{for } p \in \mathcal{U}_\alpha,\ X \in T_pM.$$

Any other connection ∇ is related to this one by a bundle map $E \to \mathrm{Hom}(TM, E)$ over $\mathcal{U}_\alpha$. We defined the Christoffel symbols by reinterpreting this as a bilinear bundle map $TM \oplus E \to E$, but we could also choose to interpret it instead as a bundle map $TM \to \mathrm{End}(E)$ over $\mathcal{U}_\alpha$. Using the trivialization to identify fibers of E with $\mathbb{F}^m$, the fibers of $\operatorname{End}(E)$ then become the space of matrices $\mathbb{F}^{m\times m}$, and we deduce the existence of a unique m-by-m matrix-valued 1-form

$$A_\alpha \in \Omega^1(\mathcal{U}_\alpha, \mathbb{F}^{m\times m})$$

such that the covariant derivative ∇ is given in the local trivialization over $\mathcal{U}_\alpha$ by the formula

$$(\nabla_X s)_\alpha(p) = ds_\alpha(X) + A_\alpha(X)s_\alpha(p), \qquad \text{for } p \in \mathcal{U}_\alpha,\ X \in T_pM, \tag{20.7}$$

often abbreviated as

$$(\nabla s)_\alpha = ds_\alpha + A_\alpha s_\alpha.$$

A word on notation: for any manifold M and any finite-dimensional (real or complex) vector space V, we will from now on denote by

$$\Omega^1(M, V) = \{\text{smooth “}V\text{-valued” 1-forms on } M\}$$

the vector space of smooth maps $\omega : TM \to V$ whose restrictions $\omega_p : T_pM \to V$ to the tangent space over each point $p \in M$ are real-linear maps. In the case $V = \mathbb{F}^{m\times m}$ seen above, elements of $\Omega^1(\mathcal{U}_\alpha, \mathbb{F}^{m\times m})$ can also be imagined as m-by-m matrices whose individual entries are smooth $\mathbb{F}$-valued 1-forms on $\mathcal{U}_\alpha$.

The existence and uniqueness of the **connection 1-form** $A_\alpha \in \Omega^1(\mathcal{U}_\alpha, \mathbb{F}^{m\times m})$ satisfying (20.7) was deduced above from Proposition 20.5, but if you prefer, you could also derive a precise formula for A_α from the Christoffel symbols:

Exercise 20.8. Given a coordinate chart $(x^1, \ldots, x^n)$ on $\mathcal{U}_\alpha$, show that at each point in $\mathcal{U}_\alpha$ and for each $i = 1, \ldots, n$, the entries $A_\alpha(\partial_i)^a{}_b$ of the matrix $A_\alpha(\partial_i) \in \mathbb{F}^{m\times m}$ are the Christoffel symbols Γ^a_{ib}.

Equation 20.7 leads to yet another somewhat untidy definition of connections that is nonetheless popular in the physics world: a connection is a choice of m-by-m matrix-valued 1-forms A_α over open subsets $\mathcal{U}_\alpha$, one for each local trivialization $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^m$, such that a certain transformation property with respect to change of trivializations on overlap regions is satisfied (see the exercise below).

Exercise 20.9. If $g = g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{GL}(m, \mathbb{F})$ is the transition map relating two trivializations Φ_α and Φ_β, show that the connection 1-forms A_α and A_β are related on $\mathcal{U}_\alpha \cap \mathcal{U}_\beta$ by

$$A_\alpha(X) = g(p)^{-1} A_\beta(X) g(p) + g(p)^{-1}\, dg(X), \qquad \text{for } p \in \mathcal{U}_\alpha \cap \mathcal{U}_\beta,\ X \in T_pM.$$

This transformation formula is often abbreviated by

$$A_\alpha = g^{-1} A_\beta g + g^{-1}\, dg \qquad \text{on } \mathcal{U}_\alpha \cap \mathcal{U}_\beta. \tag{20.8}$$

Physicists refer to (20.8) as a **gauge transformation** (*Eichtransformation*), alluding to the important role that connection 1-forms play in quantum field theory: in that context they are called *gauge fields*, and they serve to model elementary particles such as photons and other “gauge bosons” that mediate the fundamental forces of nature. The choice of the letter A to denote a connection form is in fact motivated by physics, where the *vector potential* of classical electromagnetic field theory (conventionally denoted by $\mathbf{A}$) can be interpreted as a connection form for a trivial Hermitian line bundle.

There is another reason to use connection 1-forms rather than Christoffel symbols when the vector bundle has extra structure. In this case it’s appropriate to restrict attention to a particular class of connections, and it turns out that this restriction can be expressed elegantly via the connection forms.

Definition 20.10. Let $\pi : E \to M$ be a vector bundle with a G-structure, for some Lie subgroup $G \subset \mathrm{GL}(m, \mathbb{F})$. Then a connection ∇ on E is called **G-compatible** if all parallel transport isomorphisms respect the G-structure: this means that for any path $\gamma(t) \in \mathcal{U}_\alpha$ in the domain of a G-compatible local trivialization $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^m$, the maps $P^t_\gamma : E_{\gamma(0)} \to E_{\gamma(t)}$ satisfy

$$\Phi_\alpha \circ P^t_\gamma \circ \Phi_\alpha^{-1}(\gamma(0), v) = (\gamma(t), g(t)v) \qquad \text{for all } t \text{ and } v \in E_{\gamma(0)},$$

where $g(t) \in G$ is a smooth path in G through $g(0) = \mathbb{1}$.

The definition seems less abstract when we apply it to particular structures: e.g. for $G = \mathrm{O}(m)$ or $\mathrm{U}(m)$, the structure in question is a bundle metric, and the condition above means that parallel transport maps preserve the inner products on the fibers, i.e. they are isometries. In this situation we call ∇ a **metric connection**.

Example 20.11. Recall from §18.7 that $\mathrm{GL}(m, \mathbb{C})$ can be regarded as a subgroup of $\mathrm{GL}(2m, \mathbb{R})$ so that $\mathrm{GL}(m, \mathbb{C})$-structures on a real vector bundle $E \to M$ of rank $2m$ are equivalent to complex structures on E, which make all fibers into complex m-dimensional vector spaces. A connection on the *real* bundle E is then $\mathrm{GL}(m, \mathbb{C})$-compatible if and only if the parallel transport maps are complex linear for this complex structure, and ∇ is then called a **complex connection**. Note: if we had regarded E as a complex vector bundle in the first place, then choosing a connection ∇ on that bundle would have *automatically* meant that parallel transport is complex linear, so you may be wondering why it is useful to single out a special class of "complex connections" on a real vector bundle. One answer to this question is as follows: as we will soon see, every Riemannian manifold (M, g) has a canonical connection on its tangent bundle $TM \to M$, called the *Levi-Cività connection*, which is used for defining the standard Riemannian notions of parallel vector fields and curvature. In certain situations, especially if M is also a *symplectic* manifold, it is also useful to endow M with an almost complex structure (cf. §7.1.4), meaning a bundle map $J : TM \to TM$ that satisfies $J^2 \equiv -\mathbb{1}$ everywhere, thus making $TM \to M$ into a complex vector bundle. While complex connections on TM always exist, there is no guarantee in general that the Levi-Cività connection is one—this turns out to be true if and only if g and J satisfy a very rigid compatibility condition, guaranteeing that J is integrable (cf. Exercise 8.5), hence M in this situation is a complex manifold with a special type of Riemannian metric, called a *Kähler metric*.

We will prove in the next lecture that G-compatible connections always exist. The real strength of connection 1-forms is that they give an easy characterization of the G-compatibility condition. Recall from §18.1 that the *Lie algebra* of a Lie subgroup $G \subset \mathrm{GL}(m, \mathbb{F})$ is the tangent space $\mathfrak{g} := T_{\mathbb{1}}G \subset T_{\mathbb{1}}\,\mathrm{GL}(m, \mathbb{F}) = \mathbb{F}^{m\times m}$.

Theorem 20.12. *If $E \to M$ is a vector bundle with a G-structure and ∇ is a connection on E, then ∇ is G-compatible if and only if for every G-compatible trivialization Φ_α, the corresponding connection 1-form takes values in the Lie algebra $\mathfrak{g} \subset \mathbb{F}^{m\times m}$ of G, i.e.*

$$A_\alpha \in \Omega^1(\mathcal{U}_\alpha, \mathfrak{g}).$$

Before proving the theorem, it will be helpful to deal with a minor technical point. We have occasionally mentioned the notion of a **section along a path**, meaning the following: given a path $\gamma(t) \in M$, we associate to each t in its domain a vector

$$s(t) \in E_{\gamma(t)},$$

so strictly speaking, s is a section of the pullback bundle γ^*E. While s is not quite the same thing as a section of E, there is a straightforward way to define the covariant derivative of s with respect to the parameter t: at $t = 0$ it is

$$\nabla_t s|_{t=0} := \left.\frac{d}{dt}(P_\gamma^t)^{-1}(s(t))\right|_{t=0} \in E_{\gamma(0)},$$

and one can similarly define $\nabla_t s(t_0) \in E_{\gamma(t_0)}$ for arbitrary t_0 in the domain of γ by repraparametrizing the path to $t \mapsto \gamma(t_0 + t)$, so that it passes through $\gamma(t_0)$ at time $t = 0$. The section s is then parallel along γ if and only if $\nabla_t s \equiv 0$. A slightly subtle distinction between this and the usual covariant derivative of a section of E is that s depends directly on t, not just on $\gamma(t) \in M$, so $\nabla_t s$ can be nonzero even if the path $\gamma(t)$ is stationary. For example, if γ is a constant path at a point $p \in M$, then parallel transport P_γ^t defines the identity map $T_pM \to T_pM$ for every t, and $\nabla_t s \in T_pM$ is then just the ordinary derivative of the path $s(t)$ in the vector space T_pM. On the other hand, if $\dot\gamma(0) \neq 0$, then γ is an embedding near $t = 0$ and thus traces out a smooth 1-dimensional submanifold of M; restricting γ to a suitably small neighborhood of 0, it is then easy to see that any section s along γ can be "extended" to a global section $\hat{s} \in \Gamma(E)$ such that

$$\hat{s}(\gamma(t)) = s(t) \qquad \text{and} \qquad \nabla_{\dot\gamma(t)}\hat{s} = \nabla_t s(t) \qquad \text{for all } t.$$

(Indeed, first write down the extension $\hat{s}$ on a neighborhood of $\gamma(0)$ in a slice chart for the image of γ, then extend it arbitrarily to the rest of M.) In this situation, various useful things we've proven about $\nabla\hat{s}$ will apply to $\nabla_t s$ as well: one is the formula $\nabla_X \hat{s} = K(T\hat{s}(X))$, which becomes

$$\nabla_t s(t) = K(\dot{s}(t)), \tag{20.9}$$

where we are regarding $s(t) = \hat{s}(\gamma(t))$ as a smooth path in the total space E, whose derivative is thus $\dot{s}(t) = T\hat{s}(\dot\gamma(t))$. Another is the coordinate formula (20.7) for $\nabla_X\hat{s}$ in terms of a connection 1-form: assuming $\gamma(t)$ lies in the domain $\mathcal{U}_\alpha$ of a trivialization $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^m$ and writing $\Phi_\alpha(s(t)) = (\gamma(t), s_\alpha(t))$, we obtain

$$(\nabla_t s)_\alpha(t) = \dot{s}_\alpha(t) + A_\alpha(\dot\gamma(t))s_\alpha(t) \in \mathbb{F}^m. \tag{20.10}$$

An easy continuity argument now shows that (20.9) and (20.10) are not only valid under the condition $\dot\gamma \neq 0$: they are valid for *all* smooth paths γ, since a path with $\dot\gamma(t_0) = 0$ at some point t_0 admits arbitrarily small perturbations to one with $\dot\gamma(t_0) \neq 0$, and the section s can be perturbed along with it.

Proof of Theorem 20.12. Suppose $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^m$ is a local trivialization and γ is a path in $\mathcal{U}_\alpha$ with $p := \gamma(0)$ and $X := \dot\gamma(0)$. Since parallel transport maps are linear, there exists a unique function $g(t) \in \mathrm{GL}(m, \mathbb{F})$ with $g(0) = \mathbb{1}$ such that for any parallel section $s(t) \in E_{\gamma(t)}$ along γ, the local representative $s_\alpha(t) \in \mathbb{F}^m$ defined by $\Phi_\alpha(s(t)) = (\gamma(t), s_\alpha(t))$ satisfies

$$s_\alpha(t) = g(t)s_\alpha(0).$$

By (20.10), $\nabla_t s \equiv 0$ implies that $g(t)$ is the unique solution with $g(0) = \mathbb{1}$ to the linear ODE

$$\dot{g}(t) = -A_\alpha(\dot\gamma(t))g(t).$$

It suffices then to show that g takes values in the subgroup G (implying ∇ is G-compatible over $\mathcal{U}_\alpha$) if and only if A_α takes values in its Lie algebra $\mathfrak{g}$. Assuming the former, we can plug $t = 0$ into the above equation to conclude $A_\alpha(X) = -\dot{g}(0) \in \mathfrak{g}$, which completes the proof that $A_\alpha \in \Omega^1(\mathcal{U}_\alpha, \mathfrak{g})$ since $p \in \mathcal{U}_\alpha$ and $X \in T_pM$ were chosen arbitrarily. The converse follows from Exercise 20.14 below. $\square$

Exercise 20.13. A smooth **time-dependent vector field** on a manifold M is a family of vector fields $\{X_t \in \mathfrak{X}(M)\}_{t\in I}$ parametrized by an interval $I \subset \mathbb{R}$ such that the map $I \times M \to TM : (t,p) \mapsto X_t(p)$ is smooth. A path $\gamma(t) \in M$ is called an **orbit** or **flow line** of the time-dependent vector field $\{X_t\}_{t\in I}$ if it satisfies $\dot\gamma(t) = X_t(\gamma(t))$ for every t. One can develop the theory of flows for time-dependent vector fields analogously to the time-independent case, defining in particular a smooth map φ_X^t on suitable open subsets of M such that $\gamma(t) := \varphi_X^t(p)$ is the unique orbit of $\{X_t\}_{t\in I}$ satisfying $\gamma(0) = p$. If you prefer not to redo work that has already been done, you can instead do this:

(a) Given a time-dependent vector field $\{X_t\}_{t\in I}$ on M, define a time-independent vector field $Y \in \mathfrak{X}(I \times M)$ by

$$Y(t,p) := (1, X_t(p)) \in \mathbb{R} \times T_pM = T_tI \times T_pM = T_{(t,p)}(I \times Y).$$

Use the flow of Y to deduce everything you might possibly want to know about the flow of $\{X_t\}$, e.g. that φ_X^t exists and is unique and is a diffeomorphism $M \to M$ for every $t \in \mathbb{R}$ if M is compact.

Caution: Do not try to prove the relations $\varphi_X^{s+t} = \varphi_X^s \circ \varphi_X^t$ or $\varphi_X^{-t} = (\varphi_X^t)^{-1}$, which are valid in general only for time-independent vector fields.

With these basics understood, the following observation will be helpful for Exercise 20.14 below:

(b) Suppose $N \subset M$ is a smooth submanifold and $\{X_t\}_{t\in I}$ is a time-dependent vector field on M such that $X_t(p) \in T_pN$ for every $p \in N$ and $t \in I$. Show that every flow line of $\{X_t\}$ is either contained in N or disjoint from it.

Exercise 20.14. For any Lie subgroup $G \subset \mathrm{GL}(m,\mathbb{F})$ and a smooth path of matrices $\mathbf{A}(t) \in \mathfrak{g} = T_{\mathbb{1}}G$, show that the unique solution $\mathbf{\Phi}(t) \in \mathbb{F}^{m\times m}$ to the initial value problem

$$\begin{cases} \dot{\mathbf{\Phi}}(t) &= \mathbf{A}(t)\mathbf{\Phi}(t), \\ \mathbf{\Phi}(0) &= \mathbb{1} \end{cases}$$

satisfies $\mathbf{\Phi}(t) \in G$ for all t.

Hint: Show that for any $\mathbf{A} \in \mathfrak{g}$, $X(\mathbf{B}) := \mathbf{AB} \in \mathbb{F}^{m\times m} = T_{\mathbf{B}}\,\mathrm{GL}(m,\mathbb{F})$ defines a smooth vector field on $\mathrm{GL}(m,\mathbb{F})$ that satisfies $X(\mathbf{B}) \in T_{\mathbf{B}}G$ for all $\mathbf{B} \in G$.

Remark 20.15. The notion of a G-structure on a vector bundle makes sense for any subgroup $G \subset \mathrm{GL}(m,\mathbb{F})$, i.e. the definition itself does not require the additional condition that $G \subset \mathrm{GL}(m,\mathbb{F})$ is a smooth submanifold. However, by applying Exercise 20.14, Theorem 20.12 makes crucial use of this assumption, along with the fact (used in Exercise 20.14) that the matrix multiplication map $\mathbb{F}^{m\times m} \times \mathbb{F}^{m\times m} \to \mathbb{F}^{m\times m} : (\mathbf{A},\mathbf{B}) \mapsto \mathbf{AB}$ is smooth. In other words, while G-structures on bundles can be defined for arbitrary subgroups $G \subset \mathrm{GL}(m,\mathbb{F})$, making connections compatible with these structures requires the group G to be smooth.

21. Constructions of connections

21.1. A general existence result. Let's get this out of the way first:

Theorem 21.1. *Every vector bundle $E \to M$ with a G-structure for some Lie subgroup $G \subset \mathrm{GL}(m,\mathbb{F})$ admits a G-compatible connection.*

Proof. Choose an open covering $\{\mathcal{U}_\alpha\}_{\alpha\in I}$ of M with a subordinate partition of unity $\{\varphi_\alpha : M \to [0,1]\}_{\alpha\in I}$ such that there are also local G-compatible trivializations $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^m$ for each $\alpha \in I$. Each of these determines a trivial connection D^α on $E|_{\mathcal{U}_\alpha}$, which is G-compatible since its parallel transport maps look like the identity $\mathbb{F}^m \to \mathbb{F}^m$ in the trivialization Φ_α. We can then define a global connection ∇ on E by

$$\nabla_X s := \sum_{\alpha\in I} \varphi_\alpha(p) D^\alpha_X s, \qquad \text{for } p \in M,\ X \in T_pM,$$

where it is understood that at each point p, the sum contains only the finitely many terms for which $p \in \mathrm{supp}(\varphi_\alpha) \subset \mathcal{U}_\alpha$, and D^α is thus well defined near p. That the resulting operator

$\nabla : \Gamma(E) \to \Gamma(\mathrm{Hom}(TM, E))$ is a G-compatible connection follows now by writing it down in local trivializations. Indeed, each $p \in M$ has a neighborhood $\mathcal{U} \subset M$ that is contained in all of the sets $\mathcal{U}_\alpha$ for which $p \in \mathrm{supp}(\varphi_\alpha)$, and choosing any G-compatible trivialization of $E|_{\mathcal{U}}$ identifies each of the relevant operators D^α with an operator of the form $d^\alpha_X f := df(X) + B_\alpha(X)f$ on functions $f \in C^\infty(\mathcal{U}, \mathbb{F}^m)$, where $B_\alpha \in \Omega^1(\mathcal{U}, \mathfrak{g})$ since D^α is G-compatible. In this same trivialization, ∇ then becomes

$$d_X f = \sum_\alpha [\varphi_\alpha \, df(X) + \varphi_\alpha B_\alpha(X) f] = df(X) + B(X)f, \qquad \text{where} \qquad B := \sum_\alpha \varphi_\alpha B_\alpha \in \Omega(\mathcal{U}, \mathfrak{g}),$$

and is thus G-compatible by Theorem 20.12. □

It is important to understand that Theorem 21.1 says nothing about uniqueness, and indeed, connections are in general neither unique nor canonical: in the case $G = \mathrm{GL}(m, \mathbb{F})$ for instance, one can produce an infinite-dimensional family of connections by choosing any specific connection ∇ and defining other connections by $\nabla + A$ for arbitrary bundle maps $A : E \to \mathrm{Hom}(TM, E)$. (Something similar is true for G-compatible connections with arbitrary Lie subgroups $G \subset \mathrm{GL}(m, \mathbb{F})$ if one considers only bundle maps $E \to \mathrm{Hom}(TM, E)$ that preserve the relevant structure—the space of such bundle maps is typically still infinite dimensional.) The major exception is the tangent bundle $TM \to M$ of a Riemannian or pseudo-Riemannian manifold: this bundle has structure group $\mathrm{O}(k, \ell)$ determined by the signature (k, ℓ) of its bundle metric, and while there is an infinite-dimensional family of $\mathrm{O}(k, \ell)$-compatible connections on TM, we will see in the next lecture that a canonical one can be singled out, due to the fact that $TM \to M$ is not just any vector bundle but specifically a *tangent* bundle.

21.2. Pullbacks. The next two sections will be concerned with the following question: given a finite collection of vector bundles $E^1, \ldots, E^m$ with connections and a natural operation that produces a new bundle E out of $E^1, \ldots, E^m$, how do the connections on $E^1, \ldots, E^m$ determine one on E? It will usually be obvious how the connection on E should be defined—only a little bit of effort is then required to check that the result really is a connection.

We start with pullbacks: suppose $E \to M$ is a smooth vector bundle, N is a manifold and $f : N \to M$ is a smooth map. A section

$$s \in \Gamma(f^*E)$$

of the pullback bundle $f^*E \to N$ associates to each $p \in N$ a vector $s(p) \in E_{f(p)}$, and is therefore sometimes called a **section of** E **along** f. If a connection ∇ on $E \to M$ with parallel transport maps $P^t_\gamma : E_{\gamma(0)} \to E_{\gamma(t)}$ is given, then there is an obvious way to define parallel transport maps $P^t_\gamma : (f^*E)_{\gamma(0)} \to (f^*E)_{\gamma(t)}$ for $f^*E \to N$ along a path γ in N, namely

$$P^t_\gamma := P^t_{f \circ \gamma} : (f^*E)_{\gamma(0)} = E_{f(\gamma(0))} \to E_{f(\gamma(t))} = (f^*E)_{\gamma(t)}. \tag{21.1}$$

To see that this really defines a connection on $f^*E \to N$, let us translate (21.1) into a definition of a horizontal subbundle. Confusion can sometimes arise from the fact that fibers of f^*E are also fibers of E, so it will be helpful to distinguish them by adopting the following slightly verbose notation: elements of E can be written as pairs

$$(p, v) \in E, \qquad \text{for } p \in M, \, v \in E_p,$$

while elements of f^*E are written similarly as

$$(p, v) \in f^*E, \qquad \text{for } p \in N, \, v \in E_{f(p)} = (f^*E)_p.$$

The canonical smooth linear bundle map $f^*E \to E$ covering $f : N \to M$ then takes the form

$$\Psi : f^*E \to E : (p, v) \mapsto (f(p), v).$$

Equation (21.1) can now be interpreted as saying that a section $s(t) \in (f^*E)_{\gamma(t)}$ of f^*E along a path γ in N is parallel if and only if the section $\Psi \circ s(t) \in E_{f(\gamma(t))}$ of E along $f \circ \gamma$ is parallel. Differentiating this relation with respect to t gives a corresponding relation between horizontal subbundles: $\dot{s}(t) \in T_{s(t)}(f^*E)$ should be horizontal if and only if $\partial_t(\Psi \circ s)(t) = T\Psi(\dot{s}(t)) \in T_{\Psi(s(t))}E$ is horizontal, so that $H(f^*E) \subset T(f^*E)$ must be defined by

$$H_{(p,v)}(f^*E) := (T\Psi)^{-1}\left(H_{(f(p),v)}E\right) \subset T_{(p,v)}(f^*E).$$

To see that this really does define a complement to the vertical subbundle $V_{(p,v)}(f^*E)$, notice that since Ψ defines isomorphisms between fibers of f^*E and fibers of E, its derivative $T\Psi$ defines isomorphisms between the corresponding vertical subspaces. The condition $T(f^*E) = V(f^*E) \oplus H(f^*E)$ then follows from $TE = VE \oplus HE$ via a simple linear-algebraic exercise:

EXERCISE 21.2. Suppose X, X' are vector spaces, $V \subset X$ and $V', H' \subset X'$ are linear subspaces such that $X' = V' \oplus H'$, and $A : X \to X'$ is a linear map that restricts to $V \subset X$ as an isomorphism onto V'. Show that the subspace $H := A^{-1}(H') \subset X$ is then complementary to V, i.e. $X = V \oplus H$.

Having shown that there is a well-defined horizontal subbundle $H(f^*E) \subset T(f^*E)$ corresponding to the parallel transport maps in (21.1), it follows from Lemma 19.3 that $H(f^*E)$ is a connection on $f^*E \to N$ in the sense of Definition 19.4, as the parallel transport maps are manifestly linear. It is also clear from this definition that if the connection on E is compatible with some structure group G on E, then the pullback connection is compatible with the induced G-structure on f^*E.

EXAMPLE 21.3. For a smooth path $\gamma : I \to M$ defined on an open interval $I \subset \mathbb{R}$, a section s of $E \to M$ along γ is the same thing as a section of the pullback bundle $\gamma^*E \to I$. Let ∂_t denote the standard basis vector on $T_tI = \mathbb{R}$ for each $t \in I$. Now if ∇ is a connection on $E \to M$ and we equip $\gamma^*E \to I$ with the resulting pullback connection, writing the covariant derivative of s in terms of parallel transport gives

$$\nabla_{\partial_t} s\big|_{t=0} = \frac{d}{dt}(P_\gamma^t)^{-1}\left(s(t)\right)\bigg|_{t=0}.$$

This is of course exactly the same thing as what we have previously denoted by $\nabla_t s(0)$; in other words, the covariant derivative with respect to t of a section of E along a path $\gamma(t) \in M$ is the same thing as the covariant derivative (using the pullback connection) of the corresponding section of γ^*E in the direction of the canonical unit vector field on the interval. This should not be surprising—if it had not been true, we would have concluded that we have the wrong definition of the pullback connection and then searched for a different one.

EXAMPLE 21.4. The following generalization of Example 21.3 is sometimes useful for computations: consider an open subset $\mathcal{V} \subset \mathbb{R}^d$ and a smooth map $f : \mathcal{V} \to M$. A section $s \in \Gamma(f^*E)$ of $E \to M$ along f then assigns to each tuple $(t^1, \dots, t^d) \in \mathcal{V}$ a vector $s(t^1, \dots, t^d) \in E_{f(t^1,\dots,t^d)}$, and we can define a covariant derivative with respect to each of the variables $t^1, \dots, t^d$,

$$\nabla_i s(t^1, \dots, t^d) \in E_{f(t^1,\dots,t^d)}, \qquad i = 1, \dots, d,$$

which literally means the covariant derivative (via the pullback connection) of $s \in \Gamma(f^*E)$ with respect to the standard basis vector $\partial_i \in T_{(t^1,\dots,t^d)}\mathcal{V} = \mathbb{R}^d$. This makes $\nabla_i s$ another section of f^*E, so it can be differentiated again, defining iterated covariant derivatives $\nabla_i \nabla_j s$, $\nabla_i \nabla_j \nabla_k s$ and so forth. For example, the partial derivatives of $f : \mathcal{V} \to M$ as defined in §4.1 are vector fields along f,

$$\partial_i f \in \Gamma(f^*TM), \qquad i = 1, \dots, d,$$

so if a connection on $TM \to M$ has been chosen, we can now use the pullback connection on $f^*TM \to \mathcal{V}$ to define higher (covariant) derivatives of f in the form $\nabla_j \partial_i f$, $\nabla_k \nabla_j \partial_i f$ and so forth.

Let us derive a local coordinate formula for $\nabla_t s(t)$ when $s(t) \in E_{\gamma(t)}$ is a section along a path γ in M. Assume the image of γ lies in an open subset $\mathcal{U}_\alpha \subset M$ for which there is a local trivialization $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^m$, and let $e_1, \ldots, e_m$ denote the corresponding frame for E over $\mathcal{U}_\alpha$. Assume also that $\mathcal{U}_\alpha \subset M$ admits coordinates $x^1, \ldots, x^n$, so that the Christoffel symbols Γ^a_{ib} characterizing ∇ on $E|_{\mathcal{U}_\alpha}$ are defined. Writing $s(t) = s^a(t) e_a(\gamma(t)) \in E_{\gamma(t)}$ and $\dot\gamma(t) = \dot\gamma^i(t)\,\partial_i \in T_{\gamma(t)}M$, we claim

$$(\nabla_t s)^a(t) = \dot s^a(t) + \Gamma^a_{ib}(\gamma(t))\dot\gamma^i(t) s^b(t).$$

or more succinctly,

$$(\nabla_t s)^a = \dot s^a + (\Gamma^a_{ib} \circ \gamma)\dot\gamma^i s^b. \tag{21.2}$$

It suffices to prove that this holds at $t = 0$, since the parametrization of γ can always be shifted, and in fact, we are also free to assume $\dot\gamma(0) \neq 0$, for the following reason. Unless $\dim M = 0$ (in which case the statement is trivial and there is nothing to prove), any path γ in M can be perturbed if necessary to ensure $\dot\gamma(0) \neq 0$, and the section s can be perturbed with it to a section along the perturbation of γ. If the relation (21.2) is satisfied after this perturbation, then it must have been satisfied beforehand as well, simply because both sides are continuous with respect to C^1-small perturbations of γ and s. With this understood, the condition $\dot\gamma(0) \neq 0$ allows us to assume after restricting to a suitably small neighborhood of 0 that γ is an embedding, so its image is a smooth 1-dimensional submanifold of M. One can then use a slice chart on M for this submanifold in order to construct a section $\eta \in \Gamma(E)$ such that $s(t) = \eta(\gamma(t))$ for all t, and it then follows from the definition of the covariant derivative via parallel transport that $\nabla_t s(0) = \nabla_{\dot\gamma(0)}\eta$. Writing $\dot\gamma = \dot\gamma^i \partial_i$ and $\nabla_{\dot\gamma}\eta = \dot\gamma^i \nabla_i \eta$, (20.6) now implies

$$(\nabla_t s(0))^a = \dot\gamma^i(0)\left[\partial_i \eta^a(\gamma(0)) + \Gamma^a_{ib}(\gamma(0))\eta^b(\gamma(0))\right] = \partial_t(\eta^a \circ \gamma)(0) + \Gamma^a_{ib}(\gamma(0))(\eta^b \circ \gamma)(0),$$

which justifies (21.2).

Recall from Exercise 20.8 that the connection 1-form $A_\alpha \in \Omega^1(\mathcal{U}_\alpha, \mathbb{F}^{m\times m})$ can be derived from the Christoffel symbols by $A_\alpha(\partial_i)^a{}_b = \Gamma^a_{ib}$, so for $X = X^i \partial_i$, $A_\alpha(X)^a{}_b = \Gamma^a_{ib} X^i$. The expression $\Gamma^a_{ib}(\gamma(t))\dot\gamma(t)$ in (21.2) can therefore be reinterpreted as $A_\alpha(\dot\gamma(t))^a{}_b$, and the formula thus reproduces (20.10), i.e.

$$(\nabla_t s)_\alpha = \dot s_\alpha + A_\alpha(\dot\gamma) s_\alpha.$$

In the general situation where N is an arbitrary manifold with a smooth map $f : N \to M$ and $s \in \Gamma(f^*E)$, we can compute the covariant derivative $\nabla_X s$ in any direction $X \in TN$ by choosing a path γ in N with $\dot\gamma(0) = X$ and computing $\nabla_t(s \circ \gamma)(0)$, i.e. $\nabla_X s$ is the covariant derivative at $t = 0$ of $s \circ \gamma(t) \in (f^*E)_{\gamma(t)} = E_{f\circ\gamma(t)}$, which is a section of E along the path $f \circ \gamma$. Writing $s(p) = s^a(p) e_a(f(p))$ for $p \in f^{-1}(\mathcal{U}_\alpha)$, (21.2) thus implies

$$(\nabla_X s)^a = ds^a(X) + \Gamma^a_{ib}(f(p))(f_*X)^i s^b(p), \qquad \text{for } p \in f^{-1}(\mathcal{U}_\alpha),\ X \in T_pM. \tag{21.3}$$

To rewrite this in terms of a connection 1-form, observe that the frame $e_1, \ldots, e_m$ corresponding to our trivialization Φ_α on $\mathcal{U}_\alpha$ determines a local frame for f^*E over the open set $f^{-1}(\mathcal{U}_\alpha) \subset N$, consisting of the sections $e_1 \circ f, \ldots, e_m \circ f$, and the local trivialization of f^*E corresponding to this is the one that we called

$$f^*\Phi_\alpha : (f^*E)|_{f^{-1}(\mathcal{U}_\alpha)} \to f^{-1}(\mathcal{U}_\alpha) \times \mathbb{F}^m$$

in §17.2. For $s \in \Gamma(f^*E)$, let $s_\alpha : f^{-1}(\mathcal{U}_\alpha) \to \mathbb{F}^m$ denote the local representation of s in this trivialization; then (21.3) becomes

$$(\nabla_X s)_\alpha = ds_\alpha(X) + A_\alpha(f_*X) s_\alpha(p), \qquad \text{for } p \in f^{-1}(\mathcal{U}_\alpha),\ X \in T_pM.$$

In other words, the connection 1-form for the pullback connection on f^*E with respect to the trivialization $f^*\Phi_\alpha$ is exactly what one would hope for: it is the pullback of A_α,

$$f^*A_\alpha \in \Omega^1(f^{-1}(\mathcal{U}_\alpha), \mathbb{F}^{m\times m}).$$

EXERCISE 21.5. Any section $s \in \Gamma(E)$ gives rise to a section along $f : N \to M$ in the form $s \circ f \in \Gamma(f^*E)$. Prove

$$\nabla_X(s \circ f) = \nabla_{f_*X} s \qquad \text{for all } X \in TN.$$

21.3. Algebraic operations. We shall now run through the essential items on the list of algebraic constructions of vector bundles in §17.4, and outline how to construct connections on each of them. Assume throughout that $E, F \to M$ are fixed vector bundles on which connections (both labelled ∇) have already been chosen.

21.3.1. *Direct sums.* The natural way to define parallel transport for $E \oplus F$ out of the parallel transport on E and F along a path γ is

$$P^t_\gamma(v, w) := (P^t_\gamma(v), P^t_\gamma(w)) \in E_{\gamma(t)} \times F_{\gamma(t)}, \qquad \text{for} \qquad (v, w) \in E_{\gamma(0)} \times F_{\gamma(0)}.$$

The notion of covariant differentiation on $\Gamma(E\oplus F)$ that arises from this definition is quite straightforward: under the obvious identification of $\Gamma(E \oplus F)$ with $\Gamma(E) \times \Gamma(F)$, we have

$$\nabla_X(\eta, \xi) = (\nabla_X \eta, \nabla_X \xi).$$

It is trivial to check that $\nabla : \Gamma(E \oplus F) \to \Gamma(\mathrm{Hom}(TM, E \oplus F))$ by this definition satisfies the required Leibniz rule and thus defines a connection on $E \oplus F$.

21.3.2. *The dual bundle.* The isomorphisms $P^t_\gamma : E_{\gamma(0)} \to E_{\gamma(t)}$ determine isomorphisms $P^t_\gamma : E^*_{\gamma(0)} \to E^*_{\gamma(t)}$ by dualization, i.e. for $\lambda \in E^*_{\gamma(0)}$ and $v \in E_{\gamma(t)}$, we define

$$P^t_\gamma(\lambda)v := \lambda\left((P^t_\gamma)^{-1}v\right).$$

Equivalently, this means that if $\lambda(t) \in E^*_{\gamma(t)}$ and $v(t) \in E_{\gamma(t)}$ are parallel sections along γ, then the natural pairing between them is constant, so

$$P^t_\gamma(\lambda)\left(P^t_\gamma(v)\right) = \lambda(v) \qquad \text{for all } t. \tag{21.4}$$

It follows for instance that if $e_1, \ldots, e_m$ is a frame for E near $\gamma(0)$ consisting of sections that are parallel along γ, then the sections in the dual frame $e^1_*, \ldots, e^m_*$ are also parallel along γ. From (21.4), one deduces that the covariant derivative satisfies a Leibniz rule for the pairing of E^* and E: for any sections $\lambda(t) \in E^*_{\gamma(t)}$ and $v(t) \in E_{\gamma(t)}$ along γ, we have

$$\left.\frac{d}{dt}\left[\lambda(t)\left(v(t)\right)\right]\right|_{t=0} = \left.\frac{d}{dt}(P^t_\gamma)^{-1}(\lambda(t))\left((P^t_\gamma)^{-1}(v(t))\right)\right|_{t=0} = \nabla_t\lambda(0)\left(v(0)\right) + \lambda(0)\left(\nabla_t v(0)\right),$$

which implies a statement about directional derivatives of the function $\lambda(v) \in C^\infty(M, \mathbb{F})$ for $\lambda \in \Gamma(E^*)$ and $v \in \Gamma(E)$ with respect to a vector field $X \in \mathfrak{X}(M)$, namely

$$\mathcal{L}_X\left[\lambda(v)\right] = \left(\nabla_X\lambda\right)(v) + \lambda\left(\nabla_X v\right). \tag{21.5}$$

This relation uniquely characterizes the operator $\nabla : \Gamma(E^*) \to \Gamma(\mathrm{Hom}(TM, E^*))$, and can thus be used to give an easy proof that it really does define a connection:

EXERCISE 21.6. Deduce from (21.5) that the operator $\nabla : \Gamma(E^*) \to \Gamma(\mathrm{Hom}(TM, E^*))$ satisfies the Leibniz rule required by Definition 20.4 and thus determines a connection on $E^* \to M$.

EXERCISE 21.7. In terms of the Christoffel symbols $\Gamma^a_{ib} = (\nabla_i e_b)^a$ defined with respect to a coordinate chart $(x^1, \ldots, x^n)$ and frame $e_1, \ldots, e_m$ for E over an open set $\mathcal{U} \subset M$, show that the induced connection on E^* acts on the dual frame $e^1_*, \ldots, e^m_*$ by

$$(\nabla_i e^b_*)_a = -\Gamma^b_{ia},$$

and deduce the general coordinate formula

$$(\nabla_i \lambda)_a = \partial_i \lambda_a - \Gamma^b_{ia} \lambda_b$$

for $\lambda \in \Gamma(E^*)$.

A few extra comments about the special case $E = TM$ are in order. Here a chart $(x^1, \ldots, x^n) : \mathcal{U} \to \mathbb{R}^n$ also defines a natural frame over $\mathcal{U}$, consisting of the coordinate vector fields $\partial_1, \ldots, \partial_n$, and the resulting Christoffel symbols consist of n^3 real-valued functions

$$\Gamma^i_{jk} : \mathcal{U} \to \mathbb{R}, \qquad i, j, k \in \{1, \ldots, n\}$$

given by

$$\Gamma^i_{jk} = (\nabla_j \partial_k)^i = dx^i (\nabla_j \partial_k).$$

In local coordinates, the covariant derivative of a vector field $X = X^i \partial_i$ is thus given (cf. Equation (20.6)) by

$$(\nabla_j X)^i = \partial_j X^i + \Gamma^i_{jk} X^k. \tag{21.6}$$

For the induced connection on the dual bundle T^*M, we observe that the covariant derivative $\nabla\lambda$ of a 1-form $\lambda \in \Gamma(T^*M) = \Omega^1(M)$ is a section of $\mathrm{Hom}(TM, T^*M)$ and can thus be identified in a natural way with a type $(0, 2)$ tensor field $\nabla\lambda \in \Gamma(T^0_2 M)$, i.e. we define

$$(\nabla\lambda)(X, Y) := (\nabla_X \lambda)(Y).$$

The components of $\nabla\lambda$ in local coordinates thus take the form $(\nabla\lambda)_{ij} = (\nabla\lambda)(\partial_i, \partial_j) := (\nabla_i \lambda)(\partial_j) =: (\nabla_i \lambda)_j$, and Exercise 21.7 gives

$$(\nabla\lambda)_{ij} = \partial_i \lambda_j - \Gamma^k_{ij} \lambda_k. \tag{21.7}$$

The tensor $\nabla\lambda \in \Gamma(T^0_2 M)$ is our newest and best answer to the question first posed in Lecture 8 concerning how one should go about defining the "derivative" of a tensor field, in this case specifically a 1-form. One of the answers we came up with in Lecture 8 was the exterior derivative $d\lambda \in \Omega^2(M)$, which is also a type $(0, 2)$ tensor, but it carries less information: if you compare the local coordinate formulas we have for $\nabla\lambda$ and $d\lambda$, you'll notice that the individual partial derivatives $\partial_i \lambda_j$ cannot all be derived from $d\lambda$, but from $\nabla\lambda$ they can. In that sense, $\nabla\lambda$ is a better way of defining the derivative of λ, but it has the comparative disadvantage that it depends on a choice, since connections can always be chosen but are not unique.

21.3.3. *Tensor bundles.* If $A : V \to V'$ and $B : W \to W'$ are linear maps, there is a unique linear map

$$A \otimes B : V \otimes W \to V' \otimes W'$$

defined via the condition $(A \otimes B)(v \otimes w) = Av \otimes Bw$ for all $v \in V$ and $w \in W$. This determines a natural definition for parallel transport maps $P^t_\gamma : E_{\gamma(0)} \otimes F_{\gamma(0)} \to E_{\gamma(t)} \otimes F_{\gamma(t)}$, via the condition

$$P^t_\gamma(\eta \otimes \xi) := P^t_\gamma(\eta) \otimes P^t_\gamma(\xi) \qquad \text{for all } \eta \in E_{\gamma(0)},\ \xi \in F_{\gamma(0)}.$$

In particular, the pointwise tensor product of any parallel sections of E and F along γ then becomes a parallel section of $E \otimes F$. As in §21.3.2, this gives rise to a Leibniz rule for the covariant derivative:

$$\nabla_t (\eta(t) \otimes \xi(t))|_{t=0} = \left.\frac{d}{dt} (P^t_\gamma)^{-1}(\eta(t)) \otimes (P^t_\gamma)^{-1}(\xi(t))\right|_{t=0} = \nabla_t \eta(0) \otimes \xi(0) + \eta(0) \otimes \nabla_t \xi(0),$$

implying that for any $\eta \in \Gamma(E)$, $\xi \in \Gamma(F)$ and $X \in \mathfrak{X}(M)$,

$$\nabla_X(\eta \otimes \xi) = \nabla_X \eta \otimes \xi + \eta \otimes \nabla_X \xi. \tag{21.8}$$

Once again, this uniquely characterizes the covariant derivative and can be used to prove that what we have defined really is a connection on $E \otimes F$:

EXERCISE 21.8. Deduce from (21.8) that the operator $\nabla : \Gamma(E \otimes F) \to \Gamma(\mathrm{Hom}(TM, E \otimes F))$ defined above is a connection on $E \otimes F$.

By finite iterations, one can extract from the constructions in this section and §21.3.2 a definition of a connection on any of the tensor bundles $E^k_\ell \cong E^{\otimes k} \otimes (E^*)^{\otimes \ell}$ that is uniquely determined by any choice of connection on E. Moreover, it is uniquely characterized by the property that all Leibniz rules one can reasonably think of to write down are satisfied. For example, the induced connection on E^1_2 is related to the chosen connection on E and the induced connection on E^* by

$$\mathcal{L}_X \left(S(\lambda, \eta, \xi) \right) = (\nabla_X S)(\lambda, \eta, \xi) + S(\nabla_X \lambda, \eta, \xi) + S(\lambda, \nabla_X \eta, \xi) + S(\lambda, \eta, \nabla_X \xi)$$

for all $S \in \Gamma(E^1_2)$, $\lambda \in \Gamma(E^*)$ and $\eta, \xi \in \Gamma(E)$, and this relation uniquely determines ∇S.

In the case $E = TM$, the covariant derivative of a type (k, ℓ) tensor field $S \in \Gamma(T^k_\ell M)$ can be understood as a type $(k, \ell + 1)$ tensor field $\nabla S \in \Gamma(T^k_{\ell+1} M)$ by defining

$$(\nabla S)(\lambda^1, \ldots, \lambda^k, X_0, \ldots, X_\ell) := (\nabla_{X_0} S)(\lambda^1, \ldots, \lambda^k, X_1, \ldots, X_\ell).$$

In local coordinates, the components of ∇S thus take the form

$$(\nabla S)^{i_1 \ldots i_k}{}_{j_0 \ldots j_\ell} = (\nabla_{j_0} S)^{i_1 \ldots i_k}{}_{j_1 \ldots j_\ell} = (\nabla_{\partial_{j_0}} S) \left(dx^{i_1}, \ldots, dx^{i_k}, \partial_{j_1}, \ldots, \partial_{j_\ell} \right).$$

EXERCISE 21.9. For a connection on TM with Christoffel symbols Γ^i_{jk} in some choice of local coordinates, show that the induced connection on $T^k_\ell M$ is given in the same coordinates by

$$\begin{aligned}(\nabla S)^{i_1 \ldots i_k}{}_{j_0 \ldots j_\ell} = \partial_{j_0} S^{i_1 \ldots i_k}{}_{j_1 \ldots j_\ell} &+ \Gamma^{i_1}_{j_0 a} S^{a i_2 \ldots i_k}{}_{j_1 \ldots j_\ell} + \ldots + \Gamma^{i_k}_{j_0 a} S^{i_1 \ldots i_{k-1} a}{}_{j_1 \ldots j_\ell} \\ &- \Gamma^a_{j_0 j_1} S^{i_1 \ldots i_k}{}_{a j_2 \ldots j_\ell} - \ldots - \Gamma^a_{j_0 j_\ell} S^{i_1 \ldots i_k}{}_{j_1 \ldots j_{\ell-1} a}\end{aligned}$$

for $S \in \Gamma(T^k_\ell S)$. Notice that this formula generalizes both (21.6) and (21.7).

21.3.4. *Bundles of linear maps.* Since $\mathrm{Hom}(E, F)$ is canonically isomorphic to $E^* \otimes F$, the constructions in §21.3.2 and §21.3.3 determine a natural connection on $\mathrm{Hom}(E, F)$.

EXERCISE 21.10. Show that the connection on $\mathrm{Hom}(E, F)$ is uniquely determined from the connections on E and F via the Leibniz rule

$$\nabla_X (A\eta) = (\nabla_X A)(\eta) + A(\nabla_X \eta) \qquad \text{for all } A \in \Gamma(\mathrm{Hom}(E, F)),\ \eta \in \Gamma(E) \text{ and } X \in \mathfrak{X}(M).$$

Hint: It suffices to consider bundle maps $A : E \to F$ of the form $A\eta = \lambda(\eta)\xi$ for fixed sections $\lambda \in \Gamma(E^*)$ and $\xi \in \Gamma(F)$. (Why?)

21.4. Tangent bundles, torsion and symmetry. For the rest of this lecture we specialize to the rank n real vector bundle

$$TM \to M$$

over a smooth n-manifold M. A connection on $TM \to M$ is also often referred to as a **connection on the manifold** M: it defines in particular the notion of parallel vector fields. Using the constructions in §21.2 and §21.3, it also determines connections on all of the tensor bundles $T^k_\ell M \to M$ and the pullback $f^*TM \to N$ for any smooth map $f : N \to M$. We will always assume when a connection ∇ on TM has been specified that the bundles $T^k_\ell M$ and f^*TM are endowed with the connections determined by ∇ in this way.

For covariant derivatives of vector fields, a natural question arises that would not make sense on an arbitrary vector bundle. Suppose $\mathcal{V} \subset \mathbb{R}^d$ is an open set and $f : \mathcal{V} \to M$ is a smooth map as discussed in Example 21.4, so that we can define partial derivatives $\partial_j f \in \Gamma(f^*TM)$ and then covariantly differentiate to define second derivatives $\nabla_i \partial_j f \in \Gamma(f^*TM)$. Do mixed partial derivatives in this sense commute, i.e. we do we have

$$\nabla_i \partial_j f = \nabla_j \partial_i f \qquad \text{for all } i, j?$$

The question can easily be answered via a local coordinate computation: choose a chart $(\mathcal{U}, x)$ on M with coordinates $x = (x^1, \dots, x^n)$ and, on the subset in $\mathcal{V}$ where f has image in $\mathcal{U}$, write $f^k := x^k \circ f$ for each $k = 1, \dots, n$ so that $\partial_i f = \partial_i f^k \, \partial_k$. Applying (21.3) then gives

$$\begin{aligned}(\nabla_i \partial_j f - \nabla_j \partial_i f)^k &= \partial_i \partial_j f^k - \partial_j \partial_i f^k + \Gamma^k_{ab}(\partial_i f^a)(\partial_j f^b) - \Gamma^k_{ab}(\partial_j f^a)(\partial_i f^b) \\ &= (\Gamma^k_{ab} - \Gamma^k_{ba})(\partial_i f^a)(\partial_j f^b).\end{aligned}$$

The only way to make sure this vanishes for arbitrary maps $f : \mathcal{V} \to M$ is if the Christoffel symbols satisfy the relation

$$\Gamma^k_{ab} = \Gamma^k_{ba} \qquad \text{for all } k, a, b \in \{1, \dots, n\}.$$

There is no reason why an arbitrary connection on $TM \to M$ should satisfy this; in fact, on the domain of a single chart one can always define a connection whose Christoffel symbols are any desired set of n^3 functions, which need not be related to each other in any way. But differential geometers have a favorite trick for situations like this: when we see a quantity that doesn't always vanish even though we wish it would, we make it into a tensor.

EXERCISE 21.11. Given a connection ∇ on the manifold M, prove that the bilinear map $T : \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M)$ given by

$$T(X, Y) := \nabla_X Y - \nabla_Y X - [X, Y]$$

defines a type $(1, 2)$ tensor field on M, whose components in any local coordinate system are given by

$$T^i_{\ jk} = \Gamma^i_{jk} - \Gamma^i_{kj}.$$

The tensor $T \in \Gamma(T^1_2 M)$ in Exercise 21.11 is called the **torsion** of the connection ∇, and ∇ is called **symmetric** if its torsion tensor vanishes. Note that since the Christoffel symbols Γ^i_{jk} are not the components of any globally-defined tensor field, it is at first glance far from obvious that $\Gamma^i_{jk} - \Gamma^i_{kj}$ should be. One can check using the transformation formula in Exercise 20.7 that the functions $\Gamma^i_{jk} - \Gamma^i_{kj}$ do indeed transform as a tensor, but this is tedious; we should be very grateful in Exercise 21.11 that we can instead use C^∞-linearity to write down a coordinate-invariant definition of the torsion tensor.

As soon as one knows that connections on $TM \to M$ exist and have a well-defined torsion tensor, it is not hard to see that *symmetric* connections also exist:

EXERCISE 21.12. Given any connection ∇ on M with torsion tensor $T \in \Gamma(T^1_2 M)$, show that $\hat{\nabla}_X Y := \nabla_X Y - \frac{1}{2} T(X, Y)$ defines a *symmetric* connection on M.

A second proof that symmetric connections always exist will emerge in the next lecture when we discuss connections on Riemannian manifolds. Let us conclude for now by stating the most useful property of symmetric connections, which follows immediately from the calculations above:

PROPOSITION 21.13. *A connection ∇ on a manifold M is symmetric if and only if for every open set $\mathcal{V} \subset \mathbb{R}^d$ and smooth map $f : \mathcal{V} \to M$, the relation $\nabla_i \partial_j f \equiv \nabla_j \partial_i f$ holds for all $i, j \in \{1, \dots, d\}$.* □

REMARK 21.14. Symmetry of a connection does *not* imply that one can also exchange the order of the operators ∇_i and ∇_j in higher covariant derivatives. That is not true in general, and we will come back to this subject when we discuss curvature.

22. Pseudo-Riemannian manifolds and geodesics

22.1. Geodesics and the exponential map. We will be assuming in most of this lecture that M is a Riemannian or pseudo-Riemannian manifold, but the general definition of a geodesic does not actually require so much structure; it only requires a connection ∇ on M, by which we mean a connection on the tangent bundle $TM \to M$. The defining property of a straight line $\gamma : (a, b) \to \mathbb{R}^n$ in Euclidean space is that its velocity $\dot{\gamma}(t) \in \mathbb{R}^n$ is constant. The obvious analogue of this condition for a path $\gamma : (a, b) \to M$ is that its velocity $\dot{\gamma} \in \Gamma(\gamma^* TM)$ should be *parallel* along γ, leading to the **geodesic equation**

$$\nabla_t \dot{\gamma} \equiv 0.$$

Paths $\gamma : (a, b) \to M$ that satisfy this condition are called **geodesics** (*Geodäten* or *geodätische Linien*) in M. It should be emphasized that the notion of a geodesic depends on the choice of connection, though we will see shortly that if a pseudo-Riemannian metric g is given, then the connection can be chosen canonically, so that the notion of a geodesic depends only on g.

When $\gamma : (a, b) \to M$ passes through the domain $\mathcal{U} \subset M$ of a chart $(x^1, \ldots, x^n)$, its coordinates define a path $(\gamma^1(t), \ldots, \gamma^n(t))$ in $\mathbb{R}^n$, and using (21.2), the geodesic equation then becomes a system of n second-order nonlinear differential equations for the functions $\gamma^i(t) \in \mathbb{R}$, namely

$$\ddot{\gamma}^i(t) + \Gamma^i_{jk}(\gamma(t))\dot{\gamma}^j(t)\dot{\gamma}^k(t) = 0 \qquad \text{for all } t,$$

or in succinct form,

$$\ddot{\gamma}^i + \Gamma^i_{jk}\dot{\gamma}^j\dot{\gamma}^k \equiv 0. \tag{22.1}$$

As a second-order system on an open set in $\mathbb{R}^n$, (22.1) has a unique solution near any point $t = t_0$ with any given initial position $\gamma(t_0)$ and velocity $\dot{\gamma}(t_0)$. It follows that for every $p \in M$ and $X \in T_pM$, there exists a unique geodesic

$$(a, b) \to M : t \mapsto \gamma_X(t), \qquad \text{such that} \qquad \nabla_t \dot{\gamma}_X \equiv 0,\ \gamma_X(0) = p \text{ and } \dot{\gamma}_X(0) = X.$$

Here $-\infty \leqslant a < 0 < b \leqslant \infty$, and (a, b) is assumed to be the largest possible interval on which the solution γ_X exists. The point $\gamma_X(t) \in M$ is defined for all pairs (t, X) belonging to some open subset of $\mathbb{R} \times TM$, and it depends smoothly on both t and X; this follows from the standard theorem about smooth dependence on initial conditions for ODEs.

Exercise 22.1. Show that for any geodesic $\gamma : (a, b) \to M$ and any constant $c \in \mathbb{R}$, the path defined by $\hat{\gamma}(t) := \gamma(ct)$ on the appropriate interval is also a geodesic.

An interesting consequence of Exercise 22.1 is that the point $\gamma_X(t)$ doesn't just depend smoothly on t and X, it depends in fact only on their product $tX \in TM$. Indeed, consider a pair of colinear vectors $X_1, X_2 \in T_pM$ with $X_2 = cX_1$ for some $c \in \mathbb{R}$. If γ_1 and γ_2 are the unique geodesics through p with $\dot{\gamma}_1(0) = X_1$ and $\dot{\gamma}_2(0) = X_2$, then Exercise 22.1 implies $\gamma_2(t) = \gamma_1(ct)$ for all t, hence $\gamma_1(t_1) = \gamma_2(t_2)$ whenever $t_1 = ct_2$, which means $t_2 X_2 = ct_2 X_1 = t_1 X_1$. To put this observation in its most useful form, we define the open set

$$\mathcal{O} := \{X \in TM \mid 1 \text{ is in the domain of } \gamma_X\}$$

and the smooth function

$$\exp : \mathcal{O} \to M : X \mapsto \gamma_X(1).$$

Note that the domain of exp contains the zero-section of TM since geodesics with $\dot{\gamma}(0) = 0$ can be defined for all time (they are constant). The discussion above proves:

Proposition 22.2. *For each $p \in M$ and $X \in T_pM$, $I_X := \{t \in \mathbb{R} \mid tX \in \mathcal{O}\}$ is an open interval containing 0, and $\gamma : I_X \to M : t \mapsto \exp(tX)$ is the maximal geodesic through $\gamma(0) = p$ with $\dot{\gamma}(0) = X$.* $\square$

We call $\exp : TM \supset \mathcal{O} \to M$ the **exponential map**. For a point $p \in M$, its restriction to an individual tangent space $\mathcal{O}_p := \mathcal{O} \cap T_pM$ is sometimes denoted by

$$\exp_p : \mathcal{O}_p \to M,$$

and it satisfies $\exp_p(0) = p$ since the unique geodesic γ with $\dot{\gamma}(0) = 0 \in T_pM$ is constant. Moreover, Proposition 22.2 implies that the derivative of $\exp_p$ at $0 \in T_pM$ is the identity map,

$$T_0(\exp_p) : T_0(T_pM) = T_pM \to T_pM : X \mapsto \left.\frac{d}{dt}\exp(tX)\right|_{t=0} = X,$$

so that by the inverse function theorem, $\exp_p$ maps a sufficiently small neighborhood of 0 in T_pM diffeomorphically onto a neighborhood of p in M.

The terminology "exponential map" can be motivated in part by the following example: if S^1 is regarded as the unit circle in $\mathbb{C} = \mathbb{R}^2$, then there is a natural connection on S^1 for which the geodesics passing through 1 at time $t = 0$ are precisely the paths of the form $\gamma(t) = e^{i\theta t} =: \exp(ti\theta)$ for $\theta \in \mathbb{R}$, which satisfy $\dot{\gamma}(0) = i\theta \in i\mathbb{R} = T_1S^1$.

22.2. The Levi-Cività connection. For the rest of this lecture, assume (M, g) is a pseudo-Riemannian manifold, which means the tangent bundle $TM \to M$ is equipped with a (possibly indefinite) bundle metric and thus has structure group $\mathrm{O}(k, \ell)$ for some integers $k, \ell \geqslant 0$ with $k + \ell = n = \dim M$. If $(k, \ell) = (n, 0)$, then g is positive and (M, g) is called a Riemannian manifold (without the "pseudo-"). We will sometimes need to assume this, but most of what we do in the present lecture will be equally valid for indefinite metrics. We will often use inner product notation as a synonym for g,

$$\langle\ ,\ \rangle := g(\cdot, \cdot),$$

reserving the notation $g \in \Gamma(T_2^0 M)$ mainly for situations where its role as a tensor field needs to be emphasized. The bundle metric gives $TM \to M$ structure group $\mathrm{O}(k, \ell)$, so we can speak of $\mathrm{O}(k, \ell)$-compatible connections, also known as *metric* connections.

EXERCISE 22.3. Show that the following conditions for a connection ∇ on a vector bundle $E \to M$ with bundle metric $g = \langle\ ,\ \rangle \in \Gamma(E_2^0)$ are equivalent:

(i) ∇ is a metric connection;
(ii) For all $X \in \mathfrak{X}(M)$ and $\eta, \xi \in \Gamma(E)$, $\mathcal{L}_X\langle\eta, \xi\rangle = \langle\nabla_X\eta, \xi\rangle + \langle\eta, \nabla_X\xi\rangle$;
(iii) The induced connection on E_2^0 satisfies $\nabla g \equiv 0$.

A connection on a real vector bundle with bundle metric g is said to be **compatible with** g if it is a metric connection, or equivalently, if it satisfies any of the conditions in Exercise 22.3.

The next result is sometimes called the *fundamental theorem of (pseudo-)Riemannian geometry*, because almost every other result in the subject depends on it. It is independent of our previous proof that connections on vector bundles always exist, so if you combine it with the theorem that every manifold admits a Riemannian metric, it implies a second proof of the fact that every manifold admits a symmetric connection.

THEOREM 22.4. *For any pseudo-Riemannian manifold (M, g), there exists a unique connection on $TM \to M$ that is symmetric and compatible with g.*

PROOF. We first show uniqueness: assuming ∇ is such a connection, Exercise 22.3 implies that for any vector fields X, Y and Z, we have the three relations

$$\begin{aligned}
\mathcal{L}_X\langle Y, Z\rangle &= \langle\nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle, \\
\mathcal{L}_Y\langle Z, X\rangle &= \langle\nabla_Y Z, X\rangle + \langle Z, \nabla_Y X\rangle, \\
\mathcal{L}_Z\langle X, Y\rangle &= \langle\nabla_Z X, Y\rangle + \langle X, \nabla_Z Y\rangle.
\end{aligned}$$

Adding the first two, subtracting the third and using the assumption $T(X,Y) = \nabla_X Y - \nabla_Y X - [X,Y] \equiv 0$, we find

$$\begin{aligned}&\mathcal{L}_X\langle Y,Z\rangle + \mathcal{L}_Y\langle Z,X\rangle - \mathcal{L}_Z\langle X,Y\rangle\\ &\quad= \langle \nabla_X Y + \nabla_Y X, Z\rangle + \langle Y, \nabla_X Z - \nabla_Z X\rangle + \langle X, \nabla_Y Z - \nabla_Z Y\rangle\\ &\quad= \langle 2\nabla_X Y, Z\rangle - \langle [X,Y],Z\rangle + \langle [X,Z],Y\rangle + \langle [Y,Z],X\rangle,\end{aligned}$$

thus

$$\begin{aligned}\langle \nabla_X Y, Z\rangle = \frac{1}{2}\Big(&\mathcal{L}_X\langle Y,Z\rangle + \mathcal{L}_Y\langle Z,X\rangle - \mathcal{L}_Z\langle X,Y\rangle\\ &+ \langle [X,Y],Z\rangle - \langle [X,Z],Y\rangle - \langle [Y,Z],X\rangle\Big).\end{aligned} \tag{22.2}$$

A straightforward (though slightly tedious) calculation shows that the right hand side of this expression is C^∞-linear with respect to X and Z. It therefore associates to every $Y \in \mathfrak{X}(M)$ a tensor field $S_Y \in \Gamma(T_2^0 M)$ such that (22.2) can be rewritten in succinct form as

$$\langle \nabla_X Y, \cdot\rangle = S_Y(X,\cdot).$$

This uniquely determines $\nabla_X Y$, since by the nondegeneracy of g, the map $T_pM \to T_p^*M : Z \mapsto g_p(Z,\cdot)$ is an isomorphism for every $p \in M$, implying that $\mathfrak{X}(M) \to \Omega^1(M) : Z \mapsto \langle Z,\cdot\rangle$ is also an isomorphism. If one now *defines* $\nabla_X Y$ in terms of $S_Y(X,\cdot)$ via this relation for every $Y \in \mathfrak{X}(M)$ and $X \in TM$, one can check that it satisfies the required Leibniz rule and is thus a connection on M, in addition to being symmetric and compatible with g. □

The connection in Theorem 22.4 is called the **Levi-Cività connection** on (M,g). Whenever we discuss pseudo-Riemannian manifolds from now on, we will always use the Levi-Cività connection for computations on its tangent bundle, along with the various induced connections that it determines on associated bundles such as T^*M and $T_\ell^k M$. The first hint that this might be the "right" thing to do comes from the fact that the Levi-Cività connection does not depend on any choices other than the metric; this is the first time we have seen a connection that is not some kind of arbitrary choice. The real justification for using this in preference to any other connection will come from the multitude of geometrically-motivated theorems that we can use it to prove, e.g. the fact (to be proved in the next lecture) that for the Levi-Cività connection on a Riemannian manifold, geodesics are not only the natural generalization of the notion of a "straight line" but also define *shortest* paths between nearby points.

22.3. Musical isomorphisms and coordinates. We would like to write down an explicit local coordinate formula for the Levi-Cività connection. The following algebraic remarks serve as preparation for this.

On a real vector bundle $E \to M$, any bundle metric $\langle\ ,\ \rangle$ determines a natural smooth linear bundle map

$$\flat : E \to E^* : v \mapsto v_\flat := \langle v,\cdot\rangle.$$

The nondegeneracy of $\langle\ ,\ \rangle$ implies that $\flat$ is injective on every fiber, and it is therefore a bundle isomorphism; note that this is true for any *nondegenerate* bilinear form, so in particular $\langle\ ,\ \rangle$ may be an indefinite bundle metric, it need not be positive. The inverse of $\flat$ is denoted by

$$\sharp : E^* \to E : v \mapsto v^\sharp,$$

and notation motivates terminology: we call $\flat$ and $\sharp$ the **musical isomorphisms** determined by $\langle\ ,\ \rangle$.

As an isomorphism, $\flat$ can be used to transfer all data from E to E^*, e.g. it gives a natural definition of a bundle metric on E^*, namely

$$\langle \lambda, \mu \rangle := \langle \lambda^\sharp, \mu^\sharp \rangle \qquad \text{for} \qquad \lambda, \mu \in E^* \oplus E^*. \tag{22.3}$$

EXERCISE 22.5. Assume ∇ is a metric connection on E. Show:

(a) For the induced connections on $\mathrm{Hom}(E, E^*)$ and $\mathrm{Hom}(E^*, E)$, $\nabla(\flat) \equiv 0$ and $\nabla(\sharp) \equiv 0$.
(b) The induced connection on E^* is compatible with the bundle metric (22.3).

Choose a frame $e_1, \ldots, e_m$ for E over some open set $\mathcal{U} \subset M$, let $e^1_*, \ldots, e^m_*$ denote the dual frame, and denote the resulting components of the bundle metrics on E and E^* by

$$g_{ij} := \langle e_i, e_j \rangle, \qquad g^{ij} := \langle e^i_*, e^j_* \rangle.$$

For $v = v^i e_i, w = w^i e_i \in E_p$ and $\lambda = \lambda_i e^i_*, \mu = \mu_i e^i_* \in E^*_p$ at a point $p \in \mathcal{U}$, one then has

$$\langle v, w \rangle = g_{ij} v^i w^j, \qquad \langle \lambda, \mu \rangle = g^{ij} \lambda_i \mu_j. \tag{22.4}$$

The convention for the musical isomorphisms is that for $v \in E_p$ or $\lambda \in E^*_p$, one writes the components of $v_\flat$ and $\lambda^\sharp$ with the *same* symbol but with the index raised or lowered, thus

$$\eta = \eta^i e_i \quad \Leftrightarrow \quad \eta_\flat = \eta_i e^i_*, \qquad \text{and} \qquad \lambda = \lambda_i e^i_* \quad \Leftrightarrow \quad \lambda^\sharp = \lambda^i e_i.$$

Philosophically, this means in some sense that we are considering vectors in E and dual vectors in E^* to be two distinct presentations of the same fundamental object. Since $\langle v, w \rangle = v_\flat(w) = w_\flat(v)$ and $\langle \lambda, \mu \rangle = \langle \lambda^\sharp, \mu^\sharp \rangle = \lambda(\mu^\sharp) = \mu(\lambda^\sharp)$, the bundle metrics on E and E^* can now be written in the appealing form

$$\langle v, w \rangle = v^i w_i = v_i w^i, \qquad \langle \lambda, \mu \rangle = \lambda_i \mu^i = \lambda^i \mu_i.$$

Comparing this with (22.4), you may notice that it implies explicit coordinate formulas for the maps $\flat$ and $\sharp$, namely

$$v_i = g_{ij} v^j, \qquad \text{and} \qquad \lambda^i = g^{ij} \lambda_j.$$

Since $\sharp = \flat^{-1}$, it follows that the m-by-m matrices with entries g_{ij} and g^{ij} are inverse to each other, i.e.

$$g_{ij} g^{jk} = \delta^k_i. \tag{22.5}$$

One can always deduce the components g^{ij} from this fact once the g_{ij} are known.

REMARK 22.6. It was important throughout this discussion that E is a *real* vector bundle, not complex. Several details would need to modified if E were a complex bundle, starting with the observation that $\flat$ and $\sharp$ as we defined them are no longer bundle isomorphisms, as they are complex antilinear, not complex linear.

Specializing to the case where $E = TM$ for a pseudo-Riemannian manifold (M, g), we can define the musical isomorphisms $\flat : TM \to T^*M$ and $\sharp : T^*M \to TM$ as above, use them to define a bundle metric $\langle \ , \ \rangle$ on T^*M, then fix a chart $(\mathcal{U}, x)$ and write

$$g_{ij} = \langle \partial_i, \partial_j \rangle, \qquad g^{ij} = \langle dx^i, dx^j \rangle \qquad \text{on } \mathcal{U}.$$

Setting $X := \partial_i$, $Y := \partial_j$ and $Z := \partial_k$, $\nabla_X Y$ can be expressed in terms of the Christoffel symbols using (20.5), and we thus have

$$\langle \nabla_X Y, Z \rangle = \langle \nabla_i \partial_j, \partial_k \rangle = \langle \Gamma^a_{ij} \partial_a, \partial_k \rangle = g_{ak} \Gamma^a_{ij}.$$

If ∇ is the Levi-Cività connection, then (22.2) equates this with $\frac{1}{2}\left(\partial_i g_{jk} + \partial_j g_{ki} - \partial_k g_{ij}\right)$, in which the Lie bracket terms from (22.2) do not appear since coordinate vector fields always commute with each other. Applying (22.5) now gives a formula for explicitly computing the Levi-Cività connection: its Christoffel symbols are

$$\Gamma^\ell_{ij} = \frac{1}{2} g^{k\ell} \left(\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij} \right). \tag{22.6}$$

EXAMPLE 22.7. Consider $\mathbb{R}^n$ with what we will henceforth call the **standard Euclidean metric**, meaning the Riemannian metric defined via the Euclidean inner product. The Levi-Cività connection is in this case exactly what you would expect: since the components $g_{ij} = \delta_{ij}$ of the metric are all constant, the Christoffel symbols computed via (22.6) all vanish identically, and ∇ is therefore the trivial connection on the trivial bundle $T\mathbb{R}^n = \mathbb{R}^n \times \mathbb{R}^n$. Note that this is true for *any* choice of pseudo-Riemannian metric on $\mathbb{R}^n$ whose components are constant, including indefinite metrics such as the Minkowski metric of special relativity. The geodesic equation for paths $\gamma : (a, b) \to \mathbb{R}^n$ is thus $\ddot{\gamma} = 0$, and its solutions are straight lines with constant speed.

EXERCISE 22.8. The **Poincaré half-plane** $(\mathbb{H}, h)$ is the 2-manifold

$$\mathbb{H} = \{(x, y) \in \mathbb{R}^2 \mid y > 0\} \subset \mathbb{R}^2$$

with Riemannian metric

$$h_{(x,y)}(X, Y) = \frac{1}{y^2}\langle X, Y\rangle_E \qquad \text{for } X, Y \in T_{(x,y)}\mathbb{H} = \mathbb{R}^2,$$

where $\langle\ ,\ \rangle_E$ denotes the Euclidean inner product on $\mathbb{R}^2$. As we will later see, this is an example of a surface with constant negative curvature.

(a) Using the obvious global coordinates, derive the Christoffel symbols for the Levi-Cività connection on $(\mathbb{H}, h)$ and show that for this connection, the geodesic equation can be written as

$$\ddot{x} - \frac{2}{y}\dot{x}\dot{y} = 0, \qquad \ddot{y} + \frac{1}{y}(\dot{x}^2 - \dot{y}^2) = 0$$

for a smooth path $\gamma(t) = (x(t), y(t))$.

(b) Show that for any constants $x_0 \in \mathbb{R}$ and $r > 0$, the geodesic equation in part (a) has solutions of the form

$$\gamma(t) = (x_0, y(t)), \qquad \text{or} \qquad \gamma(t) = (x_0 + r\cos\theta(t), r\sin\theta(t))$$

for appropriately chosen functions $y(t) > 0$ and $\theta(t) \in (0, \pi)$.

(c) Show that any two points in $(\mathbb{H}, h)$ are connected by a unique geodesic segment $\gamma : [a, b] \to M$, and compute the length of this segment, meaning the integral

$$\ell^b_a(\gamma) := \int_a^b |\dot{\gamma}(t)|\, dt := \int_a^b \sqrt{\langle\dot{\gamma}(t), \dot{\gamma}(t)\rangle}\, dt.$$

22.4. Arc length and the energy functional. We shall now begin exploring the relationship between the geodesics of the Levi-Cività connection and the problem of finding paths of minimal length between fixed points. This uses some basic concepts from the *calculus of variations*, which deals with optimization problems on infinite dimensional spaces. Fix two points $p, q \in M$ and real numbers $a < b$. We denote by

$$C^\infty([a, b], M\,; p, q)$$

the set of all smooth paths $\gamma : [a, b] \to M$ such that $\gamma(a) = p$ and $\gamma(b) = q$. Given a Riemannian metric $g = \langle\ ,\ \rangle$, we denote

$$|X| := \sqrt{\langle X, X\rangle}$$

and define the **length functional** on $C^\infty([a,b], M\,;\, p, q)$ by

$$\ell_a^b(\gamma) = \int_a^b |\dot\gamma(t)|\, dt.$$

A related functional is the **energy functional**,

$$E_a^b(\gamma) = \int_a^b \langle \dot\gamma(t), \dot\gamma(t)\rangle\, dt,$$

which also makes sense for an arbitrary pseudo-Riemannian metric, i.e. there is no need to assume g is positive. The geometric meaning of ℓ_a^b is clear: $\ell_a^b(\gamma)$ is the length of the path traced out by γ, as measured with respect to the Riemannian metric g. As such, it depends only on the image of γ, and is thus invariant under reparametrizations, i.e. for any diffeomorphism $\varphi : [a,b] \to [a', b']$ and smooth path $\gamma \in C^\infty([a', b'], M; p, q)$, we have

$$\ell_a^b(\gamma \circ \varphi) = \ell_{a'}^{b'}(\gamma).$$

It is less obvious what geometric meaning the energy functional may have, especially in the indefinite case, but we will find it convenient as a computational tool in order to understand the length functional better.

We wish to view $C^\infty([a,b], M\,;\, p, q)$ informally as an infinite-dimensional manifold, and E_a^b and ℓ_a^b as "smooth functions" on this manifold which can be differentiated. There will be no need to define this in formal terms, because for the type of optimization problem we have in mind, it always suffices to consider the values of each functional along paths in $C^\infty([a,b], M\,;\, p, q)$. In general, for a given functional

$$F : C^\infty([a,b], M\,;\, p, q) \to \mathbb{R},$$

the first goal of the calculus of variations is to find necessary conditions on a smooth path $\gamma \in C^\infty([a,b], M\,;\, p, q)$ so that $F(\gamma)$ may attain a minimal or maximal value among all paths $\gamma_\epsilon \in C^\infty([a,b], M\,;\, p, q)$ close to γ. This condition will typically take the form of a differential equation that γ must satisfy. To make this precise, we say that a **smooth 1-parameter family** of paths from p to q is a collection $\gamma_s \in C^\infty([a,b], M\,;\, p, q)$ for $s \in (-\epsilon, \epsilon)$ such that the map $(s,t) \mapsto \gamma_s(t)$ is smooth. Informally, we think of this as a smooth path in $C^\infty([a,b], M\,;\, p, q)$ through γ_0, and its "velocity vector" at $s = 0$ is then given by the partial derivatives $\partial_s \gamma_s(t)|_{s=0} \in T_{\gamma_0(t)}M$ for all t, which define a vector field along γ_0,

$$\eta := \partial_s \gamma_s|_{s=0} \in \Gamma(\gamma_0^* TM),$$

such that $\eta(a) = 0$ and $\eta(b) = 0$. We therefore think of the vector space

$$\{\eta \in \Gamma(\gamma^* TM) \mid \eta(a) = 0 \text{ and } \eta(b) = 0\}$$

as the "tangent space" to $C^\infty([a,b], M\,;\, p, q)$ at γ. It is now clear how one should define a "directional derivative" of F in a direction defined by a section of $\gamma^* TM$. This motivates the following definition, which generalizes the notion of a critical point of a real-valued function in finite dimensions.

Definition 22.9. The path $\gamma \in C^\infty([a,b], M\,;\, p, q)$ is called **stationary** for the functional $F : C^\infty([a,b], M\,;\, p, q) \to \mathbb{R}$ if for every smooth 1-parameter family $\gamma_s \in C^\infty([a,b], M\,;\, p, q)$ with $\gamma_0 = \gamma$,

$$\frac{d}{ds} F(\gamma_s)\Big|_{s=0} = 0. \tag{22.7}$$

Note that for an arbitrary functional, it is not *a priori* clear that the derivatives in (22.7) will always exist. This is however true in many cases of interest, and in such a situation, it's easy to see that (22.7) is a necessary condition for F to attain an extremal value at γ.

PROPOSITION 22.10. *The energy functional E_a^b is stationary at γ if and only if γ satisfies the geodesic equation for the Levi-Cività connection.*

PROOF. Pick any smooth 1-parameter family $\gamma_s \in C^\infty([a,b], M\,;\, p,q)$ with $\gamma_0 = \gamma$ and denote $\eta = \partial_s \gamma_s|_{s=0} \in \Gamma(\gamma^* TM)$. In the following calculation, we regard $\partial_s \gamma_s(t)$ and $\dot{\gamma}_s(t) := \partial_t \gamma_s(t)$ as defining vector fields along the map $(s,t) \mapsto \gamma_s(t) \in M$, which can then be covariantly differentiated using the pullback connection. Differentiating under the integral sign and using the properties of the Levi-Cività connection, we have

$$\begin{aligned} \frac{d}{ds} E_a^b(\gamma_s)\bigg|_{s=0} &= \int_a^b \frac{\partial}{\partial s} \big\langle \partial_t \gamma_s(t), \partial_t \gamma_s(t) \big\rangle \bigg|_{s=0} dt \\ &= \int_a^b \Big(\big\langle \left. \nabla_s \partial_t \gamma_s(t) \right|_{s=0}, \dot{\gamma}(t) \big\rangle + \big\langle \dot{\gamma}(t), \left. \nabla_s \partial_t \gamma_s(t) \right|_{s=0} \big\rangle \Big) \, dt \\ &= 2 \int_a^b \big\langle \dot{\gamma}(t), \nabla_t \left. \partial_s \gamma_s(t) \right|_{s=0} \big\rangle \, dt = 2 \int_a^b \big\langle \dot{\gamma}(t), \nabla_t \eta(t) \big\rangle \, dt, \end{aligned}$$

where in the last line we've used the symmetry of the connection to replace $\nabla_s \partial_t$ with $\nabla_t \partial_s$. We now perform a geometric version of integration by parts, using the fact that $\eta(t)$ vanishes at the end points. It follows indeed from the fundamental theorem of calculus that

$$\begin{aligned} 0 = \langle \dot{\gamma}(b), \eta(b) \rangle - \langle \dot{\gamma}(a), \eta(a) \rangle &= \int_a^b \frac{d}{dt} \big\langle \dot{\gamma}(t), \eta(t) \big\rangle \, dt \\ &= \int_a^b \big\langle \nabla_t \dot{\gamma}(t), \eta(t) \big\rangle \, dt + \int_a^b \big\langle \dot{\gamma}(t), \nabla_t \eta(t) \big\rangle \, dt, \end{aligned}$$

thus

$$\frac{d}{ds} E_a^b(\gamma_s)\bigg|_{s=0} = -2 \int_a^b \big\langle \nabla_t \dot{\gamma}(t), \eta(t) \big\rangle \, dt.$$

Since choosing arbitrary 1-parameter families γ_s leads to arbitrary sections $\eta \in \Gamma(\gamma^* TM)$ with $\eta(a) = 0$ and $\eta(b) = 0$, this expression will vanish for all such choices if and only if $\nabla_t \dot{\gamma} \equiv 0$. $\square$

To see what this tells us about the length functional, suppose now that the metric $\langle \ , \ \rangle$ is positive, so that $|X| = \sqrt{\langle X, X \rangle}$ can be defined and interpreted as the length of any tangent vector $X \in TM$. The **speed** of a path $\gamma : (a,b) \to M$ at time t is then the length of its velocity vector $\dot{\gamma}(t)$, and another easy observation about geodesics follows from the fact that ∇ is a metric connection: we have

$$\partial_t |\dot{\gamma}(t)|^2 = \partial_t \langle \dot{\gamma}(t), \dot{\gamma}(t) \rangle = 2 \langle \nabla_t \dot{\gamma}(t), \dot{\gamma}(t) \rangle,$$

so if the geodesic equation is satisfied, the speed $|\dot{\gamma}(t)|$ is constant. We claim: every immersed path $\gamma \in C^\infty([a,b], M\,;\, p,q)$ has a unique reparametrization $\beta = \gamma \circ \varphi \in C^\infty([a,b], M\,;\, p,q)$ that has constant speed. Indeed, to derive β, we can first figure out which constant $v := |\dot{\beta}(t)|$ needs to be: since $\ell_a^b(\gamma) = \ell_a^b(\beta) = \int_a^b v \, dt = v(b-a)$, we must have $v = \ell_a^b(\gamma)/(b-a)$. If we then assume $\beta = \gamma \circ \varphi : [a,b'] \to [a,b]$ for some $b' > a$ and a strictly increasing diffeomorphism $\varphi : [a,b] \to [a,b']$, the condition $|\dot{\beta}(t)| = v$ is satisfied if and only if φ satisfies the differential equation $\dot{\varphi}(t) = v / |\dot{\gamma}(\varphi(t))|$. The right hand side of this equation is positive and bounded away from 0, so after imposing the initial condition $\varphi(a) = a$, there will be a unique solution φ on some interval $[a, b']$ with $b' > a$ uniquely determined by the condition $\varphi(b') = b$. Appealing again to reparametrization invariance, we then find

$$\ell_a^b(\gamma) = \ell_a^{b'}(\beta) = \int_a^{b'} v\,dt = (b'-a)v = \frac{b'-a}{b-a}\ell(\gamma),$$

and thus conclude $b' = b$, proving the claim.

The reparametrization-invariance of ℓ_a^b implies that whenever a path $\gamma \in C^\infty([a,b], M\,;\,p,q)$ is stationary for ℓ_a^b, all its reparametrizations are as well. Now if $\gamma_s \in C^\infty([a,b], M\,;\,p,q)$ is a smooth 1-parameter family of *immersed* paths for which $\gamma := \gamma_0$ happens to have constant speed $v := |\dot\gamma|$, we find

$$\begin{aligned}\left.\frac{d}{ds}\ell_a^b(\gamma_s)\right|_{s=0} &= \int_a^b \left.\frac{\partial}{\partial s}\sqrt{\langle\dot\gamma_s(t), \dot\gamma_s(t)\rangle}\right|_{s=0} dt = \int_a^b \frac{1}{2\sqrt{\langle\dot\gamma_0(t), \dot\gamma_0(t)\rangle}} \left.\frac{\partial}{\partial s}\langle\dot\gamma_s(t), \dot\gamma_s(t)\rangle\right|_{s=0} dt \\ &= \frac{1}{2v}\left.\frac{d}{ds}E_a^b(\gamma_s)\right|_{s=0}.\end{aligned}$$

It follows that if γ is stationary for ℓ_a^b, then it has a reparametrization with constant speed that is stationary for E_a^b, and is therefore a geodesic. Conversely, every geodesic is stationary for ℓ_a^b, and also has constant speed. This proves:

Corollary 22.11. *In a Riemannian manifold (M, g), an immersed path $\gamma \in C^\infty([a,b], M\,;\,p,q)$ is a geodesic if and only if it is stationary for the length functional ℓ_a^b and has constant speed.* $\square$

We conclude that any path $\gamma \in C^\infty([a,b], M\,;\,p,q)$ which minimizes the length $\ell_a^b(\gamma)$ among all nearby paths from p to q can be parametrized by a geodesic. We will discuss a "local" converse to this in the next lecture.

Remark 22.12. One can extend our discussion of the length functional to the indefinite case with the following modifications. The argument above that $\langle\dot\gamma, \dot\gamma\rangle$ is constant for any geodesic γ is valid for metrics of arbitrary signature, so it makes sense to distinguish between cases where this constant is positive, negative or zero. In general relativity, where the metric has signature $(1,3)$,[63] one calls a geodesic **time-like** if $\langle\dot\gamma, \dot\gamma\rangle > 0$, **space-like** if it is negative and **light-like** if it vanishes. Space-like geodesics represent paths in spacetime that would be perceived by a three-dimensional observer to move faster than the speed of light, while time-like geodesics move slower, and light-like geodesics move (unsurprisingly) at precisely the speed of light. According to the known laws of physics, all freely moving objects with positive mass traverse time-like geodesics in spacetime, and massless objects traverse light-like geodesics. Nothing can traverse a space-like geodesic; its "speed" $|\dot\gamma| := \sqrt{\langle\dot\gamma, \dot\gamma\rangle}$ as measured by the Lorentzian metric $g = \langle\ ,\ \rangle$ would be imaginary.

With this understood, the length functional makes sense for time-like paths, and Corollary 22.11 remains true on a Lorentzian manifold if one restricts attention to time-like geodesics.

23. More on geodesics

23.1. Normal coordinates. Assume M is a smooth manifold without boundary, with a connection ∇. In the previous lecture, we defined the exponential map

$$\exp : TM \supset \mathcal{O} \to M,$$

which is defined on an open subset $\mathcal{O} \subset TM$ containing the zero-section and can be characterized by the property that for each $X \in TM$, $\gamma(t) := \exp(tX)$ is the unique geodesic ($\nabla_t\dot\gamma \equiv 0$) satisfying

[63] Many authors also prefer to take $(3,1)$ as the signature of a spacetime manifold, in which case the definitions of the terms "space-like" and "time-like" should be modified by a sign.

$\dot\gamma(0) = X$, defined for t in the largest possible interval. We also observed that for each $p \in M$, the restriction

$$\exp_p := \exp|_{\mathcal{O}_p} : \mathcal{O}_p \to M \qquad \text{for } \mathcal{O}_p := \mathcal{O} \cap T_pM$$

satisfies $\exp_p(0) = p$ and has derivative equal to the identity map $T_pM \to T_pM$ at $0 \in \mathcal{O}_p$, implying that it maps a neighborhood of $0 \in T_pM$ diffeomorphically onto a neighborhood of $p \in M$. This means that $\exp_p$ can be used to define local coordinates near p, i.e. if we choose any basis $X_1, \dots, X_n$ of T_pM, then

$$\varphi(t^1, \dots, t^n) := \exp_p(t^i X_i)$$

defines a diffeomorphism from some neighborhood of $0 \in \mathbb{R}^n$ to some neighborhood $\mathcal{U} \subset M$ of p, and its inverse $x = (x^1, \dots, x^n) : \mathcal{U} \to x(\mathcal{U}) \subset \mathbb{R}^n$ is therefore a chart sending p to 0.

Charts defined via the exponential map as described above are often referred to as **normal coordinates** about p. They have the following special property. By construction, any path through p that looks in normal coordinates like a straight line with constant velocity through the origin is a geodesic, and any path of this form is also a flow line (through p) of some vector field $Y \in \mathfrak{X}(M)$ that has constant components near p in normal coordinates. The geodesic equation thus implies that all vector fields with this property satisfy

$$\nabla_{Y(p)} Y = 0.$$

This applies in particular to the coordinate vector fields $\partial_1, \dots, \partial_n$, as well as their linear combinations with constant coefficients, such as $\partial_i + \partial_j$. We therefore have

$$0 = \nabla_{\partial_i + \partial_j}(\partial_i + \partial_j) = \nabla_i \partial_i + \nabla_j \partial_j + \nabla_i \partial_j + \nabla_j \partial_i = \nabla_i \partial_j + \nabla_j \partial_i \quad \text{at } p,$$

implying that the Christoffel symbols satisfy

$$\Gamma^k_{ij} + \Gamma^k_{ji} = 0 \qquad \text{at } p.$$

If the connetion is symmetric, this implies that the Christoffel symbols vanish at p, and we've proved:

PROPOSITION 23.1. *For any symmetric connection ∇ on M, the Christoffel symbols vanish in any normal coordinate system about p.* □

To take this a step further, suppose (M, g) is a pseudo-Riemannian manifold with signature (k, ℓ) and ∇ is the Levi-Cività connection. In this setting we can require the basis $X_1, \dots, X_n \in T_pM$ in the construction above to be orthonormal, meaning

$$\langle X_i, X_j \rangle = \eta_{ij} := \begin{cases} 1 & \text{if } i = j \leqslant k, \\ -1 & \text{if } i = j > k, \\ 0 & \text{if } i \neq j, \end{cases}$$

and normal coordinates about p under this extra condition are called **Riemann normal coordinates**. The vectors $X_1, \dots, X_n$ match the coordinate vector fields $\partial_1, \dots, \partial_n$ at p, so the components $g_{ij} = \langle \partial_i, \partial_j \rangle$ of the metric now match η_{ij} at p, meaning that $\langle\ ,\ \rangle$ matches the "standard" inner product of signature (k, ℓ) on $\mathbb{R}^n$ at that one point. The vanishing of the Christoffel symbols at that point implies moreover that

$$\partial_k g_{ij}(p) = \partial_k \langle \partial_i, \partial_j \rangle\big|_p = \langle \nabla_k \partial_i, \partial_j \rangle\big|_p + \langle \partial_i, \nabla_k \partial_j \rangle\big|_p = 0$$

for all i, j, k, since the covariant derivatives of the coordinate vector fields all vanish at p. This proves:

PROPOSITION 23.2. *In any Riemann normal coordinate system about a point p in a pseudo-Riemannian manifold (M, g), the components g_{ij} of the metric satisfy*

$$g_{ij}(p) = \eta_{ij}, \qquad \text{and} \qquad \partial_k g_{ij}(p) = 0$$

for all $i, j, k \in \{1, \dots, n\}$. $\square$

Riemann normal coordinates are sometimes useful for calculations, but their existence also has theoretical importance, for the following reason. The simplest example of a pseudo-Riemannian manifold with signature (k, ℓ) is $\mathbb{R}^n$ with a metric whose components are given by the constants η_{ij} in the obvious global coordinates. In fact, the classification of quadratic forms implies (cf. §18.5) that any pseudo-Riemannian metric on $\mathbb{R}^n$ with constant components can be turned into this one by a global linear change of coordinates. When the signature is $(n, 0)$, this is what we call the *Euclidean* metric; the case of signature $(1, n-1)$ or $(n-1, 1)$ is called the **Minkowski metric**, and is important in special relativity. Anticipating the relevance of curvature to this discussion, we shall refer to this metric for arbitrary signatures as the **flat metric** on $\mathbb{R}^n$. The significance of Riemann normal coordinates according to Proposition 23.2 is that they make an arbitrary metric g look more like the flat one, at least at a single point—its value and first derivative at that point match the flat case. We will see when we discuss curvature that, in general, one cannot do better than this: arbitrary pseudo-Riemannian manifolds cannot be made to look like flat space on open neighborhoods of a point just by choosing the right coordinates. Attempting to do this will run into problems as soon as one tries to make the *second* derivatives of g_{ij} vanish, and this impossibility is one of the things that curvature measures.

REMARK 23.3. Another nice trick one can play with the exponential map is to obtain standardized models for arbitrary smooth submanifolds. The result, known as the **tubular neighborhood theorem**, says that if $N \subset M$ is a submanifold and M and N both have empty boundary, then there is a diffeomorphism Φ from some neighborhood $\mathcal{U} \subset M$ of N to a neighborhood $\mathcal{O} \subset \nu N$ of the zero-section in the total space of its normal bundle (see Example 17.15), and Φ identifies N itself with the zero-section. This result is useful because vector bundles of a given rank over a given manifold are typically not so hard to classify up to bundle isomorphism, thus one obtains manageable lists of models that can describe neighborhoods of all possible embeddings of N into M. For example, one can show that all orientable vector bundles over S^1 are trivial, so one concludes that all embeddings of S^1 into an orientable n-manifold have neighborhoods that look like $S^1 \times \mathbb{D}^{n-1}$; this fact is crucial in knot theory. We refer to [**Hir94**, Chapter 4] for a general discussion of the tubular neighborhood theorem, including a version for manifolds with nonempty boundary. The case of a compact submanifold without boundary is easier, and is Exercise 23.4 below.

EXERCISE 23.4. Suppose N is a compact smooth submanifold of M, where ∂N and ∂M are empty. Choose a Riemannian metric $g = \langle\ ,\ \rangle$ on M and recall from Exercise 17.16 that the subbundle $TN^\perp \subset TM|_N$ is isomorphic to the normal bundle of N. Prove:

(a) For any $\epsilon > 0$ sufficiently small, the set

$$\mathbb{D}_\epsilon(TN)^\perp := \left\{ X \in TN^\perp \;\middle|\; |X| < \epsilon \right\}$$

is contained in the domain $\mathcal{O}$ of the exponential map $\exp : TM \supset \mathcal{O} \to M$.
Hint: For every $p \in M$, the zero vector in T_pM has a neighborhood in TM that belongs to the domain of exp. *Use the fact that N is compact.*

(b) The derivative of the smooth map

$$\Psi := \exp|_{\mathbb{D}_\epsilon(TN)^\perp} : \mathbb{D}_\epsilon(TN)^\perp \to M$$

is invertible at every zero vector in $TN^\perp$. It follows via the inverse function theorem that for each $p \in N$, Ψ restricts to a diffeomorphism from some neighborhood of $0 \in T_pM$ in $\mathbb{D}_\epsilon(TN)^\perp$ to a neighborhood of p in M.

(c) After possibly shrinking $\epsilon > 0$ further, the map Ψ in part (b) is a diffeomorphism onto an open neighborhood of N in M.

23.2. The shortest path between nearby points. [64]

Assume (M, g) is a Riemannian manifold and ∇ is the Levi-Cività connection. As you know, if (M, g) is Euclidean space, then the shortest path between any two distinct points $p, q \in M$ is a straight line, also known as a geodesic. Is this true in all Riemannian manifolds? We saw in the previous lecture for instance that any path with constant speed that is a local minimum of the length functional on paths from p to q must be a geodesic. Various subtleties can arise, however, because it is not always true that there is a unique geodesic from p to q, nor must every geodesic from p to q be shorter than all other paths; the easiest example to imagine here is the unit sphere $S^2 \subset \mathbb{R}^3$, which we'll discuss in more detail in the next lecture. More can be said however if we assume that p and q are sufficiently close to each other:

THEOREM 23.5. *For every point p in a Riemannian manifold (M, g), there is a neighborhood $\mathcal{U} \subset M$ of p such that for each $q \in \mathcal{U}$, there exists an embedded geodesic segment $\gamma : [0, 1] \to M$ from $\gamma(0) = p$ to $\gamma(1) = q$ that is strictly shorter than all paths from p to q other than the reparametrizations of γ.*

The existence of the neighborhood $\mathcal{U} \subset M$ in this theorem is the easy part: we have already seen that $\exp_p$ maps some neighborhood $\mathcal{O} \subset T_pM$ of 0 diffeomorphically to a neighborhood $\mathcal{U} \subset M$ of p. Every $q \in \mathcal{U}$ can then be written as $q = \exp_p(X)$ for a unique $X \in \mathcal{O}$, and we may as well assume $\mathcal{O} \subset T_pM$ is a **star-shaped** neighborhood, meaning that $tX \in \mathcal{O}$ for every $X \in \mathcal{O}$ and $t \in [0, 1]$, so that the geodesic segment $\gamma : [0, 1] \to M : t \mapsto \exp(tX)$ from p to q is also contained in $\mathcal{U}$. Note that this might not necessarily be the *only* geodesic segment connecting p to q, though it is certainly the only one that is fully contained in $\mathcal{U}$. The goal is to show that this particular geodesic segment and its reparametrizations are strictly shorter than all other paths from p to q.

The key turns out to be an observation that sounds eminently plausible in our geometric intuition, but is a bit tricky to prove: every geodesic emerging from p is *orthogonal* to the spheres of constant radius around p. By "spheres of constant radius", we mean more precisely the following: for each $r \in \mathbb{R}$, consider the set

$$\Sigma_r := \left\{ \exp_p(X) \in M \;\middle|\; X \in \mathcal{O} \text{ and } \langle X, X \rangle = r \right\} \subset \mathcal{U}. \tag{23.1}$$

The definition also makes sense when the metric is indefinite, so we have allowed r to be any real number, not just $r > 0$. The condition $\langle X, X \rangle = r$ cuts out a smooth hypersurface in T_pM for any $r \neq 0$, and this is also true at $r = 0$ with the exception of a singular point at the origin, thus Σ_r is a smooth hypersurface in M for every $r \neq 0$, and so is Σ_0 except at the isolated singular point $p \in \Sigma_0$, which we will exclude.[65]

PROPOSITION 23.6 (Gauss lemma). *Assume (M, g) is a pseudo-Riemannian manifold without boundary, $0 \in \mathcal{O} \subset T_pM$ denotes the star-shaped neighborhood described above with $\mathcal{U} = \exp_p(\mathcal{O})$, and $\Sigma_r \subset \mathcal{U}$ is the hypersurface defined in (23.1). Then for every $r \in \mathbb{R}$, any geodesic segment of*

---

[64] All details concerning indefinite metrics in this section should be considered inessential to the course and are included only out of interest—they were not covered in the lecture.

[65] If the metric is positive, then Σ_0 consists only of the point p and is thus excluded from this discussion. But Σ_0 is more interesting in the indefinite case: imagine for instance the standard indefinite inner product of signature $(1, 1)$ on $\mathbb{R}^2$, so that $\langle X, X \rangle = 0$ is equivalent to the equation $x^2 - y^2 = 0$. This cuts out a smooth submanifold with an isolated singularity at the origin.

the form $\gamma(t) = \exp_p(tX)$ *for* $X \in \mathcal{O}\backslash\{0\}$ *with* $\gamma(1) \in \Sigma_r$ *hits* Σ_r *orthogonally, i.e.* $\langle \dot\gamma(1), Y\rangle = 0$ *for all* $Y \in T_{\gamma(1)}\Sigma_r$.

Proof. Suppose $\exp_p(X) = q \in \Sigma_r$, meaning $\langle X, X\rangle = r$, and pick any $Y \in T_q\Sigma_r$. The latter can be realized as $Y = \partial_t f(1,0)$ for a smooth map of the form

$$f : [0, 1+\epsilon) \times (-\epsilon, \epsilon) \to M : (s,t) \mapsto \exp_p(sX(t)) \in \mathcal{U}$$

with $\epsilon > 0$ chosen sufficiently small and $X(t) \in T_pM$ a smooth path with $X(0) = X$ and $\langle X(t), X(t)\rangle = r$ for all t. The lemma will thus follow from the claim that for any map of this form,

$$\langle \partial_s f, \partial_t f\rangle \equiv 0.$$

When $s = 0$ this is immediate, because $f(0,t) = p$ for all t and thus $\partial_t f(0,t) = 0$. Using the properties of the Levi-Cività connection and the fact that $s \mapsto f(s,t) = \exp_p(sX(t))$ is a geodesic for each fixed t, we also have

$$\partial_s\langle \partial_s f, \partial_t f\rangle = \langle \nabla_s\partial_s f, \partial_t f\rangle + \langle \partial_s f, \nabla_s\partial_t f\rangle = \langle \partial_s f, \nabla_t \partial_s f\rangle. \tag{23.2}$$

Next observe that for each t, the "speed squared"[66] $\langle \partial_s f, \partial_s f\rangle$ of the geodesic $s \mapsto f(s,t)$ is a constant independent of s, because $\nabla_s\partial_s f \equiv 0$ implies $\partial_s\langle \partial_s f, \partial_s f\rangle \equiv 0$ and thus $\langle \partial_s f(s,t), \partial_s f(s,t)\rangle = \langle \partial_s f(0,t), \partial_s f(0,t)\rangle = \langle X(t), X(t)\rangle = r$. This proves

$$0 = \partial_t\langle \partial_s f, \partial_s f\rangle = 2\langle \nabla_t\partial_s f, \partial_s f\rangle,$$

so that (23.2) now vanishes, thus establishing that $\langle \partial_s f(s,t), \partial_t f(s,t)\rangle = \langle \partial_s f(0,t), \partial_t f(0,t)\rangle = 0$ for all (s,t). ☐

Remark 23.7. The $r = 0$ case of the Gauss lemma is vacuous when the metric is positive, and what it says in the indefinite case is slightly counterintuitive: observe that if $X \in T_pM$ is a nonzero vector with $\langle X, X\rangle = 0$, then also $\langle tX, tX\rangle = 0$ for every t and the geodesic $t \mapsto \exp(tX)$ is therefore *contained* in Σ_0, in addition to being (according to the statement of the proposition) orthogonal to it. This is not a contradiction, because while $\Sigma_0\backslash\{p\}$ is a well-defined submanifold of M, it is not what we would call a *pseudo-Riemannian submanifold* of (M,g), i.e. the restriction of $\langle\ ,\ \rangle$ to Σ_0 is degenerate and thus fails to be a pseudo-Riemannian metric. As a consequence, for $q \in \Sigma_0\backslash\{p\}$, the "orthogonal complement" $(T_q\Sigma_0)^\perp := \{Y \in T_qM \mid \langle Y, X\rangle = 0 \text{ for all } X \in T_q\Sigma_0\}$ has the correct dimension but is not actually *complementary* to $T_q\Sigma_0$, but is instead contained in it. The content of Proposition 23.6 is then that for each $q \in \Sigma_0$, $(T_q\Sigma_0)^\perp$ is the 1-dimensional subspace of $T_q\Sigma_0$ spanned by the tangent vector of a geodesic connecting p to q.

Proof of Theorem 23.5. Fix $\mathcal{O} \subset T_pM$ and $\mathcal{U} \subset M$ as in Proposition 23.6, assuming additionally that the metric $\langle\ ,\ \rangle$ is positive and $\mathcal{O}$ has the form of a ball,

$$\mathcal{O} = \{X \in T_pM \mid \langle X, X\rangle < R^2\}$$

for some $R > 0$. Given $q = \exp_p(X) \in \mathcal{U}\backslash\{p\}$ with $X \in \mathcal{O}\backslash\{0\}$, the geodesic segment $\gamma_0 : [0,1] \to \mathcal{U} : t \mapsto \exp_p(tX)$ has length $\ell_0^1(\gamma_0) = |X| =: \sqrt{r}$. For any other smooth path $\gamma : [0,1] \to \mathcal{U}$ from $\gamma(0) = p$ and $\gamma(1) = q$, let us assume after a small perturbation that $\gamma(t) \neq p$ for all $t \neq 0$, in which case we can write

$$\gamma(t) = \exp_p(\rho(t)X(t)) \qquad \text{for all } t \in (0,1],$$

with uniquely-determined smooth paths $\rho(t) > 0$ and $X(t) \in \Sigma_r$ satisfying $\lim_{t\to 0}\rho(t) = 0$, $\rho(1) = 1$ and $X(1) = X$. Write $f(s,t) = \exp_p(sX(t))$ as in the proof of Prop. 23.6, so the proposition implies $\langle \partial_s f, \partial_t f\rangle \equiv 0$, and since $s \mapsto f(s,t)$ is a geodesic with constant speed starting at $X(t)$

[66] The *speed* $|\dot\gamma(t)|$ of a geodesic γ only makes sense when the metric is positive, but "speed squared" $|\dot\gamma(t)|^2 := \langle \dot\gamma(t), \dot\gamma(t)\rangle$ can also be defined in the indefinite case, with the understanding that it might be negative.

for each t, $|\partial_s f(s,t)| = |X(t)| = \sqrt{r}$ for every t. Now since $\gamma(t) = f(\rho(t),t)$, we have $\dot\gamma(t) = \partial_s f(\rho(t),t)\dot\rho(t) + \partial_t f(\rho(t),t)$, and using the Pythagorean theorem,

$$|\dot\gamma(t)|^2 = |\partial_s f(\rho(t),t)\dot\rho(t)|^2 + |\partial_t f(\rho(t),t)|^2 \geqslant r|\dot\rho(t)|^2,$$

with strict inequality unless the path $X(t)$ is constant. The latter would mean $\gamma(t) = \exp_p(\rho(t)X) = \gamma_0(\rho(t))$, so that γ traces out the same image as γ_0, with a strictly longer length unless $\rho : (0,1] \to (0,1]$ is a diffeomorphism, which means γ is a reparametrization of γ_0. Now if $X(t)$ is not constant, we have

$$|\dot\gamma(t)| > \sqrt{r}|\dot\rho(t)| \geqslant \sqrt{r}\dot\rho(t) \qquad \text{for all } t \in (0,1]$$

and thus

$$\ell_0^1(\gamma) = \int_0^1 |\dot\gamma(t)|\,dt > \sqrt{r}\int_0^1 \dot\rho(t)\,dt = \sqrt{r} = \ell_0^1(\gamma_0).$$

This proves that all paths from p to q contained in $\mathcal{U}$ are strictly longer than the reparametrizations of γ_0. Any path that is *not* contained in $\mathcal{U}$ is obviously also longer, because it must cover a distance of at least $R > \sqrt{r}$ after starting at p before it can exit $\mathcal{U}$. □

REMARK 23.8. It is not straightforward to formulate variants of Theorem 23.5 with indefinite metrics, but on a pseudo-Riemannian manifold with Lorentz signature $(1, n-1)$ one can say the following. Recall from Remark 22.12 that the length functional is well-defined on time-like paths γ since their velocities satisfy $\langle\dot\gamma,\dot\gamma\rangle > 0$. It is not really appropriate to call it "length" in this situation, though; physicists prefer to call it the **proper time**, because in a Lorentzian 4-manifold representing spacetime, the proper time of a time-like path is the actual amount of time elapsed on a clock that is carried along that path through spacetime. Let us therefore denote

$$\tau_a^b(\gamma) := \int_a^b \sqrt{\langle\dot\gamma(t),\dot\gamma(t)\rangle}\,dt,$$

and consider the proper time of a time-like path from p to a point $q = \exp_p(X)$ that is "nearby" in the sense that $X \in T_pM$ is close to 0. The set of points q that can be reached in this way from p is called the **light cone** of p; it is an open subset bounded by light-like paths, i.e. paths that represent objects moving at the speed of light. An interesting detail arises here that is completely unlike anything in the Riemannian case: if you look at the standard Lorentzian inner product in an orthonormal basis so that it takes the form

$$\langle X, Y\rangle = X^1Y^1 - \sum_{j=2}^n X^jY^j,$$

you may notice that the set of time-like vectors (satisfying $\langle X,X\rangle > 0$) has two connected components, and as a result, the light cone of p is guaranteed to have two components if the neighborhoods $\mathcal{O} \subset T_pM$ and $\mathcal{U} \subset M$ are chosen sufficiently small. This is a symptom of the fact that in the physical world, there is a distinction between time-like paths moving forward or backward in time. We can therefore label the two components of the light cone C_p^+ and C_p^-, call them the *positive* and *negative* light cone respectively, and say $q \in C_p^+$ if and only if q is in the *future* of p.

One can now ask the following: how does the proper time of the geodesic segment $\gamma_0(t) = \exp_p(tX)$ compare with that of all other future-directed time-like paths from p to q?

The following detail is important to understand first: according to Proposition 23.6, time-like paths will pass orthogonally through hypersurfaces $\Sigma_r \subset \mathcal{U}$ with $r > 0$, and the restriction of the Lorentzian metric $\langle\ ,\ \rangle$ to these hypersurfaces is *negative*, i.e. it is $-h_r$ for a Riemannian metric h_r on Σ_r. One can deduce this from the fact that each geodesic $\gamma(t) = \exp(tX)$ hits Σ_r orthogonally: given that $\langle\dot\gamma(t),\dot\gamma(t)\rangle > 0$, the only way for $\langle\ ,\ \rangle$ to have signature $(1, n-1)$ at the intersection point is if it is negative-definite on $T\Sigma_r$.

Now, imitating the proof of Theorem 23.5, the assumption that $q = \exp_p(X) \in C_p^+$ implies $\langle X, X \rangle = [\tau_0^1(\gamma_0)]^2 =: r$, hence $q \in \Sigma_r$, and we can consider arbitrary paths from p to q of the form $\gamma(t) = \exp_p(\rho(t)X(t)) = f(\rho(t), t)$, where $f(s,t) = \exp_p(sX(t))$, $X(t) \in \Sigma_r$, $X(1) = X$, $\rho(1) = 1$ and $\lim_{t \to 0} \rho(t) = 0$. We still have $\langle \partial_s f, \partial_t f \rangle \equiv 0$, but the big difference from Theorem 23.5 is now that $\langle \partial_t f, \partial_t f \rangle \leqslant 0$, with strict inequality unless $\partial_t X = 0$, so our previous application of the Pythagorean theorem becomes

$$\langle \dot{\gamma}(t), \dot{\gamma}(t) \rangle = r|\dot{\rho}(t)|^2 - |\partial_t f(\rho(t), t)|^2 \leqslant r|\dot{\rho}(t)|^2,$$

again with equality only if $X(t)$ is constant. Requiring γ to be time-like then imposes the condition $r|\dot{\rho}(t)|^2 > |\partial_t f(\rho(t), t)| \geqslant 0$, so in contrast to the Riemannian case, we can only consider paths for which $\dot{\rho}(t) \geqslant 0$. The end result is that either γ is a reparametrization of γ_0 or

$$\tau_0^1(\gamma) = \int_0^1 \sqrt{\langle \dot{\gamma}(t), \dot{\gamma}(t) \rangle}\, dt < \sqrt{r} \int_0^1 \dot{\rho}(t)\, dt = \sqrt{r} = \tau_0^1(\gamma_0),$$

thus the geodesic from p to q *maximizes* the proper time among time-like paths from p to q.

One can use this calculation to explain the famous "twins paradox" in relativity—it is not a paradox, but merely a result of the fact that the proper time is not the same for all time-like paths between two points in spacetime. The scenario is that Albert and Henry are born at the same time, but Albert stays for his whole life on Earth, while Henry becomes an astronaut and travels several light-years across the universe and back, travelling at nearly the speed of light in both directions. On return, Henry has barely aged at all, but Albert is twenty years older. The reason is that by staying on Earth, Albert followed a geodesic in spacetime, but Henry did not: his path was at best a *piecewise* smooth geodesic, because he had to accelerate abruptly in order to reverse course and return to Earth. As a result, Albert's path experienced more proper time than Henry's.

REMARK. This lecture took place online and was interrupted multiple times by internet outages, as a result of which, there was no time to cover the two sections below. The contents of §23.3 will be covered briefly in the next online lecture, and §23.4 will be the main topic in one of this week's problem sessions.

23.3. Geodesic completeness. For an arbitrary pseudo-Riemannian manifold (M, g), the domain of the exponential map is an open subset $\mathcal{O} \subset TM$, and we say that (M, g) is **geodesically complete** if $\mathcal{O} = TM$. An equivalent condition is that for every $p \in M$ and $X \in T_pM$, the unique maximal geodesic $\gamma : (a, b) \to M$ with $\gamma(0) = p$ and $\dot{\gamma}(0) = X$ is defined for all time, i.e. the interval (a, b) is $\mathbb{R}$. It is easy to find examples for which this is not true, e.g. take $\mathbb{R}^n$ with a flat metric but remove a point to define $M := \mathbb{R}^n \backslash \{p\}$; then there exist geodesics in M that are not defined for all time because they collide at some time with the missing point p. In general, this would seem to be a danger whenever M is noncompact, because any given geodesic could potentially "escape to infinity" in finite time. We will see below that there is still a danger in general even when M is compact. On the other hand, $\mathbb{R}^n$ with a flat metric is an obvious example of a noncompact geodesically complete manifold: its geodesics are precisely the straight paths $\gamma(t) = \mathbf{v} + t\mathbf{w}$, which are defined for all $t \in \mathbb{R}$.

In order to say something general about completeness, it is useful to reformulate the problem in terms of the flow of a vector field. It will not be a vector field on M, because the geodesic equation is second order, so solutions are determined by more than just their initial position; the initial velocity is also required. This suggests defining a vector field instead on the tangent bundle TM, such as

$$\xi(X) := \mathrm{Hor}_X(X) \in T_X(TM), \tag{23.3}$$

where for $p \in M$ and $X \in T_pM$, $\mathrm{Hor}_X : T_pM \to T_X(TM)$ denotes the horizontal lift map for some connection ∇ on TM. Suppose $Y(t) \in TM$ is a flow line of $\xi \in \mathfrak{X}(TM)$, and using the bundle projection $\pi : TM \to M$, let $\gamma(t) := \pi \circ Y(t) \in M$, so γ is a path in M and Y is a vector field along γ. The condition $\partial_t Y(t) = \xi(Y(t)) = \mathrm{Hor}_{Y(t)}(Y(t))$ then implies

$$\dot{\gamma}(t) = \pi_* \left(\partial_t Y(t)\right) = \pi_* \mathrm{Hor}_{Y(t)}(Y(t)) = Y(t)$$

and, using the vertical projection $K : T(TM) \to TM$ to write the covariant derivative via (20.9),

$$\nabla_t Y(t) = K(\partial_t Y(t)) = 0.$$

In other words, Y is the velocity of γ and it is parallel along γ, hence γ is a geodesic. Conversely, if γ is any geodesic in M, then the path $Y(t) := \dot{\gamma}(t)$ in TM satisfies $\nabla_t Y(t) = 0$, implying that $\partial_t Y(t)$ is horizontal, which it means it can only be the horizontal lift of $\dot{\gamma}(t)$ and thus satisfies $\partial_t Y(t) = \xi(Y(t))$. This proves:

PROPOSITION 23.9. *Suppose ∇ is any connection on the tangent bundle $\pi : TM \to M$, and $\xi \in \mathfrak{X}(TM)$ is the vector field defined by (23.3) in terms of this connection. Then the exponential map has the same domain as the time 1 flow φ_ξ^1 of ξ, and $\exp = \pi \circ \varphi_\xi^1$.* □

The flow of the vector field ξ on TM is called the **geodesic flow** for M with connection ∇. It becomes an especially useful tool if we specialize to the Levi-Cività connection of a Riemannian metric:

THEOREM 23.10. *Every compact Riemannian manifold without boundary is geodesically complete.*

PROOF. Assuming (M, g) is a compact Riemannian manifold, we define the vector field $\xi \in \mathfrak{X}(TM)$ as in (23.3) using the Levi-Cività connection and notice that it has the following useful property: for each $r > 0$, ξ is tangent to the smooth hypersurface

$$S_r TM := \left\{ X \in TM \;\middle|\; \langle X, X \rangle = r^2 \right\}.$$

This follows from the fact that geodesics have constant speed, thus all flow lines of ξ are confined to hypersurfaces of this form.

Now observe that since M is compact, $S_r TM$ is also compact for every $r > 0$: indeed, the intersection of $S_r TM$ with each fiber T_pM is a compact $(n-1)$-sphere in T_pM, thus for any sufficiently small compact neighborhood $K \subset M$ of p, one can use a local trivialization of $TM|_K$ to show that $S_r TM \cap \pi^{-1}(K)$ is homeomorphic to the compact set $K \times S^{n-1}$. Then if $X_k \in S_r TM$ is any sequence and we write $p_k := \pi(X_k) \in M$, the compactness of M implies after restricting to a subsequence that p_k converges to some point $p \in M$, so that X_k for large k lies in a neighborhood homeomorphic to a compact set of the form $K \times S^{n-1}$ and therefore also has a convergent subsequence.

For any given $X \in TM$, one can now define $r := |X|$ and regard ξ as a vector field on the compact manifold $S_r TM$ instead of TM; since every vector field on a compact manifold has a global flow, the theorem follows. □

Theorem 23.10 depends rather crucially on the assumption that ∇ is the Levi-Cività connection for a *positive* metric g. A negative metric would also be fine, but the trouble with signatures (k, ℓ) with $k, \ell > 0$ is that while the geodesic flow $\xi \in \mathfrak{X}(TM)$ is tangent to hypersurfaces of the form

$$\left\{ X \in TM \;\middle|\; \langle X, X \rangle = c \right\}$$

for constants $c \in \mathbb{R}$, these hypersurfaces are not compact, even if M is. The problem is clearly visible if you look at the intersection of this hypersurface with a single fiber T_pM: choosing an orthonormal basis on T_pM so that $\langle X, Y \rangle = \sum_{j=1}^k X^j Y^j - \sum_{j=k+1}^n X^j Y^j$, the set of vectors

$X \in T_pM$ with $\langle X, X\rangle = c$ is not a sphere, it is a hyperboloid, which is definitely not compact. As such, there is no reason to expect the geodesic flow on TM to be globally defined, and in general, it is not. There are simple examples of indefinite pseudo-Riemannian manifolds that are compact but not geodesically complete.[67]

On the flip side, there are plenty of interesting Riemannian manifolds that are geodesically complete despite being noncompact; we will discuss some important examples in the next lecture. We do not have space here to prove the main result on this subject, but we plan to do so next semester, so consider the following statement a preview:

THEOREM (Hopf-Rinow theorem). *A connected Riemannian manifold (M, g) is geodesically complete if and only if it is a complete metric space with respect to the metric defined as the infimum of lengths of paths between points. Moreover, if it is complete, then for every pair of points $p, q \in M$, there exists a (not necessarily unique) geodesic segment from p to q that minimizes the length among all paths from p to q.*

23.4. Geodesics as a Hamiltonian system. The notion of the geodesic flow on TM can be placed into a wider context that connects it with symplectic geometry (cf. Lecture 14). To see this, we start with the observation that for any smooth manifold M, the cotangent bundle T^*M admits a canonical symplectic form. One defines it as follows: first let $\pi : T^*M \to M$ denote the bundle projection for the cotangent bundle, whose derivative gives a map $T\pi : T(T^*M) \to TM$ sending $T_\alpha(T^*M)$ linearly to T_qM for each $q \in M$ and $\alpha \in T_q^*M$. We can thus define a 1-form $\lambda \in \Omega^1(T^*M)$ by

$$\lambda_\alpha(\xi) := \alpha(T\pi(\xi)). \tag{23.4}$$

This is called the **tautological 1-form** on T^*M, and we will see below that

$$\omega := d\lambda \in \Omega^2(T^*M)$$

is a symplectic form. Recall from Lecture 14: this would mean that every point in T^*M has a neighborhood on which there exists a chart of the form $(p^1, q^1, \ldots, p^n, q^n)$ such that $\omega = \sum_{j=1}^n dp^j \wedge dq^j$. We claim in fact that any chart $(\mathcal{U}, (x^1, \ldots, x^n))$ on M naturally gives rise to a chart with this property on the open set $T^*M|_{\mathcal{U}} = \pi^{-1}(\mathcal{U}) \subset T^*M$. Indeed, we define n of the required $2n$ coordinates on $T^*M|_{\mathcal{U}}$ by

$$q^i := x^i \circ \pi, \qquad i = 1, \ldots, n.$$

For the remaining n coordinates, observe that the coordinates $x^1, \ldots, x^n$ give us a natural basis for each of the cotangent spaces over $\mathcal{U}$, namely the coordinate differentials, so let us define $p^1, \ldots, p^n$ on $T^*M|_{\mathcal{U}}$ by

$$(p^1, \ldots, p^n)\left(a_i\, dx^i\right) := (a_1, \ldots, a_n) \in \mathbb{R}^n.$$

Now observe: if a path $s(t) \in T^*M|_{\mathcal{U}}$ has constant coordinates $q^1, \ldots, q^n$, it means that $s(t)$ is moving within a single fiber, thus the velocity vectors $\dot{s}(t)$ belong to the vertical subbundle $V(T^*M) \subset T(T^*M)$, and in particular, this applies to the coordinate vector fields

$$\frac{\partial}{\partial p^1}, \ldots, \frac{\partial}{\partial p^n} \in V(T^*M).$$

On the other hand, for any path $s(t) \in T^*M|_{\mathcal{U}}$ whose coordinates are all constant except for one particular q^i, it follows that $\pi \circ s(t) \in \mathcal{U}$ has constant coordinates except for x^i, and thus

$$\pi_* \frac{\partial}{\partial q^i} = \frac{\partial}{\partial x^i} \qquad \text{for } i = 1, \ldots, n.$$

[67] See for instance the *Clifton-Pohl torus*: https://en.wikipedia.org/wiki/Clifton-Pohl_torus.

One sees now from the definition of $\lambda \in \Omega^1(T^*M)$ that it annihilates all vertical vectors, thus if we denote by $\alpha \in T^*M|_{\mathcal{U}}$ the point with some particular value of the coordinates $q^1, \ldots, q^n, p^1, \ldots, p^n$, then $\lambda_\alpha\left(\frac{\partial}{\partial p^i}\right) = 0$, and

$$\lambda_\alpha\left(\frac{\partial}{\partial q^i}\right) = \alpha\left(\pi_* \frac{\partial}{\partial q^i}\right) = \alpha\left(\frac{\partial}{\partial x^i}\right) = p^i.$$

The formula for λ in our chosen coordinates is therefore

$$\lambda = \sum_{i=1}^n p^i \, dq^i,$$

and it follows that $\omega = \sum_{i=1}^n dp^i \wedge dq^i$, so ω is symplectic.

Remark 23.11. Symplectic geometers sometimes abbreviate the tautological 1-form λ and symplectic form $\omega = d\lambda$ on T^*M by "$p\,dq$" and "$dp \wedge dq$" respectively, where the symbols p and q are each meant as shorthand for n separate coordinates. It is a somewhat remarkable fact that $p\,dq$ turns out to be the same 1-form no matter how one chooses the local coordinates $q^1, \ldots, q^n$; what makes this possible is the fact that while the coordinates $q^1, \ldots, q^n$ are arbitrarily chosen on some open subset of M, the remaining n coordinates $p^1, \ldots, p^n$ are not at all arbitrary, in fact they are completely determined by $q^1, \ldots, q^n$. One cannot assume in general that any of these coordinates are globally defined, but $p\,dq$ does make sense globally, because one can also express it as in (23.4) without choosing any coordinates.

The symplectic structure of T^*M provides a natural framework for viewing second-order dynamical systems on M as Hamiltonian systems on T^*M, and the geodesic flow is the simplest interesting example of this. In order to see it clearly, it will help to adopt the following notation: let us denote elements of T^*M as pairs

$$(q,p) \in T^*M, \qquad \text{where} \qquad q \in M \text{ and } p \in T_q^*M,$$

thus explicitly keeping track of horizontal motion via the symbol q and vertical motion via p. The easiest way to understand the tangent spaces $T_{(q,p)}(T^*M)$ then comes from choosing a connection ∇ on $\pi : T^*M \to M$, as it gives rise to a horizontal/vertical splitting

$$T_{(q,p)}(T^*M) = H_{(q,p)}(T^*M) \oplus V_{(q,p)}(T^*M)$$

such that $V_{(q,p)}(T^*M)$ is canonically isomorphic to the fiber T_q^*M and π_* gives a natural isomorphism of $H_{(q,p)}(T^*M)$ with the tangent space T_qM. The connection thus determines a natural isomorphism

$$T_{(q,p)}(T^*M) \cong T_qM \oplus T_q^*M,$$

and with this in mind, we shall write elements of $T_{(q,p)}(T^*M)$ as pairs of the form (Y, η) with $Y \in T_qM$ and $\eta \in T_q^*M$. For a path $\gamma(t) = (q(t), p(t)) \in T^*M$, the derivative $\dot\gamma(t) \in T_{\gamma(t)}(T^*M)$ is now written as a pair (Y, η) where $Y = \dot q(t) \in T_{q(t)}M$, and $\eta \in T_{q(t)}^*M$ is literally the projection of $\dot\gamma(t) \in T_{\gamma(t)}(T^*M)$ along the horizontal subspace to the vertical subspace, which means the covariant derivative, hence

$$\partial_t(q(t), p(t)) = (\dot q(t), \nabla_t p(t)) \in T_{q(t)}M \oplus T_{q(t)}^*M \cong T_{(q(t),p(t))}(T^*M).$$

Once one gets used to these natural isomorphisms, computations in T^*M become fairly straightforward.

Now suppose $g = \langle \ , \ \rangle$ is a pseudo-Riemannian metric on M, and the connection ∇ on $T^*M \to M$ is the connection induced on T^*M by the Levi-Cività connection of $TM \to M$. Using the musical isomorphisms to define a corresponding bundle metric on T^*M, Exercise 22.5

implies that our connection on $T^*M \to M$ is compatible with this bundle metric. The simplest Hamiltonian function that might be interesting to consider on T^*M is

$$H : T^*M \to \mathbb{R} : (q,p) \mapsto \frac{1}{2}\langle p, p\rangle. \tag{23.5}$$

There is some physical motivation to look at this particular function: in the special case of $M = \mathbb{R}^n$ with the standard Euclidean metric, H has an interpretation as the classical kinetic energy of a moving particle with mass 1. If we assume this is also the *total* energy, meaning there is no potential energy and thus no forces acting on the particle, then the motion of the particle is along straight lines in $\mathbb{R}^n$, and these are the geodesics in Euclidean space. It is not unreasonable to hope that the same correspondence might hold on a general pseudo-Riemannian manifold, and indeed:

PROPOSITION 23.12. *The Hamiltonian vector field for the function H in (23.5) is given by $X_H(q,p) = (p^\sharp, 0) \in T_qM \oplus T_q^*M \cong T_{(q,p)}(T^*M)$.*

Before proving the proposition, let us see what the flow of X_H looks like. For a path $\gamma(t) = (q(t), p(t)) \in T^*M$, $\dot\gamma(t) = X_H(\gamma(t))$ now means

$$\dot q(t) = p(t)^\sharp \qquad \text{and} \qquad \nabla_t p(t) = 0.$$

By Exercise 22.5, $\nabla_t \dot q = \nabla_t(p^\sharp) = (\nabla_t p)^\sharp = 0$, so this implies that the path $q(t)$ is a geodesic and $p(t)$ is simply the image of its velocity under the musical isomorphism $\flat : T_{q(t)}M \to T^*_{q(t)}M$. This proves a "Hamiltonian version" of the main result about the geodesic flow:

PROPOSITION 23.13. *Given a pseudo-Riemannian manifold (M, g) with Levi-Cività connection ∇ and the function $H : T^*M \to \mathbb{R}$ defined in (23.5) via the metric, the exponential map on TM is related to the flow of the Hamiltonian vector field X_H on T^*M by $\exp(Y) = \pi \circ \varphi^1_{X_H}(Y_\flat)$, where π is the bundle projection $T^*M \to M$.* □

Turning toward the proof of Proposition 23.12, it will be useful to have a coordinate-independent formula for ω that is more direct than calling it the exterior derivative of λ.

LEMMA 23.14. *Using the isomorphism $T_{(q,p)}(T^*M) \cong T_qM \oplus T_q^*M$, the canonical symplectic form ω on T^*M is given by*

$$\omega_{(q,p)}((Y,\eta),(Y',\eta')) = \eta(Y') - \eta'(Y).$$

PROOF. Using bilinearity and antisymmetry, it suffices to prove three more specific formulae:
(i) $\omega_{(q,p)}((0,\eta),(0,\eta')) = 0$ for all $\eta, \eta' \in T_q^*M$;
(ii) $\omega_{(q,p)}((Y,0),(Y',0)) = 0$ for all $Y, Y' \in T_qM$;
(iii) $\omega_{(q,p)}((Y,0),(0,\eta)) = -\eta(Y)$ for all $Y \in T_qM$ and $\eta \in T_q^*M$.

For all three, we will use the relation

$$d\lambda(\partial_s f, \partial_t f) = \partial_s\left[\lambda(\partial_t f)\right] - \partial_t\left[\lambda(\partial_s f)\right], \tag{23.6}$$

which is valid for any smooth map $\mathbb{R}^2 \overset{\text{open}}{\supset} \mathcal{V} \to M : (s,t) \mapsto f(s,t)$. Indeed, this is actually just a computation of $f^*d\lambda(\partial_s, \partial_t) = d(f^*\lambda)(\partial_s, \partial_t)$, and since the coordinate vector fields ∂_s and ∂_t on $\mathcal{V} \subset \mathbb{R}^2$ commute, the relation follows from our original definition of the exterior derivative in §8.2.

With this understood, let us first prove (i). Given $\eta, \eta' \in T_q^*M$, define

$$f(s,t) = (q, p + s\eta + t\eta') \in T^*M,$$

so $f(0,0) = (q,p)$, $\partial_s f(0,0) = (0,\eta)$ and $\partial_t f(0,0) = (0,\eta')$. Since $\partial_s f$ and $\partial_t f$ are both vertical vectors for every (s,t), λ annihilates them both and both terms in (23.6) therefore vanish, thus proving $d\lambda(\partial_s f, \partial_t f) = 0$.

Next is (ii): choose f in the form

$$f(s,t) = (\gamma(s,t), \sigma(s,t)) \in T^*M$$

such that $\gamma(0,0) = q$, $\partial_s\gamma(0,0) = Y$ and $\partial_t\gamma(0,0) = Y'$, and σ is a section of T^*M along γ that satisfies $\sigma(0,0) = p$ and has vanishing covariant derivative at $(s,t) = (0,0)$. (To see that the latter is possible, one can e.g. first define $\sigma(s,0)$ by parallel transporting p along the path $s \mapsto \gamma(s,0)$, then define $\sigma(s,t)$ by parallel transporting $\sigma(s,0)$ along the path $t \mapsto \gamma(s,t)$ for each fixed s.) We now have $\partial_s f(0,0) = (Y,0)$ and $\partial_t f(0,0) = (Y',0)$, and by (23.6),

$$d\lambda(\partial_s f, \partial_t f) = \partial_s\left[\sigma(\partial_t\gamma)\right] - \partial_t\left[\sigma(\partial_s\gamma)\right] = (\nabla_s\sigma)(\partial_t\gamma) - (\nabla_t\sigma)(\partial_s\gamma) + \sigma(\nabla_s\partial_t\gamma - \nabla_t\partial_s\gamma).$$

The last term in this expression vanishes identically because the Levi-Cività connection is symmetric, and the first two terms vanish specifically at $s = t = 0$ because $\nabla\sigma = 0$ at that point, so (i) is proven.

For (iii), we choose f in the form

$$f(s,t) = (\gamma(s), \sigma(s) + t\xi(s)) \in T^*M$$

such that $\gamma(0) = q$, $\gamma'(0) = Y$, and σ and ξ are parallel sections of T^*M along γ with $\sigma(0) = p$ and $\xi(0) = \eta$, thus $\partial_s f(0,0) = (Y,0)$ and $\partial_t f(0,0) = (0,\eta)$. Since $\partial_t f(s,t) = (0,\xi(s))$ is always vertical, $\lambda(\partial_t f) \equiv 0$, and (23.6) thus gives

$$d\lambda(\partial_s f(s,t), \partial_t f(s,t)) = -\partial_t\left[\lambda(\partial_s f(s,t))\right] = -\partial_t\left[(\sigma(s) + t\xi(s))(\gamma'(s))\right] = -\xi(s)\left(\gamma'(s)\right),$$

which is $-\eta(Y)$ at $s = 0$. □

PROOF OF PROPOSITION 23.12. The function $H(q,p) = \frac{1}{2}\langle p,p\rangle$ is constant in horizontal directions since parallel transport preserves the bundle metric, and in vertical directions, its differential is simply the differential at $(q,p) \in T^*M$ of the quadratic function $T_q^*M \to \mathbb{R} : p \mapsto \frac{1}{2}\langle p,p\rangle$, giving

$$dH(q,p)(Y,\eta) = \langle p,\eta\rangle.$$

Plugging $X_H(q,p) = (p^\sharp, 0)$ into the formula of Lemma 23.14 for ω gives

$$\omega(X_H(q,p),(Y,\eta)) = -\eta(p^\sharp) = -\langle \eta^\sharp, p^\sharp\rangle = -\langle p,\eta\rangle,$$

thus we've proven that X_H satisfies the defining equation $\omega(X_H,\cdot) = -dH$ of a Hamiltonian vector field. □

Proposition 23.12 opens the door toward using methods from symplectic geometry in the study of geodesics, and this is a fairly large topic in modern research. As a very simple illustration, we will now give a second proof of the result that compact Riemannian manifolds are geodesically complete. Recall that a map $\varphi : X \to Y$ between topological spaces is called **proper** if the preimage of every compact set is compact.

EXERCISE 23.15. Show that $H(q,p) = \frac{1}{2}\langle p,p\rangle$ is a proper function on T^*M if and only if the bundle metric $\langle\ ,\ \rangle$ is (positive or negative) definite.

The exercise combines with the following result to give another proof of Theorem 23.10.

THEOREM 23.16. *On any symplectic manifold (W,ω) with a smooth proper function $H : W \to \mathbb{R}$, the flow of the Hamiltonian vector field X_H exists globally.*

PROOF. One of the fundamental properties of Hamiltonian systems is that energy is conserved: "energy" in this case means the value of the Hamiltonian, and this value does not change along flow lines of X_H since

$$dH(X_H) = -\omega(X_H, X_H) \equiv 0.$$

It follows that every flow line of X_H stays within a level set $H^{-1}(c) \subset W$ for some $c \in \mathbb{R}$, and that set is compact if H is proper, thus the flow line can be continued for all time. $\square$

24. Euclidean and non-Euclidean geometries

In this lecture we will look at three specific examples of Riemannian manifolds whose properties lend considerable intuition to the rest of the subject. Our main goal for now will be to understand the behavior of the geodesics on these three examples, and certain qualitative differences will become apparent when we do this. We will later see that these differences are symptomatic of the distinction between positive, negative and zero curvature.

A bit of preparation is necessary before we discuss the actual examples, mainly because as a second-order nonlinear differential equation, the geodesic equation is generally not so easy to solve. We will first develop some tools that—at least in fortunate situations—make it easier.

24.1. Notation: how to write down a pseudo-Riemannian metric. In local coordinates $x^1, \ldots, x^n$, a pseudo-Riemannian metric $g = \langle \ , \ \rangle$ on a manifold M is a type $(0, 2)$ tensor field, and thus takes the form

$$g = g_{ij}\, dx^i \otimes dx^j,$$

where the components satisfy the relation $g_{ij} = g_{ji}$ since $\langle \ , \ \rangle$ is symmetric. One often sees this written in the form

$$g = \sum_{i \leqslant j} g_{ij}\, dx^i\, dx^j,$$

in which the summation avoids unnecessary repetition of matching components by using the abbreviation

$$dx^i\, dx^j := \frac{1}{2}\left(dx^i \otimes dx^j + dx^j \otimes dx^i\right). \tag{24.1}$$

So for example, the Euclidean metric on $\mathbb{R}^2$ can now be written in Cartesian coordinates (x, y) as

$$g_E = dx^2 + dy^2,$$

while the metric on the Poincaré half-plane in Exercise 22.8 becomes

$$h = \frac{1}{y^2}\left(dx^2 + dy^2\right).$$

On $\mathbb{R}^n$ in the standard coordinates $(x^1, \ldots, x^n)$, the Euclidean metric is now

$$g_E = (dx^1)^2 + \ldots + (dx^n)^2,$$

and changing some signs gives us the standard flat pseudo-Riemannian metric of signature (k, ℓ), written as

$$(dx^1)^2 + \ldots + (dx^k)^2 - (dx^{k+1})^2 - \ldots - (dx^n)^2.$$

Exercise 24.1. On $\mathbb{R}^2$ with coordinates (x, y), show that the pseudo-Riemannian metric $dx\, dy$ has signature $(1, 1)$, and find a new global coordinate system (s, t) in which it takes the form $ds^2 - dt^2$.

Remark 24.2. Algebraically, (24.1) can be regarded as a **symmetric product**, which is analogous to the wedge product but without all the minus signs. On an arbitrary vector space V, one can define a commutative product with values in $V \otimes V$ by symmetrizing the usual tensor product, thus writing $vw := \frac{1}{2}(v \otimes w + w \otimes v)$. The values of this product belong to the subspace consisting of symmetric bilinear maps $V^* \times V^* \to \mathbb{F}$, or equivalently, the kernel of the projection $\mathrm{Alt} : V \otimes V \to \Lambda^2 V$. If you don't care so much about algebra, don't worry about this.

24.2. Isometries and conformal transformations. A diffeomorphism $\varphi : M \to N$ from one pseudo-Riemannian manifold (M, g) to another (N, h) is called an **isometry** if

$$\varphi^* h = g.$$

We say in this situation that (M, g) and (N, h) are **isometric**, and indicate that φ is an isometry by writing

$$\varphi : (M, g) \to (N, h).$$

In more concrete terms, the condition means

$$h_{\varphi(p)}(\varphi_* X, \varphi_* Y) = g_p(X, Y) \qquad \text{for all } p \in M \text{ and } X, Y \in T_pM,$$

so in other words, the derivative $\varphi_* : T_pM \to T_{\varphi(p)}N$ of φ at every point $p \in M$ preserves the scalar products on these tangent spaces. In the Riemannian case (i.e. when the scalar products are positive), this has a simple geometric interpretation: one defines the lengths $|X|, |Y| \geqslant 0$ and angle $\theta \in [0, \pi]$ between two vectors $X, Y \in T_pM$ in this case by

$$|X| := \sqrt{\langle X, X \rangle}, \qquad |Y| := \sqrt{\langle Y, Y \rangle}, \qquad \theta = \arccos\left(\frac{\langle X, Y \rangle}{|X| \cdot |Y|}\right), \tag{24.2}$$

and preserving inner products thus means preserving lengths of tangent vectors and angles between them. It follows that a diffeomorphism is an isometry if and only if it preserves lengths of paths and angles between intersecting paths.

Isometry is the natural notion of equivalence in the category of pseudo-Riemannian manifolds, thus it preserves all meaingful notions that are defined in terms of pseudo-Riemannian metrics. For example, it preserves geodesics, i.e. if $\varphi : (M, g) \to (N, h)$ is an isometry, then a path $\gamma : (a, b) \to M$ is a geodesic if and only if $\varphi \circ \gamma : (a, b) \to N$ is a geodesic. One easy way to see this is via the energy functional from §22.4: it is straightforward to check that $E(\varphi \circ \gamma) = E(\gamma)$ for all paths γ in M, and that γ is therefore stationary for the energy functional on $C^\infty([a, b], M\,; p, q)$ if and only if $\varphi \circ \gamma$ is stationary for the energy functional on $C^\infty([a, b], N\,; \varphi(p), \varphi(q))$.

The set of all isometries $(M, g) \to (M, g)$ forms a group, denoted by

$$\operatorname{Isom}(M, g) \subset \operatorname{Diff}(M).$$

This group has the useful property that it maps geodesics to geodesics. On the other hand, one should not expect this group to be nontrivial in general, as preserving distances and angles turns out to be a very stringent condition on a diffeomorphism. One can show that $\operatorname{Isom}(M, g)$ is always a Lie group (see e.g. [**Kob95**]), so in particular, it is a smooth finite-dimensional manifold. The following result imposes an absolute upper bound on its dimension: if $\dim M = n$, then $\dim \operatorname{Isom}(M, g)$ can never be larger than

$$n + \dim \mathrm{O}(k, \ell) = n + \frac{1}{2}(n-1)n = \frac{1}{2}n(n+1). \tag{24.3}$$

This is, namely, the dimension of M plus the dimension of the space of all linear maps $T_pM \to T_qM$ for two points $p, q \in M$ that preserve the scalar product.

THEOREM 24.3. *Suppose (M, g) is a connected pseudo-Riemannian manifold, $p, q \in M$ are two points and $X_1, \dots, X_n \in T_pM$ and $Y_1, \dots, Y_n \in T_qM$ are orthonormal bases. Then there exists at most one isometry $\varphi \in \operatorname{Isom}(M, g)$ such that*

$$\varphi(p) = q \qquad \textit{and} \qquad \varphi_* X_i = Y_i \quad \textit{for all } i = 1, \dots, n.$$

PROOF. By looking at isometries of the form $\psi^{-1} \circ \varphi$, it is equivalent to show that the only isometry $f : (M, g) \to (M, g)$ satisfying $f(p) = p$ and $T_pf = \mathbb{1} : T_pM \to T_pM$ is the identity map. Since each geodesic through p is determined by its velocity at p, and f maps geodesics to geodesics, the condition $T_pf = \mathbb{1}$ implies that f is the identity map on the open neighborhood $\mathcal{U} \subset M$ of p

consisting of all points that can be reached via geodesics from p. Now suppose $q \in M$. Since M is connected, we can find a continuous path $\gamma : [0,1] \to M$ from $\gamma(0) = p$ to $\gamma(1) = q$, and the interval $[0,1]$ can then be partitioned into a finite union of subintervals with end points

$$0 =: t_0 < t_1 < \ldots < t_{N-1} < t_N := 1$$

such that for each $j = 1, \ldots, N$, $\gamma(t_j)$ can be reached via a geodesic starting at $\gamma(t_{j-1})$. It follows that f is also the identity on a neighborhood of $\gamma(t_1)$ and therefore $T_{\gamma(t_1)} f = \mathbb{1}$. Repeating the same argument N times then extends this conclusion to a neighborhood of $\gamma(t_N) = q$, and since the point q was arbitrary, f is therefore the identity map everywhere. □

Remark 24.4. Theorem 24.3 guarantees uniqueness, but not existence, thus (24.3) shows the dimension of $\mathrm{Isom}(M,g)$ if there exist as many isometries as the theorem allows, but in general $\dim \mathrm{Isom}(M,g)$ may be smaller. Once we have proved the basic theorems about curvature, it will begin to seem obvious that $\mathrm{Isom}(M,g)$ should be trivial for "most" pseudo-Riemannian manifolds, as the existence of nontrivial isometries will imply conditions on the curvature that are not usually satisfied.

You may recall from linear algebra that for two inner product spaces V and W, every linear map $A : V \to W$ that preserves lengths of vectors automatically preserves the inner product, and therefore also angles: this follows from the bilinearity of the inner product after expanding the relation $\langle A(v+w), A(v+w)\rangle = \langle v+w, v+w\rangle$. As a consequence, every smooth distance-preserving map between Riemannian manifolds is necessarily an isometry, and thus also an angle-preserving map. The converse however is false:

Lemma 24.5. *For two (positive) finite-dimensional inner product spaces V and W, a linear map $A : V \to W$ preserves angles if and only if there is a constant $c > 0$ such that $\langle Av, Aw\rangle = c\langle v, w\rangle$ for all $v, w \in V$.*

Proof. It is clear from (24.2) that the condition $\langle Av, Aw\rangle = c\langle v, w\rangle$ implies angles are preserved. Conversely, if $A : V \to W$ preserves angles, then it maps any orthonormal basis of V to a set of the form $\lambda_1 e_1, \ldots, \lambda_n e_n$ where the $\lambda_1, \ldots, \lambda_n$ are positive numbers and $e_1, \ldots, e_n$ is an orthonormal basis of W. After fixing appropriate bases, we can therefore assume without loss of generality that $V = W = \mathbb{R}^n$, both endowed with the standard Euclidean inner product, and A is represented by a diagonal matrix with positive entries $\lambda_1, \ldots, \lambda_n$. Writing $\mathbf{e}_1, \ldots, \mathbf{e}_n \in \mathbb{R}^n$ for the standard basis, the orthogonal vectors $\mathbf{e}_i + \mathbf{e}_j$ and $\mathbf{e}_i - \mathbf{e}_j$ for any $i \neq j$ must then be mapped by A to two orthogonal vectors, implying

$$0 = \langle A(\mathbf{e}_i + \mathbf{e}_j), A(\mathbf{e}_i - \mathbf{e}_j)\rangle = \langle \lambda_i \mathbf{e}_i + \lambda_j \mathbf{e}_j, \lambda_i \mathbf{e}_i - \lambda_j \mathbf{e}_j\rangle = \lambda_i^2 - \lambda_j^2,$$

and thus $\lambda_1 = \ldots = \lambda_n =: \lambda$. This proves $\langle Av, Aw\rangle = \lambda\langle v, w\rangle$ for all v, w. □

With this lemma in mind, a diffeomorphism $\varphi : M \to N$ is called a **conformal transformation** $(M,g) \to (N,h)$ if it satisfies

$$\varphi^* h = fg \qquad \text{for some smooth function } f : M \to (0,\infty),$$

where we should emphasize that the function f need not be specified in advance. This condition means that for every $p \in M$ and $X, Y \in T_pM$,

$$h_{\varphi(p)}(\varphi_* X, \varphi_* Y) = f(p) \cdot g_p(X,Y),$$

hence in the Riemannian case, one can say that the linear map $\varphi_* : T_pM \to T_{f(p)}N$ preserves angles (but not necessarily lengths), and the conformal transformations are therefore regarded as precisely those diffeomorphisms that preserve all angles between intersecting curves. The set of conformal transformations $(M, g) \to (M, g)$ also forms a group, denoted by

$$\mathrm{Conf}(M, g) \subset \mathrm{Diff}(M),$$

and it contains $\mathrm{Isom}(M, g)$ since every isometry is also a conformal transformation. The converse is false in general: for instance, for the Euclidean metric $g_E = dx^2 + dy^2$ and the Poincaré metric $h = \frac{1}{y^2}(dx^2 + dy^2)$ on the upper half-plane $\mathbb{H} = \{(x, y) \in \mathbb{R}^2 \mid y > 0\}$, the identity map $(\mathbb{H}, g_E) \to (\mathbb{H}, h)$ is conformal, but is not an isometry. This example also shows that conformal transformations do not preserve geodesics in general. (For the geodesics on $(\mathbb{H}, h)$, see Exercise 22.8.)

Conformal transformations arise naturally in complex analysis, due to the following exercise.

Exercise 24.6. Identify $\mathbb{C}$ with $\mathbb{R}^2$ via $x + iy \leftrightarrow (x, y)$ and endow it with the standard Euclidean metric. Show that a diffeomorphism $f : \mathcal{U} \to \mathcal{V}$ between two open subsets $\mathcal{U}, \mathcal{V} \subset \mathbb{C}$ is a conformal transformation if and only if it is either holomorphic or antiholomorphic (meaning the map $\bar{f} : \mathcal{U} \to \mathbb{C}$ is holomorphic). In particular, the group of orientation-preserving conformal transformations from an open region in $\mathbb{C}$ to itself is the same as its group of holomorphic automorphisms.

24.3. Pseudo-Riemannian submanifolds. Many interesting examples of Riemannian manifolds occur as hypersurfaces in flat space, so the question arises: if Σ is a submanifold of a pseudo-Riemannian manifold (M, g) whose geodesic flow we already understand, can we compute from it the geodesics on Σ? In fortunate cases this is possible, but there are a few subtleties to be aware of. First is the metric on Σ: we would obviously like to define it as the restriction of $g = \langle\ ,\ \rangle$ to $T\Sigma \subset TM$, or equivalently, the pullback $j^*g \in \Gamma(T^0_2\Sigma)$ via the inclusion map $j : \Sigma \hookrightarrow M$. This is fine if g is positive, because the restriction will then also satisfy $\langle X, X\rangle > 0$ for all nontrivial $X \in T\Sigma$, but in the indefinite case, the nondegeneracy of g does not immediately imply the same for its restriction j^*g. There is a simple exercise in linear algebra to be done before we continue.

Suppose V is a finite-dimensional real vector space and $\langle\ ,\ \rangle$ is a nondegenerate symmetric bilinear form on V; recall that "nondegenerate" in this situation means the map $V \to V^* : v \mapsto \langle v, \cdot\rangle$ is an isomorphism. By analogy with the case of a positive-definite inner product, we can associate to any linear subspace $W \subset V$ its orthogonal "complement"

$$W^\perp := \left\{v \in V \mid \langle v, w\rangle = 0 \text{ for all } w \in W\right\}.$$

We put the word "complement" in parentheses here because if $\langle\ ,\ \rangle$ is not positive-definite, there is no guarantee in general that W and $W^\perp$ will actually be complementary, i.e. they might have nontrivial intersection.

Lemma 24.7. *For any finite-dimensional real vector space V with a nondegenerate symmetric bilinear form $\langle\ ,\ \rangle$ and a subspace $W \subset V$:*

(1) $\dim W + \dim W^\perp = \dim V$,
(2) $(W^\perp)^\perp = W$,
(3) The restriction of $\langle\ ,\ \rangle$ to W is nondegenerate if and only if $W \cap W^\perp = \{0\}$, which is true if and only if $V = W \oplus W^\perp$.

Proof. Degeneracy of $\langle\ ,\ \rangle|_W$ means there exists a nontrivial vector $v \in W$ such that $\langle v, \cdot\rangle|_W = 0$, which is the same thing as saying $v \in W \cap W^\perp$. To show that $\dim W + \dim W^\perp = \dim V$, it suffices to view $W^\perp \subset V$ as the kernel of the linear map

$$V \to W^* : v \mapsto \langle v, \cdot\rangle|_W,$$

and observe that this map is surjective since every linear functional $\lambda : W \to \mathbb{R}$ can be extended to a linear functional on V and then presented as $\lambda = \langle v, \cdot\rangle$ for a unique $v \in V$, due to the nondegeracy of $\langle \ , \ \rangle$ on V. Since $W \subset (W^\perp)^\perp$ by definition, this also implies $W = (W^\perp)^\perp$, since both subspaces have the same dimension. $\square$

DEFINITION 24.8. In a pseudo-Riemannian manifold (M, g), a submanifold $\Sigma \subset M$ with inclusion map $j : \Sigma \hookrightarrow M$ is called a **pseudo-Riemannian submanifold** if j^*g is nondegenerate, so that it defines a pseudo-Riemannian metric on Σ. We call Σ a **Riemannian submanifold** if j^*g is positive.

Lemma 24.7 implies:

COROLLARY 24.9. *A submanifold Σ in (M, g) is a pseudo-Riemannian submanifold if and only if for every $p \in \Sigma$, $T_pM = T_p\Sigma \oplus (T_p\Sigma)^\perp$.* $\square$

The condition in Corollary 24.9 is satisfied for *every* submanifold $\Sigma \subset M$ if (M, g) is a Riemannian manifold, but this is not true in the indefinite case. For example, light-like paths (see Remark 22.12) in a Lorentzian manifold (M, g) trace out smooth 1-dimensional submanifolds $\Sigma \subset M$, but (Σ, j^*g) is not a pseudo-Riemannian submanifold, as j^*g in this case vanishes.

REMARK 24.10. Lemma 24.7 remains true without significant changes if $\langle \ , \ \rangle$ is assumed antisymmetric instead of symmetric, and this observation is important in symplectic geometry. In particular, an analogue of Corollary 24.9 holds for symplectic submanifolds of a symplectic manifold.

By Corollary 24.9, every pseudo-Riemannian submanifold $\Sigma \subset (M, g)$ comes with a well-defined **orthogonal projection**

$$\pi_\Sigma : TM|_\Sigma \to T\Sigma,$$

which projects each tangent space T_pM for $p \in \Sigma$ to $T_p\Sigma \subset T_pM$ along the complementary subspace $(T_p\Sigma)^\perp \subset T_pM$.

PROPOSITION 24.11. *If ∇ is the Levi-Cività connection on (M, g) and $\Sigma \subset M$ is a pseudo-Riemannian submanifold with inclusion $j : \Sigma \hookrightarrow M$, the Levi-Cività connection on (Σ, j^*g) is uniquely determined by the relation*

$$\hat{\nabla}_X Y = \pi_\Sigma(\nabla_X Y), \qquad \text{for } p \in \Sigma,\ X \in T_p\Sigma \text{ and } Y \in \mathfrak{X}(M) \text{ with } Y(\Sigma) \subset T\Sigma.$$

PROOF. Any vector field on Σ near p can be extended to a vector field on M using a slice chart, thus the stated relation uniquely determines a connection on Σ if we can prove that the operator $\pi_\Sigma \circ \nabla_X$ satisfies the required Leibniz rule. And it does: for $f \in C^\infty(\Sigma)$, $Y \in \mathfrak{X}(\Sigma)$ and $X \in T_p\Sigma$, we extend f and Y arbitrarily to a smooth function and vector field respectively on M, and use the Leibniz rule for ∇ to compute

$$\pi_\Sigma(\nabla_X(fY)) = \pi_\Sigma((\mathcal{L}_X f)Y + f\nabla_X Y) = (\mathcal{L}_X f)\pi_\Sigma(Y) + f\pi_\Sigma(\nabla_X Y) = (\mathcal{L}_X f)Y + f\pi_\Sigma(\nabla_X Y),$$

where we have written $\pi_\Sigma(Y) = Y$ since $Y(\Sigma) \subset T\Sigma$. This proves that $\pi_\Sigma \circ \nabla$ defines a connection on Σ, and to see that it is also compatible with the restricted metric j^*g, we take two vector fields $Y, Z \in \mathfrak{X}(\Sigma)$, extend them smoothly to vector fields on M, and then use the fact that ∇ is compatible with g, plus the fact that $Y(p)$ and $Z(p)$ are both orthogonal to $(T_p\Sigma)^\perp$:

$$\mathcal{L}_X \langle Y, Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle = \langle \pi_\Sigma(\nabla_X Y), Z\rangle + \langle Y, \pi_\Sigma(\nabla_X Z)\rangle.$$

Finally, we observe that for vector fields $Y, Z \in \mathfrak{X}(M)$ with values in $T\Sigma$ along Σ, the Lie bracket $[Y, Z] \in \mathfrak{X}(M)$ necessarily also has this property, so the torsion of $\pi_\Sigma \circ \nabla$ at $p \in \Sigma$ is

$$\begin{aligned}\hat{T}(Y(p), Z(p)) &:= \pi_\Sigma(\nabla_{Y(p)} Z) - \pi_\Sigma(\nabla_{Z(p)} Y) - [Y, Z](p) = \pi_\Sigma\left(\nabla_{Y(p)} Z - \nabla_{Z(p)} Y - [Y, Z](p)\right) \\ &= \pi_\Sigma(T(Y(p), Z(p))) = 0,\end{aligned}$$

since ∇ is symmetric. The result now follows from the uniqueness of the Levi-Cività connection. $\square$

COROLLARY 24.12. *Assume $\Sigma \subset (M,g)$ is a pseudo-Riemannian submanifold with inclusion $j : \Sigma \hookrightarrow M$, and ∇ denotes the Levi-Cività connection on (M,g). Then a path $\gamma : (a,b) \to \Sigma$ is a geodesic on (Σ, j^*g) if and only if $\nabla_t \dot\gamma(t)$ is orthogonal to $T_{\gamma(t)}\Sigma$ for all t.*

PROOF. According to Proposition 24.11, the geodesic equation on (Σ, j^*g) is $\pi_\Sigma(\nabla_t\dot\gamma) \equiv 0$. $\square$

24.4. Three examples of Riemannian manifolds.

24.4.1. *Euclidean space.* We have already mentioned that the Christoffel symbols on $M := \mathbb{R}^n$ with the Euclidean metric

$$g = g_E := (dx^1)^2 + \ldots + (dx^n)^2$$

vanish identically, thus the geodesic equation becomes $\ddot\gamma = 0$ and the geodesics are straight lines. You may think there is not much more to say about this example, but that didn't stop Euclid from writing a treatise about $(\mathbb{R}^2, g_E)$ that was regarded as the basis of Western mathematics for 2000 years. Here is a modern reformulation of the first two of Euclid's five postulates, on which all of his propositions about plane geometry are based:

(E1) For every pair of distinct points $p, q \in M$, there exists a unique geodesic segment $\gamma : [0,1] \to M$ with $\gamma(0) = p$ and $\gamma(1) = q$.
(E2) Every geodesic in (M,g) exists for all time, i.e. (M,g) is geodesically complete.

Before continuing, let us mention another property of $(\mathbb{R}^n, g_E)$ that Euclid uses constantly without mentioning it, but that is actually a quite nontrivial property for a Riemannian manifold to have. The isometry group $\mathrm{Isom}(\mathbb{R}^n, g_E)$ is as large as possible, i.e. every isometry that is permitted by Theorem 24.3 actually exists. Indeed, the isometry group of $(\mathbb{R}^n, g_E)$ contains all the translations $\mathbf{x} \mapsto \mathbf{x} + \mathbf{v}$ by vectors $\mathbf{v} \in \mathbb{R}^n$, as well as the orthogonal transformations $\mathbf{A} \in \mathrm{O}(n)$, and one can combine these to produce a transformation that takes any given point p to another given point q while effecting an arbitrary rotation or reflection on their tangent spaces. In Euclid's argumentation, this fact is used for congruence proofs, e.g. two triangles in $\mathbb{R}^2$ are seen to be "the same" because one can be overlaid upon another, which means in modern terms that there is an isometry $\mathbb{R}^2 \to \mathbb{R}^2$ mapping one to the other. One can use this to justify the notion that "all right angles are the same," which is essentially the content of Euclid's fourth postulate (E4): if α_1, α_2 are two geodesics that intersect at a right angle at $\alpha_1(s_1) = \alpha_2(s_2) =: p$ and β_1, β_2 is another pair of geodesics with a right-angle intersection at $\beta_1(t_1) = \beta_2(t_2) =: q$, then there exists an isometry sending $p \mapsto q$, $\alpha_1(\mathbb{R}) \mapsto \beta_1(\mathbb{R})$ and $\alpha_2(\mathbb{R}) \mapsto \beta_2(\mathbb{R})$. This property is the reason why angles in the plane can be measured and meaningfully compared, even if they appear at different points. Euclid's third postulate (E3) is not so much a *property* of $(\mathbb{R}^2, g_E)$ as a "recipe" for constructing circles, which in our n-dimensional context would mean spheres: for every pair of distinct points p, q, there is a unique "$(n-1)$-sphere" centered at p containing q, which we would define as

$$\left\{ \exp_p(X) \;\middle|\; X \in T_pM \text{ such that } |X| = |X_0| \text{ where } \exp_p(X_0) = q \right\}.$$

Note that the vector $X_0 \in T_pM$ in this definition is unique due to the uniqueness of the geodesic segment in (E1), and $\exp_p(X)$ is always well defined due to (E2). The point of (E3) is that it gives rise to *constructive* arguments, e.g. Euclid's proposition on bisecting triangles provides not just the *existence* of bisections but an actual *recipe* to construct them with a ruler and compass.

The most famous of Euclid's postulates is the fifth, which is better known in a reformulation that was stated by John Playfair in 1795 and shown to be equivalent to Euclid's fifth postulate whenever the first four also hold:

(E5) For any geodesic $\gamma : \mathbb{R} \to M$ and a point $p \in M$ not on the image of γ, there exists at most one (up to parametrization) geodesic through p that does not intersect γ.

This is the **parallel postulate**, and historically, it has caused a lot of trouble. We'll come back to that shortly.

24.4.2. *Spheres.* The natural Riemannian metric on the unit sphere $S^n \subset \mathbb{R}^{n+1}$ is the one that it inherits by restriction from the Euclidean metric on $\mathbb{R}^{n+1}$. For the latter, the Levi-Cività connection ∇ is the trivial one, thus according to Corollary 24.12, a path $\gamma : (a,b) \to S^n$ is a geodesic in S^n if and only if

$$\ddot{\gamma}(t) \in (T_{\gamma(t)}S^n)^\perp = \mathbb{R}\gamma(t) \qquad \text{for all } t.$$

It is easy to find paths that have this property, e.g. for any $p \in S^n$ and $\mathbf{v} \in T_pS^n = p^\perp$ with $|\mathbf{v}| = 1$, the path

$$\gamma(t) = (\cos t)p + (\sin t)\mathbf{v} \in S^n \subset \mathbb{R}^{n+1}$$

is an example since $\ddot{\gamma}(t) = -\gamma(t) \in \mathbb{R}\gamma(t)$ for all t. Geodesics of this form exist for all $t \in \mathbb{R}$, and they can be chosen so that $\gamma(0) = p$ is an arbitrary point in S^n and $\dot{\gamma}(0) = \mathbf{v}$ is an arbitrary unit vector in $T_{\gamma(0)}S^n$. It follows that *all* geodesics on S^n are either paths of this form or (depending on their speed) reparametrizations of them: their images are the intersections of S^n with arbitrary 2-dimensional subspaces (spanned by the vectors $p, \mathbf{v} \in \mathbb{R}^{n+1}$), and are known as **great circles**.

Just like Euclidean space, the sphere S^n has the largest possible isometry group: any matrix in $\mathrm{O}(n+1)$ defines a transformation $\mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ that preserves S^n. If $\mathbf{e}_1 \in \mathbb{R}^{n+1}$ denotes the first standard basis vector, then for any other $\mathbf{v} \in S^n$, one can find $\mathbf{A} \in \mathrm{O}(n+1)$ with $\mathbf{A}\mathbf{e}_1 = \mathbf{v}$ by defining the columns of $\mathbf{A}$ to be any orthonormal basis $\mathbf{v}_1, \ldots, \mathbf{v}_{n+1}$ with $\mathbf{v}_1 = \mathbf{v}$. This construction allows considerable freedom in the choice of $\mathbf{v}_2, \ldots, \mathbf{v}_{n+1}$, and this freedom is sufficient to realize any desired orthogonal transformation on the subspace $T_{\mathbf{v}}S^n = \mathbf{v}^\perp$.

Let's see how Euclid's axioms are doing. All the geodesics mentioned above are defined for all $t \in \mathbb{R}$, so (E2) is fine. There is a problem with (E1), though: while it is certainly possible to connect any two distinct points $p, q \in S^n$ by a geodesic segment, this segment is *never* unique: every geodesic on S^n is periodic, so you can always find another segment from p to q just by traversing the circle more times. In some cases you can find a lot more: for instance, antipodal points on S^2 are connected by an infinite family of geodesics, e.g. the longitudes that connect the north and south poles on the Earth. Another consequence of this ambiguity is that a geodesic from p to q is definitely *not* always the shortest path on S^n from p to q, nor must it be a local minimum of the length functional: if you imagine for instance a path that traverses most of a great circle in order to move from p to a nearby point q, it is easy to find non-geodesic paths nearby that are shorter. We proved in §22.4 that geodesics are stationary (i.e. critical points) for the length functional, but indeed, not every critical point must be a local minimum.

On S^2, the parallel postulate is true for a stupid reason: no two geodesics are parallel, i.e. they always must intersect! In summary, classical geometry on S^2 is an interesting subject, but it has very little to do with Euclid's postulates.

24.4.3. *Hyperbolic space.* The third example gives a reason to care about indefinite metrics even if you have no interest in physics and really just want to understand Riemannian manifolds. The idea is to do the same thing as in the previous subsection, but with the Euclidean metric on $\mathbb{R}^{n+1}$ replaced by a metric with Lorentz signature: we will call it the **Minkowski metric**, and write it in coordinates $X = (\tau, x^1, \ldots, x^n) = (\tau, \mathbf{x}) \in \mathbb{R}^{n+1}$ as

$$g_M := -d\tau^2 + (dx^1)^2 + \ldots + (dx^n)^2.$$

The sphere was obtained as a regular level set for the Euclidean metric, but using the Minkowski metric instead gives a hyperboloid:

$$\left\{X \in \mathbb{R}^{n+1} \;\middle|\; \langle X, X\rangle = -1\right\} = \left\{(\tau, \mathbf{x}) \in \mathbb{R} \times \mathbb{R}^n \;\middle|\; \tau^2 - |\mathbf{x}|^2 = 1\right\}.$$

This hypersurface has two connected components, distinguished by the conditions $\tau \geqslant 1$ and $\tau \leqslant -1$, so we pick one of them to define a connected n-manifold called **hyperbolic n-space**

$$H^n := \left\{X = (\tau, \mathbf{x}) \in \mathbb{R}^{n+1} \;\middle|\; \tau^2 - |\mathbf{x}|^2 = 1 \text{ and } \tau > 0\right\}.$$

We claim that this is in fact a *Riemannian* submanifold of $(\mathbb{R}^{n+1}, g_M)$, i.e. the restriction of the Minkowski metric to H^n is positive-definite. To see this, note that as (a component of) a regular level set of the function $f(X) := \langle X, X\rangle$, the tangent space to H^n at any point $p \in H^n$ is the kernel of $Df(p) : \mathbb{R}^{n+1} \to \mathbb{R}$, where the latter is $Df(p)Y = 2\langle p, Y\rangle$, hence

$$T_pH^n = p^\perp \subset \mathbb{R}^{n+1}.$$

One needs to be careful not to use too much Euclidean intuition in reading equations like this: the symbol $\perp$ in this case is defined relative to the Minkowski metric, which is indefinite, so it is not even automatic that $p \notin p^\perp$. On the other hand, the Minkowski inner product is negative (and therefore nondegenerate) on the 1-dimensional subspace spanned by p, so it follows from Lemma 24.7 that $\mathbb{R}p \oplus p^\perp = \mathbb{R}^{n+1}$. Since $\mathbb{R}p = (T_pH^n)^\perp$, Corollary 24.12 then implies that $H^n \subset (\mathbb{R}^{n+1}, g_M)$ is a pseudo-Riemannian submanifold. Its signature can be deduced from the fact that g_M has signature $(n,1)$ and is negative on $(TH^n)^\perp$: this is only possible if g_M restricts positively to TH^n. We therefore have a natural Riemannian metric on H^n.

Remark 24.13. You may have wondered why we defined H^n as a component of the level set with $\langle X, X\rangle = -1$ instead of $\langle X, X\rangle = 1$, as the latter might have seemed more obviously analogous to the sphere. The reason is that we specifically wanted a *Riemannian* submanifold: the hyperboloid $\langle X, X\rangle = 1$ is also a pseudo-Riemannian submanifold, one that even has the advantage of being connected, but it has signature $(n-1, 1)$.

What are the geodesics? Here it is useful to note that the Levi-Cività connection ∇ on Minkowski space is the same one as on Euclidean space: it is the trivial connection, as is true for every pseudo-Riemannian metric with constant coefficients. One can then write down the geodesics on H^n in almost exactly the same way as on S^n, the trick is just to replace cos and sin by their hyperbolic counterparts. Given any $p \in H^n$ and $\mathbf{v} \in T_pH^n = p^\perp \subset \mathbb{R}^{n+1}$ with $|\mathbf{v}| = \sqrt{\langle \mathbf{v}, \mathbf{v}\rangle} = 1$, the path

$$\gamma(t) := (\cosh t)p + (\sinh t)\mathbf{v} \in \mathbb{R}^{n+1}$$

satisfies

$$\begin{aligned}\langle \gamma(t), \gamma(t)\rangle &= \langle (\cosh t)p + (\sinh t)\mathbf{v}, (\cosh t)p + (\sinh t)\mathbf{v}\rangle = (\cosh^2 t)\langle p, p\rangle + (\sinh^2 t)\langle \mathbf{v}, \mathbf{v}\rangle \\ &= -\cosh^2 t + \sinh^2 t = -1,\end{aligned}$$

so it lies in H^n, and its image is the intersection of H^n with the 2-dimensional subspace of $\mathbb{R}^{n+1}$ spanned by p and $\mathbf{v}$. Moreover,

$$\ddot{\gamma}(t) = \gamma(t) \in (T_{\gamma(t)}H^n)^\perp,$$

so Corollary 24.12 implies that γ is a geodesic. Since $\gamma(0) = p \in H^n$ and $\dot{\gamma}(0) = \mathbf{v} \in T_pH^n$ can each be chosen arbitrarily (subject to the condition $|\mathbf{v}| = 1$), every geodesic in H^n is a reparametrization of one of these.

And the isometries? The group of linear transformations on $\mathbb{R}^{n+1}$ preserving the Minkowski metric is the Lorentz group $\mathrm{O}(n,1)$, and its transformations preserve the submanifold $H^n \subset \mathbb{R}^{n+1}$. Analogously to the action of $\mathrm{O}(n+1)$ on S^n, one can show that there is a Lorentz transformation sending any point in H^n to any other one, while realizing any desired rotation or reflection on the tangent spaces. The isometry group of H^n is therefore as large as possible: in particular, for any two geodesics on H^n with the same speed, there exists an isometry identifying one with the other.

The **hyperbolic plane** H^2 made a splash when it was first discovered in the 19th century. The reason has to do with Euclid's postulates: H^2 satisfies the first four, so a large portion of Euclid's propositions on congruence, bisection of triangles etc. works just as well in hyperbolic as in Euclidean geometry. But not the fifth postulate:

Exercise 24.14. Find a pair of intersecting geodesics on H^2 and a third geodesic that intersects neither of them.

The parallel postulate was always perceived to be a less obviously "fundamental" statement than Euclid's first four postulates, and the belief remained popular for 2000 years after Euclid that it should be possible to deduce it logically from the other four, if only one could find the right argument. Several illustrious figures even claimed at various times to have achieved this, though their proofs invariably turned out to rely on unjustified intuitive assumptions that do not follow from the first four postulates. (For more on this history, see [**Lee13b**].) The example of the hyperbolic plane revealed finally that this effort was fruitless: the fifth postulate cannot be deduced from the other four, because there exists a geometry that satisfies those four but not the fifth.

Exercise 24.15. Let $B^n \subset \mathbb{R}^n$ denote the open ball of radius 1. There is a natural diffeomorphism $\varphi : B^n \to H^n$ defined via *stereographic projection*, which means the following: for $\mathbf{x} \in B^n$, define $\varphi(\mathbf{x}) \in H^n$ as the unique intersection of H^n with the line in $\mathbb{R}^{n+1} = \mathbb{R} \times \mathbb{R}^n$ that passes through the points $(-1, 0)$ and $(0, \mathbf{x})$. The pullback $\varphi^* g_M$ thus defines a Riemannian metric on B^n making it isometric to H^n. Prove $\varphi^* g_M$ is related to the Euclidean metric $g_E = (dx^1)^2 + \ldots + (dx^n)^2$ by

$$\varphi^* g_M = \frac{4}{(1 - |\mathbf{x}|^2)^2} g_E.$$

This is called the *Poincaré disk model* of hyperbolic space.

The Poincaré disk model in Exercise 24.15 reveals that hyperbolic space is conformally flat, i.e. the metric $\varphi^* g_M$ on B^n defines the same notion of angles as the Euclidean metric. This observation becomes especially useful in the case $n = 2$, where we can use the bijection $\mathbb{R}^2 \ni (x, y) \leftrightarrow x + iy =: z \in \mathbb{C}$ to identify B^2 with

$$D := \left\{ z \in \mathbb{C} \;\middle|\; |z| < 1 \right\}$$

and write $\varphi^* g_M = \frac{4}{(1-|z|^2)^2} \left(dx^2 + dy^2 \right)$.

Exercise 24.16. The classical *Cayley transform* is the holomorphic map $f(z) := \frac{z-i}{z+i}$, which defines a conformal transformation from the open upper half-plane $\mathbb{H} := \left\{ x + iy \in \mathbb{C} \;\middle|\; y > 0 \right\}$ to D. Prove

$$f^* \varphi^* g_M = \frac{1}{y^2} \left(dx^2 + dy^2 \right),$$

hence the Poincaré half-plane from Exercise 22.8 is another model of the hyperbolic plane.

By a standard theorem in complex analysis, the group of holomorphic automorphisms of the disk $D \subset \mathbb{C}$ consists of all maps of the form

$$z \mapsto e^{i\theta} \frac{z - a}{1 - \bar{a} z}, \qquad \text{for any } \theta \in \mathbb{R},\ a \in D.$$

Or if you prefer the Poincaré half-plane model, the holomorphic automorphisms of $\mathbb{H}$ are the fractional linear transformations

$$z \mapsto \frac{az + b}{cz + d}, \qquad \text{for any } a, b, c, d \in \mathbb{R} \text{ with } ad - bc = 1.$$

These are two alternate perspectives on the same thing, and in either case, we have a 3-dimensional group of conformal transformations, containing exactly one that maps any given point to any other given point while also realizing any desired rotation. But since both of these are isometric to the hyperbolic plane, we can say the same thing about the orientation-preserving isometries: all of the latter are of course conformal transformations, and they are therefore *all* of the conformal transformations. This proves a rather surprising fact about the hyperbolic plane:[68]

THEOREM 24.17. *On H^2, every conformal transformation is an isometry.* □

This result plays a fundamental role in the theory of Riemann surfaces, due to the fact that choosing a complex structure on a surface is equivalent to choosing an orientation and a conformal structure, i.e. a conformal equivalence class of metrics. It implies that outside of a finite set of exceptions, the category of Riemann surfaces is essentially equivalent to the category of oriented surfaces with hyperbolic metrics, so that results from 2-dimensional Riemannian geometry have nontrivial consequences for complex 1-manifolds.

One of the standard theorems derivable from Euclid's five postulates is that the sum of the angles in every triangle is π. This is one of the things you lose if you remove the fifth postulate:

EXERCISE 24.18. Using whichever model you prefer, show that for any $\epsilon > 0$, H^2 contains a compact region bounded by three geodesics, each intersecting each of the others exactly once, such that the sum of the angles at the three intersections is less than ϵ.

25. Integrability and the Frobenius theorem

In this lecture we begin talking about curvature: we will consider first the setting of a general vector bundle with an arbitrary connection, and once this is understood, specialize to the tangent bundle of a pseudo-Riemannian manifold with the Levi-Cività connection. We assume as usual that

$$\pi : E \to M$$

is a smooth vector bundle, and the symbol ∇ will always mean a connection on this bundle.

25.1. Flat sections and connections. One can motivate the topic of curvature by asking three questions whose answers in the setting of ordinary differentiation (i.e. for the trivial connection on a trivial bundle) are either obvious or are well-known results from first-year analysis. The answers turn out to be much less obvious for an arbitrary connection ∇ on E.

QUESTIONS 25.1. *Choose any point $p \in M$ in the base of the vector bundle $\pi : E \to M$.*

(1) *Given $v \in E_p$, is v the value at p of any* parallel *section $s : \mathcal{U} \to E$ defined on a neighborhood $\mathcal{U} \subset M$ of p , i.e. a section satisfying $\nabla s \equiv 0$?*
(2) *Does p have a neighborhood $\mathcal{U} \subset M$ on which for every smooth path $\gamma : [0,1] \to \mathcal{U}$ with $\gamma(0) = \gamma(1) = p$, the parallel transport map $P^1_\gamma : E_p \to E_p$ is the identity?*
(3) *Given a coordinate chart $(x^1, \ldots, x^n)$ on a neighborhood of p, do the partial covariant derivative operators $\nabla_i := \nabla_{\frac{\partial}{\partial x^i}}$ and $\nabla_j := \nabla_{\frac{\partial}{\partial x^j}}$ for $i \neq j$ commute at p?*

The answer to all three questions is clearly yes if ∇ is the trivial connection with respect to some local trivialization of E near p. This is always the case if $\dim M = 1$, in particular, since p then has a neighborhood parametrized by a path, so parallel transport along that path can be used to define a trivialization in which the parallel sections are represented by constant functions, hence ∇ is the trivial connection. But for $\dim M \geqslant 2$, we will see that the answer to all three questions is no in general.

If you think of parallel sections as the generalization to vector bundles of the notion of a constant function, then it seems surprising at first that there might not exist one on any neighborhood of a point. Of course, parallel sections *along a path* do always exist; we get them from parallel transport. But if $\dim M \geqslant 2$ so that no neighborhood of p can be parametrized by a single path,

[68]There was no time to mention Theorem 24.17 in the lecture, so it is included here only for information.

then the effort to find a parallel section runs into trouble precisely because the answer to question 25.1(2) might be no: if a parallel section $s : \mathcal{U} \to E$ on some neighborhood $\mathcal{U} \to M$ of p exists with any given value $s(p) = v$, then paths $\gamma : [0,1] \to \mathcal{U}$ will satisfy $P^t_\gamma(v) = s(\gamma(t))$ for every t, and parallel transport along a ***loop*** in $\mathcal{U}$ therefore always brings us back to v. But we've already seen an example where the latter is impossible: parallel transport using the Levi-Cività connection on $TS^2 \to S^2$ along certain closed "triangular" paths in S^2 does not produce the identity map; see Figure 8 in Lecture 19. (You've learned in the mean time that the edges of the triangle in that picture are geodesic segments, and you could then deduce from the compatibility of the Levi-Cività connection with the metric that the vector field drawn along these edges really is parallel.) It follows that no parallel vector field exists on any neighborhood of that triangle.

Remark 25.2. We intentionally phrased all three of the questions in 25.1 so that they are *local* in nature, i.e. they depend on the connection only in an arbitrarily small neighborhood of p. This is the one problem with Figure 8, since the triangle in that picture cannot be called a "small" neighborhood of anything. The reason to focus only on neighborhoods of a point is that for arbitrary paths $\gamma : [0,1] \to M$ with $\gamma(0) = \gamma(1)$ in a manifold M, it might happen for topological reasons that P^1_γ is not the identity map even if local parallel sections always exist (see e.g. Exercise 25.6 below). One can show however (see Exercise 25.7) that if local parallel sections always exist, then P^1_γ depends only on the homotopy class of γ. From this fact we can still conclude via Figure 8 that local parallel vector fields cannot always exist on S^2, because S^2 is simply connected, so the loop in the picture is homotopic to a constant loop (which would of course give the identity as a parallel transport map).

We now give some formal definitions. We will continue to use the word **parallel** to describe any section $s : \mathcal{U} \to E$ on an open subset $\mathcal{U} \to E$ such that $\nabla s \equiv 0$. The terms **flat**, **horizontal** and **covariantly constant** are sometimes used as synonyms for "parallel" when applied to sections.

Definition 25.3. A connection ∇ on the bundle $E \to M$ is called **flat** if for every $p \in M$ and $v \in E_p$, there exists a neighborhood $\mathcal{U} \subset M$ of p and a flat section $s \in \Gamma(E|_{\mathcal{U}})$ with $s(p) = v$.

Proposition 25.4. *A connection ∇ on $E \to M$ is flat if and only if every point $p \in M$ has a neighborhood $\mathcal{U} \subset M$ with a local trivialization $\Phi : E|_{\mathcal{U}} \to \mathcal{U} \times \mathbb{F}^m$ in which ∇ looks like the trivial connection (see Example 20.1).*

Proof. In one direction this is obvious, since the trivial connection clearly admits flat sections (they look constant in the trivialization). Conversely, if ∇ is flat, then for any $p \in M$, we can choose a basis $v_1, \ldots, v_m$ of E_p and flat sections $e_1, \ldots, e_m \in \Gamma(E|_{\mathcal{U}})$ on some neighborhood $\mathcal{U} \subset M$ of p such that $e_i(p) = v_i$ for $i = 1, \ldots, m$; after possibly shrinking the neighborhood $\mathcal{U}$, we can assume that these also span the fiber E_q for every $q \in \mathcal{U}$, thus they form a frame for E over $\mathcal{U}$. Writing an arbitrary section $s \in \Gamma(E)$ on $\mathcal{U}$ in terms of its components as $s = s^i e_i$ with respect to the frame $e_1, \ldots, e_m$, the Leibniz rule then gives

$$\nabla_X s = ds^i(X) e_i(q) + s^i(q) \nabla_X e_i = ds^i(X) e_i(q) \qquad \text{for every } q \in \mathcal{U},\ X \in T_q M,$$

showing that the covariant derivative is represented in this frame by the differentials of the components. This means that $e_1, \ldots, e_m$ corresponds to a local trivialization in which ∇ is the trivial connection. □

It follows from Proposition 25.4 that for any flat connection, the answers to questions 25.1(2) and (3) are both affirmative.

Exercise 25.5. Prove that if $\dim M = 1$, then every connection on $E \to M$ is flat.

EXERCISE 25.6. Recall the nontrivial real line bundle $\ell \to S^1$ in Example 16.23. Exercise 25.5 implies that any connection ∇ on $\ell \to S^1$ is flat since $\dim S^1 = 1$. Show however that for a path $\gamma : [0,1] \to S^1$ that winds once around the circle and ends at its starting point $\gamma(1) = \gamma(0) =: p$, $P^1_\gamma : \ell_p \to \ell_p$ can never be the identity map.
Hint: This has to do with the fact that $\ell \to S^1$ is a non-orientable bundle.
Remark: The nontriviality of P^1_γ in this example is detecting a topological property of $\ell \to S^1$ that has nothing to do with the connection. This is why we confine the loop in Question 25.1(2) to an arbitrarily small neighborhood of a point instead of allowing arbitrary loops.

EXERCISE 25.7. Suppose ∇ is a flat connection on $E \to M$.

(a) Show that for any smooth map $f : N \to M$, the pullback of ∇ to a connection on $f^*E \to N$ is also flat.
(b) Show that if $\{\gamma_s : [0,1] \to M\}_{s\in[0,1]}$ is a smooth family of paths with fixed end points $\gamma_s(0) = p$ and $\gamma_s(1) = q$ for all $s \in [0,1]$, then the maps $P^1_{\gamma_0}, P^1_{\gamma_1} : E_p \to E_q$ are identical.
Hint: Write $h(s,t) := \gamma_s(t)$ and use the fact that the pullback connection on $h^*E \to [0,1] \times [0,1]$ is also flat. Can you construct a global flat section of h^*E, and if so, how does it behave on the subsets $[0,1] \times \{0\}$ and $[0,1] \times \{1\}$?[69]

EXERCISE 25.8. Prove:

(a) If ∇ a connection on $E \to M$ and $(\mathcal{U}_\alpha, \Phi_\alpha)$ and $(\mathcal{U}_\beta, \Phi_\beta)$ are two overlapping local trivializations in which ∇ looks like the trivial connection, then the transition functions relating these two trivializations are locally constant.
Hint: Think in terms of local frames that are built out of flat sections. If $v = v^i e_i$ where $\nabla v \equiv 0$ and $\nabla e_i \equiv 0$ for every i, what can you conclude from the Leibniz rule?
(b) Show that for any finite subgroup $G \subset \mathrm{GL}(m, \mathbb{F})$, every G-structure on $E \to M$ naturally determines a flat connection, and conversely, if M is compact, then every flat connection on $E \to M$ arises in this way from a G-structure for some finite subgroup $G \subset \mathrm{GL}(m, \mathbb{F})$.

Exercise 25.8 makes the existence of a flat connection seem like a rather restrictive condition: one would not expect the structure group of a vector bundle to be reducible in every case to a finite subgroup. Our main goal in this lecture is to formulate precise conditions for identifying whether a connection is flat. Along the way, we will be able to solve a related problem which is of independent interest and has nothing intrinsically to do with bundles: it leads to the theorem of Frobenius on integrable distributions.

25.2. Integrable frames. The word *integrability* has a variety of meanings in different contexts. Generally it refers to questions in which one is given some data of a linear nature, and would like to find some nonlinear data which produce the given linear data as a form of "derivative". The problem of finding antiderivatives of a smooth function f on $\mathbb{R}$ is the simplest example: it can always be solved (at least in principle) by writing down an antiderivative as a definite integral of f, and is thus not very interesting for the present discussion. A more interesting example is the generalization of this question to higher dimensions, which we examined in Lecture 13:

QUESTION 25.9. *Given a k-form ω on an n-manifold M, under what conditions is ω locally the exterior derivative of a $(k-1)$-form?*

[69]For the purposes of Exercise 25.7, you are safe in pretending that $[0,1] \times [0,1]$ is a smooth manifold, rather than something exotic like a "manifold with boundary and corners". If this worries you, assume that the family of paths $\gamma_s : [0,1] \to M$ is defined for $s \in \mathbb{R}$ instead of just $s \in [0,1]$; this does not change the situation in any significant way.

Including the word "locally" in this question removes topological issues from the discussion: we've seen for instance that certain 1-forms λ on S^1 cannot be differentials of functions because $\int_{S^1} \lambda \neq 0$, but that is a symptom of the fact that the topological invariant $H^1_{\mathrm{dR}}(S^1)$ is nontrivial, and does not stop every 1-form $\lambda \in \Omega^1(S^1)$ from being presentable on a neighborhood $\mathcal{U} \subset S^1$ of any given point as df for some function $f : \mathcal{U} \to \mathbb{R}$. The answer to the question comes of course from the Poincaré lemma, which states that the "integrability condition"

$$d\omega = 0$$

is not only necessary but also sufficient for ω to admit local primitives.

Here is another integrability question whose answer will have some important applications.

QUESTION 25.10. *Suppose M is an n-manifold and $X_1, \ldots, X_n$ is a frame for TM over some open subset $\mathcal{U} \subset M$. Under what conditions does there exist for every point $p \in \mathcal{U}$ a chart $(\mathcal{U}', (x^1, \ldots, x^n))$ with $p \in \mathcal{U}' \subset \mathcal{U}$ such that $X_i = \frac{\partial}{\partial x^i}$ on $\mathcal{U}'$ for every $i = 1, \ldots, n$?*

In other words, every chart naturally gives rise to a local frame for TM, but we want to know when this process can be reversed: when can a frame for TM be "upgraded" locally to a chart?

Let's start with some good news: the answer in the case $n = 1$ is *always*. Indeed, the assumption in this case is that M is a 1-manifold and X_1 is a nowhere-zero vector field on some open subset $\mathcal{U} \subset M$, so a suitable chart $(\mathcal{U}', x)$ on some neighborhood $\mathcal{U}' \subset \mathcal{U}$ of any given point $p \in \mathcal{U}$ can be defined in terms of any local solution $\gamma : (-\epsilon, \epsilon) \to M$ to the initial value problem

$$\dot{\gamma}(t) = X(\gamma(t)), \qquad \gamma(0) = p,$$

namely $\mathcal{U}' := \gamma((-\epsilon, \epsilon)) \subset M$ and $x := \gamma^{-1} : \mathcal{U}' \to \mathbb{R}$. This example provides further justification for the term "integrability": solving an ordinary differential equation is sometimes referred to as *integrating* the equation, and since every ODE admits unique local solutions, every nowhere-zero vector field X_1 on a 1-manifold is integrable in this sense. More generally, it is reasonable to call a local frame $X_1, \ldots, X_n$ for TM *integrable* if it arises from a chart as described in Question 25.10.

It is easy to see on the other hand that for $n \geqslant 2$, not every local frame for TM is integrable, and the Lie bracket gives an obvious obstruction. Indeed, the coordinate vector fields induced by a single chart always commute with each other (see Example 6.6), so $X_1, \ldots, X_n$ clearly cannot be coordinate vector fields, even in arbitrarily small neighborhoods of any given point p, unless $[X_i, X_j] \equiv 0$ for every $i, j = 1, \ldots, n$. One can easily find local frames that do not satisfy this condition, e.g. on $\mathbb{R}^2$ with coordinates (x, y), $(\partial_x, f\partial_x + g\partial_y)$ defines a frame for $T\mathbb{R}^2$ whenever $f, g : \mathbb{R}^2 \to \mathbb{R}$ are smooth functions with g never vanishing, but using Exercise 6.2, one finds

$$[\partial_x, f\partial_x + g\partial_y] = (\partial_x f)\partial_x + (\partial_x g)\partial_y,$$

which does not vanish unless $f(x, y)$ and $g(x, y)$ are both independent of x.

The really good news is that the condition on vanishing Lie brackets is not just necessary, but also sufficient:

THEOREM 25.11. *Suppose $X_1, \ldots, X_n \in \mathfrak{X}(M)$ are vector fields that all commute with each other. Then for any $p \in M$ at which $X_1(p), \ldots, X_n(p)$ form a basis of T_pM, there exists a chart $(\mathcal{U}, x)$ on M with $p \in \mathcal{U}$ such that $X_i = \frac{\partial}{\partial x^i}$ on $\mathcal{U}$ for every $i = 1, \ldots, n$.*

PROOF. For sufficiently small $\epsilon > 0$, we can use the flows of the vector fields $X_1, \ldots, X_n$ to define a smooth map

$$\psi : (-\epsilon, \epsilon)^n \to M : (t^1, \ldots, t^n) \mapsto \varphi^{t^1}_{X_1} \circ \ldots \circ \varphi^{t^n}_{X_n}(p).$$

By Theorem 6.9, the condition $[X_i, X_j] \equiv 0$ implies that the flows $\varphi^s_{X_i}$ and $\varphi^t_{X_j}$ commute with each other, thus for each $j \in \{1, \ldots, n\}$, one can reorder the composition of flows in the above definition so that $\varphi^{t^j}_{X_j}$ comes first, in which case the definition of the flow gives

$$\partial_j \psi(t^1, \dots, t^n) = X_j(\psi(t^1, \dots, t^n)).$$

Since $\psi(0, \dots, 0) = p$, and the vectors $\partial_1 \psi(0, \dots, 0), \dots, \partial_n \psi(0, \dots, 0)$ form a basis of $T_p M$, Lemma 4.2 implies that after possibly shrinking $\epsilon > 0$, ψ is the inverse of a chart on some neighborhood of p. That chart is the one we were looking for. □

25.3. Integrability of distributions. We now return to the question of how to identify when a connection ∇ on the bundle $E \to M$ is flat. From a geometric perspective, a section $s : \mathcal{U} \to E$ over some open subset $\mathcal{U} \subset M$ can be characterized purely in terms of its image

$$\Sigma := s(\mathcal{U}) \subset E,$$

which is a submanifold of the total space E having exactly one intersection point with each of the fibers $E_p \subset E$ for $p \in \mathcal{U}$. The condition $\nabla s \equiv 0$ then holds if and only if this submanifold is always tangent to the horizontal subbundle $HE \subset TE$ determined by the connection, that is,

$$T_v \Sigma = H_v E \qquad \text{for all } v \in \Sigma.$$

With this picture in mind, we can now reframe the flatness question in a somewhat wider context.

Definition 25.12. A smooth k-dimensional **distribution** on a manifold M is a smooth subbundle $\xi \subset TM$ of rank k. It is also sometimes called a **k-plane field**. Given such a distribution, an **integral submanifold** for ξ is a smooth k-dimensional submanifold $\Sigma \subset M$ such that

$$T_p \Sigma = \xi_p \qquad \text{for all } p \in \Sigma.$$

The distribution ξ is called **integrable** if for every point $p \in M$, ξ has an integral submanifold containing p.

Since we will not consider non-smooth distributions in this course, we will usually omit the word "smooth" and just refer to them as "distributions". Note that integral submanifolds do not need to be large in any sense, i.e. noncompact submanifolds diffeomorphic to a k-ball are fine, so the integral submanifold through $p \in M$ may be contained in an arbitrarily small neighborhood of p, and in this sense integrability of a distribution is a purely local condition.

Thinking in terms of distributions and integral submanifolds makes possible a slight reformulation of our goal:

Proposition 25.13. *A connection on a vector bundle $E \to M$ is flat if and only if its horizontal subbundle $HE \subset TE$ is an integrable distribution on the total space E.* □

Example 25.14. For any vector bundle $\pi : E \to M$, the vertical subbundle $VE \subset TE$ is also a distribution on the total space E, and it is *always* integrable. Indeed, the integral submanifolds of VE are the fibers of $\pi : E \to M$, and there is indeed one through every point.

The integrability problem for distributions bears several similarities to the frames considered in the previous section. One is that the 1-dimensional case is trivial: every 1-dimensional distribution (in a manifold of arbitrary dimension) is integrable. To see this on a neighborhood of any given point $p \in M$, one need only choose a vector field $X \in \mathfrak{X}(M)$ that is nonzero at p and takes values in ξ near p, as the flow lines of that vector field then trace out integral submanifolds of ξ, one of which passes through p. For a k-dimensional distribution $\xi \subset TM$ with $k \geqslant 2$, however, it is harder to see why integral submanifolds should exist, and in general they don't. Figure 9 for instance shows a 2-dimensional distribution on $\mathbb{R}^3$ consisting of 2-planes that "twist" in a way that would seem to prevent any surface from being tangent to them at every point. As with the frames in §25.2, there is in fact a necessary condition that can be stated easily, and it involves the Lie bracket:

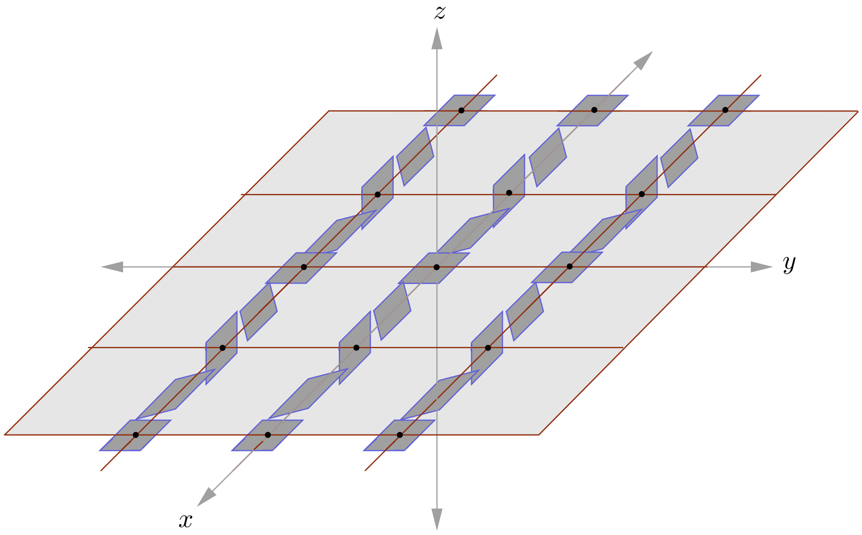

FIGURE 9. A non-integrable 2-dimensional distribution on $\mathbb{R}^3$.

LEMMA 25.15. *If $\xi \subset TM$ is an integrable distribution, then for every pair of vector fields $X, Y \in \mathfrak{X}(M)$ that take their values in ξ, the bracket $[X, Y] \in \mathfrak{X}(M)$ also takes its values in ξ.*

PROOF. Choose any point $p \in M$ and suppose $\Sigma \subset M$ is an integral submanifold containing p. Since $T_q\Sigma = \xi_q$ for all $q \in \Sigma$, vector fields $X, Y \in \mathfrak{X}(M)$ with values in ξ then define vector fields on Σ by restriction, and $[X|_\Sigma, Y|_\Sigma]$ is then (obviously) also a vector field on Σ, which necessarily matches the restriction of $[X, Y] \in \mathfrak{X}(M)$ to Σ. (You can check this by applying the operators $\mathcal{L}_X$ and $\mathcal{L}_Y$ to arbitrary smooth functions on M and their restrictions to Σ.) It follows that $[X, Y](p) \in T_p\Sigma = \xi_p$, and since p was chosen arbitrarily, $[X, Y](q)$ is therefore in ξ_q for every $q \in M$. □

The best integrability theorems are those for which the obviously necessary condition is also sufficient, and that turns out to be the case here as well. The result is known as the *Frobenius integrability theorem.*

THEOREM 25.16 (Frobenius). *A distribution $\xi \subset TM$ on M is integrable if and only if for every pair of vector fields $X, Y \in \mathfrak{X}(M)$ taking values in ξ, $[X, Y] \in \mathfrak{X}(M)$ also takes values in ξ.*

The easy direction of this theorem is Lemma 25.15 above. To prove the converse, it will be more convenient at first to consider the special case where our manifold is the total space of a vector bundle $\pi : E \to M$ and the distribution is a horizontal subbundle $HE \subset TE$, meaning any subbundle of TE that is complementary to the vertical subbundle,

$$TE = VE \oplus HE.$$

We will not assume any more than this, so in particular, HE does not need to satisfy the other condition in our first definition of a connection (Definition 19.4), which was meant to guarantee that the resulting parallel transport maps are linear. As in Lemma 19.1, HE determines horizontal lift isomorphisms

$$\mathrm{Hor}_v : T_{\pi(v)}M \xrightarrow{\cong} H_vE \subset T_vE \qquad \text{for each } v \in E,$$

and every vector field $X \in \mathfrak{X}(M)$ therefore has a horizontal lift $X^h \in \Gamma(HE) \subset \mathfrak{X}(E)$, defined by

$$X^h(v) := \mathrm{Hor}_v(X(p)) \qquad \text{for } p \in M,\ v \in E_p.$$

We will denote by

$$H : TE \to HE$$

the bundle map that projects each T_vE linearly to H_vE along V_vE.

Exercise 25.17. Show that for any $X \in \mathfrak{X}(M)$ and $f \in C^\infty(M)$, $\mathcal{L}_{X^h}(f \circ \pi) = (\mathcal{L}_X f) \circ \pi$.

Lemma 25.18. *If $\eta, \xi \in \Gamma(HE) \subset \mathfrak{X}(E)$ satisfy $\mathcal{L}_\eta(f \circ \pi) \equiv \mathcal{L}_\xi(f \circ \pi)$ for every function $f \in C^\infty(M)$, then $\eta \equiv \xi$.*

Proof. If $\eta(v) \neq \xi(v)$ for some $v \in E_p$ at $p \in M$, then $\pi_*\eta(v) \neq \pi_*\xi(v)$ since $\pi_* : TE \to TM$ maps H_vE isomorphically to T_pM; we shall assume without loss of generality that $\pi_*\xi(v) \neq 0$. Then there exists a smooth function $f : M \to \mathbb{R}$ satisfying $df(\pi_*\xi(v)) \neq 0$ and $df(\pi_*\eta(v)) = 0$, which means $\mathcal{L}_\eta(f \circ \pi)(v) = 0 \neq \mathcal{L}_\xi(f \circ \pi)(v)$. □

Lemma 25.19. *For any $X, Y \in \mathfrak{X}(M)$, $[X,Y]^h = H \circ [X^h, Y^h]$.*

Proof. Observe first that for any $\xi \in \mathfrak{X}(E)$ and $f \in C^\infty(M)$,

$$\mathcal{L}_\xi(f \circ \pi) = \mathcal{L}_{H \circ \xi}(f \circ \pi),$$

i.e. ξ can be replaced with its horizontal part or vice versa since the difference between them is vertical, and $d(f \circ \pi)|_{VE} \equiv 0$. Then for $X, Y \in \mathfrak{X}(M)$, using Exercise 25.17,

$$\begin{aligned}
\mathcal{L}_{H\circ[X^h,Y^h]}(f \circ \pi) &= \mathcal{L}_{[X^h,Y^h]}(f \circ \pi) = \mathcal{L}_{X^h}\mathcal{L}_{Y^h}(f \circ \pi) - \mathcal{L}_{Y^h}\mathcal{L}_{X^h}(f \circ \pi) \\
&= \mathcal{L}_{X^h}\left((\mathcal{L}_Y f) \circ \pi\right) - \mathcal{L}_{Y^h}\left((\mathcal{L}_X f) \circ \pi\right) = (\mathcal{L}_X\mathcal{L}_Y f) \circ \pi - (\mathcal{L}_Y\mathcal{L}_X f) \circ \pi \\
&= (\mathcal{L}_{[X,Y]} f) \circ \pi.
\end{aligned}$$

Likewise, again applying Exercise 25.17,

$$\mathcal{L}_{[X,Y]^h}(f \circ \pi) = (\mathcal{L}_{[X,Y]} f) \circ \pi = \mathcal{L}_{H\circ[X^h,Y^h]}(f \circ \pi),$$

so the result follows from Lemma 25.18 □

We now come to the main step in the proof of the Frobenius theorem.

Lemma 25.20. *Suppose that for every pair of vector fields $X, Y \in \mathfrak{X}(M)$, the vector field $[X^h, Y^h] \in \mathfrak{X}(E)$ takes values in HE. Then $HE \subset TE$ is an integrable distribution on E.*

Proof. Since the question is purely local, we lose no generality if we replace M with a small neighborhood of some point $p \in M$ on which a chart $(x^1, \ldots, x^n)$ can be defined. Denote the resulting coordinate vector fields by $X_j := \partial_j \in \mathfrak{X}(M)$ for $j = 1, \ldots, n$. By assumption $[X_i^h, X_j^h]$ is horizontal for every i and j, thus by Lemma 25.19,

$$[X_i^h, X_j^h] = H \circ [X_i^h, X_j^h] = [X_i, X_j]^h = 0.$$

It follows that for any $v \in E_p$, we can construct an integral submanifold through v via the commuting flows of X_i^h: it is parametrized by the map

$$\psi(t^1, \ldots, t^n) = \varphi^{t^1}_{X_1^h} \circ \ldots \circ \varphi^{t^n}_{X_n^h}(v) \tag{25.1}$$

for real numbers $t^1, \ldots, t^n$ sufficiently close to 0. □

Exercise 25.21. Verify that the map (25.1) parametrizes an embedded integral submanifold of HE.

The last step is to observe that while the distribution we've been considering in this discussion looks like a special case, there is no actual loss of generality.

Proof of Theorem 25.16. We assume $\xi \subset TM$ is a k-dimensional distribution such that $[X, Y] \in \Gamma(\xi)$ whenever $X, Y \in \Gamma(\xi)$. Since the integrability question is purely local, we can choose a chart near some point $p \in M$ so as to assume without loss of generality that M is an open subset $\mathcal{U} \subset \mathbb{R}^n$, and after a linear change of coordinates, we can also arrange that $\xi_p \subset \mathbb{R}^n$ is complementary to the subspace $\{0\} \times \mathbb{R}^{n-k} \subset \mathbb{R}^n$. After possibly shrinking the neighborhood $\mathcal{U}$, it follows that ξ_q is also complementary to this same subspace for every $q \in \mathcal{U}$. We can now view $\mathcal{U}$ as an open subset in the total space of the trivial vector bundle $\mathbb{R}^n = \mathbb{R}^k \times \mathbb{R}^{n-k} \to \mathbb{R}^k : (q, v) \mapsto q$, in which fibers take the form $\{q\} \times \mathbb{R}^{n-k}$, and ξ is therefore a horizontal subbundle. The stated condition on Lie brackets then establishes the hypothesis of Lemma 25.20, implying ξ is integrable. $\square$

Exercise 25.22. While integral submanifolds of a distribution $\xi \subset M$ through a given point $p \in M$ are not guaranteed to exist, show that they are unique in the following sense: if $\Sigma_1, \Sigma_2 \subset M$ are two integral submanifolds containing p, then there exist neighborhoods $\mathcal{U}_1 \subset \Sigma_1$ and $\mathcal{U}_2 \subset \Sigma_2$ of p in each such that $\mathcal{U}_1 = \mathcal{U}_2$.
Hint: It may help to think only about the special case $\xi = HE \subset TE$ for a vector bundle $\pi : E \to M$, since every case locally looks like this one. Remember that a horizontal subbundle always uniquely determines parallel transport along paths.

25.4. A tensorial characterization of flatness. In preparation for the general definition of curvature in the next lecture, we can now associate to every connection ∇ on a bundle $\pi : E \to M$ a tensor field whose vanishing is equivalent to the integrability condition in the Frobenius theorem. We continue to denote by $H : TE \to HE$ the projection along VE, and define also the complementary projection

$$V : TE \to VE,$$

which projects $T_v E$ linearly along $H_v E$ to $V_v E$ for each $v \in E$. We use these to define a bilinear map $\widehat{\Omega}_K : \mathfrak{X}(E) \times \mathfrak{X}(E) \to \Gamma(VE)$ by

$$\widehat{\Omega}_K(\eta, \xi) := -V\left([H(\eta), H(\xi)]\right).$$

The Frobenius theorem is equivalent to the statement that this map vanishes if and only if $HE \subset TE$ is an integrable distribution: indeed, every vector field on E with values in HE can be written as $H(\eta)$ for some $\eta \in \mathfrak{X}(E)$, and an arbitrary $\eta \in \mathfrak{X}(E)$ takes values in HE if and only if $V(\eta) \equiv 0$. The real reason why $\widehat{\Omega}_K$ is useful is that in addition to characterizing the flatness of a connection, it defines a tensor:

Lemma 25.23. *The bilinear map $(\eta, \xi) \mapsto \widehat{\Omega}_K(\eta, \xi)$ is C^∞-linear in both η and ξ.*

Proof. Since $\widehat{\Omega}_K$ is clearly antisymmetric, it suffices to show that it is C^∞-linear with respect to η. We use the formula $[fX, Y] = f[X, Y] - (\mathcal{L}_Y f)X$ from Exercise 6.4: for any $\eta, \xi \in \mathfrak{X}(E)$ and $f \in C^\infty(E)$,

$$\begin{aligned}\widehat{\Omega}_K(f\eta, \xi) &= -V\left([fH(\eta), H(\xi)]\right) = -V\left(f[H(\eta), H(\xi)] - \mathcal{L}_{H(\xi)} f \cdot H(\eta)\right) \\ &= -fV\left([H(\eta), H(\xi)]\right) = f\widehat{\Omega}_K(\eta, \xi),\end{aligned}$$

where the term $V\left(\mathcal{L}_{H(\xi)} f \cdot H(\eta)\right) = \mathcal{L}_{H(\xi)} f \cdot V(H(\eta))$ disappears because $H(\eta)$ takes horizontal values and is therefore annihilated by V. $\square$

The lemma implies that $\widehat{\Omega}_K$ can be interpreted as defining a bilinear bundle map

$$\widehat{\Omega}_K : TE \oplus TE \to VE,$$

and since it is antisymmetric, we also think of it as a *bundle-valued* differential 2-form on E, and write

$$\widehat{\Omega}_K \in \Omega^2(E, VE).$$

This is one version of an object called the *curvature* 2*-form* determined by the connection ∇ on E; you can now regard the subscript K as either a reference to the projection $K : TE \to E$ that determines the connection (Definition 19.5), or simply as an abbreviation for the word *Krümmung*. In the next lecture we will discuss a somewhat more user-friendly variant of $\widehat{\Omega}_K$ that packages the same information. Let us first record the following consequence of the Frobenius theorem:

COROLLARY 25.24. *A connection on a vector bundle $\pi : E \to M$ is flat if and only if the bundle-valued 2-form $\widehat{\Omega}_K \in \Omega^2(E, VE)$ vanishes.* □

25.5. Addendum: integrability in general. Integrability theorems are ubiquitous in differential geometry, and one should learn to recognize them. They can take different forms depending on the context in which they arise, but most fit the following paradigm: we have a manifold M whose tangent bundle TM carries some extra geometric structure defining a preferred class of local frames, which are guaranteed to exist on neighborhoods of any point. A preferred class of frames determines a preferred class of charts, namely $(x^1, \dots, x^n)$ such that the frame formed by the coordinate vector fields $\partial_1, \dots, \partial_n$ belongs to the preferred class. But as we saw in §25.2, not every frame comes from a chart, so it is typically harder to find a preferred chart than a preferred frame, and they don't always exist: typically some nontrivial *integrability condition* is required before the local existence of preferred charts can be guaranteed.

Theorem 25.11 fits this paradigm in a trivial way: in this case the extra geometric structure is the frame itself, and the question is whether that particular frame can arise locally from a chart.

The Frobenius theorem can also be recast in this language. Here the extra geometric structure is a distribution $\xi \subset TM$, i.e. a subbundle of the tangent bundle, and the preferred class of frames comes from Proposition 17.12: every point $p \in M$ has a neighborhood $\mathcal{U} \subset M$ on which there is a frame $X_1, \dots, X_n$ for TM such that ξ is the span of $X_1, \dots, X_k$ at every point $\mathcal{U}$. A frame arising naturally from a chart $x = (x^1, \dots, x^n) : \mathcal{U} \to \mathbb{R}^n$ will have this property if and only if ξ is spanned at every point by the first k coordinate vector fields $\partial_1, \dots, \partial_k$, in which case integral submanifolds obviously exist through every point: they take the form $x^{-1}(\mathbb{R}^k \times \{q\})$ for constants $q \in \mathbb{R}^{n-k}$. The existence of charts of this form is in fact *equivalent* to integrability:

PROPOSITION 25.25. *A k-dimensional distribution $\xi \subset TM$ is integrable if and only if every point $p \in M$ admits a neighborhood $\mathcal{U} \subset M$ with a chart $x : \mathcal{U} \to \mathbb{R}^n$ such that the sets $x^{-1}(\mathbb{R}^k \times \{q\})$ are integral submanifolds of ξ for each $q \in \mathbb{R}^{n-k}$.*

EXERCISE 25.26. Prove Proposition 25.25.
Hint: This may seem easier if you think of ξ as a horizontal subbundle in TE for some vector bundle E.

Proposition 25.25 shows that for an integral distribution, the integral submanifolds are not just locally unique (cf. Exercise 25.22), but they also fit together into a locally-defined smooth family of smooth submanifolds. This gives rise to a decomposition of M called a **foliation** (*Blätterung*), and every connected subset $\Sigma \subset M$ formed as a union of overlapping connected integral submanifolds is called a **leaf** (*Blatt*) of the foliation. By construction, every point in M belongs to a unique leaf of the foliation, and unless $\xi = TM$, there are always uncountably many distinct leaves. It is pleasant to picture them as a smooth family of disjoint submanifolds whose union is M, though this description is not always completely accurate: the following example shows that leaves are not always submanifolds, at least not globally.

EXERCISE 25.27. Given a constant $(a, b) \in \mathbb{R}^2 \backslash \{0\}$, consider a distribution ξ on $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$ such that ξ at every point is the subspace spanned by the constant vector field $X = a\partial_x + b\partial_y$, where ∂_x, ∂_y are the usual coordinate vector fields of $\mathbb{R}^2$ (which are also well-defined on $\mathbb{T}^2$ since its tangent spaces are all canonically isomorphic to $\mathbb{R}^2$). This distribution is always integrable since it is 1-dimensional. Draw pictures of some representative leaves of the resulting foliation in the cases where (a, b) is $(1, 0)$, $(0, 1)$, $(1, 1)$ and $(2, 1)$. In these cases all leaves are 1-dimensional submanifolds of $\mathbb{T}^2$. Show however that if one of a or b is rational and the other is irrational, then every leaf of the foliation is dense in $\mathbb{T}^2$, and therefore cannot be a submanifold.

For your information, here are some additional examples of integrability results, most of which we will not cover in this course, though the first is an important exception.

EXAMPLE 25.28. If (M, g) is a pseudo-Riemannian manifold with signature (k, ℓ), the orthonormal frames define a preferred class of local frames for TM, equivalent to $\mathrm{O}(k, \ell)$-compatible local trivializations. Such a frame arises from a chart $(x^1, \ldots, x^n)$ if and only if the metric g has constant components $g_{ij} \equiv \eta_{ij}$ in this chart (cf. the discussion following Proposition 23.2). We will show in the next lecture that charts of this form exist if and only if the Levi-Cività connection on (M, g) is flat, i.e. its curvature vanishes. You can already see that this is a necessary condition, since having constant components g_{ij} in some chart implies that the connection is trivial in the corresponding local trivialization.

EXAMPLE 25.29. We did not include symplectic structures among the list of "G-structures" in Lecture 18, but we could have done. The *standard symplectic structure* of $\mathbb{R}^{2m}$ is the 2-form $\omega_{\mathrm{std}} := \sum_{j=1}^m dp^j \wedge dq^j$ written in global coordinates $(p^1, q^1, \ldots, p^m, q^m)$. A linear transformation $A : \mathbb{R}^{2m} \to \mathbb{R}^{2m}$ is called **symplectic** if it preserves this structure, meaning $\omega_{\mathrm{std}}(AX, AY) = \omega_{\mathrm{std}}(X, Y)$ for all $X, Y \in \mathbb{R}^{2m}$, and the set of all such transformations forms the **symplectic linear group** $\mathrm{Sp}(2m) \subset \mathrm{GL}(2m, \mathbb{R})$. The 2-form ω_{std} is nondegenerate, meaning $\omega_{\mathrm{std}}(X, \cdot) \neq 0 \in (\mathbb{R}^{2m})^*$ for every $X \neq 0 \in \mathbb{R}^{2m}$. Conversely, it is a simple exercise in symplectic linear algebra to show that for any real $2m$-dimensional vector space V with a nondegenerate alternating 2-form $\omega \in \Lambda^2 V^*$, there exists a basis $(P_1, Q_1, \ldots, P_m, Q_m)$ such that

$$\begin{aligned} &\omega(P_j, Q_j) = 1 && \text{for all } j, \\ &\omega(P_i, Q_j) = 0 && \text{for all } i \neq j, \\ &\omega(P_i, P_j) = \omega(Q_i, Q_j) = 0 && \text{for all } i, j, \end{aligned} \tag{25.2}$$

so this basis produces an isomorphism $V \cong \mathbb{R}^{2m}$ that identifies ω with ω_{std}. A procedure for finding the basis is as follows: first choose any linearly-independent P_1, Q_1 such that $\omega(P_1, Q_1) = 1$, which is possible because ω is nondegenerate and alternating. The restriction of ω to the subspace $V_1 \subset V$ spanned by P_1 and Q_1 is then nondegenerate, so by a straightforward analogue of Lemma 24.7, its *symplectic orthogonal complement*

$$V_1^{\omega\perp} := \left\{ v \in V \;\middle|\; \omega(v, \cdot)|_{V_1} = 0 \right\}$$

satisfies $\mathbb{R}^{2m} = V_1 \oplus V_1^{\omega\perp}$, and $\omega|_{V_1^{\omega\perp}}$ is also nondegenerate. Now repeat the same argument on $V_1^{\omega\perp}$, which is 2 dimensions smaller than V, and keep repeating until there are no dimensions left. In summary, every nondegenerate alternating 2-form is equivalent to the standard symplectic form via a choice of basis.

On a real vector bundle $E \to M$ of even rank $2m$, an $\mathrm{Sp}(2m)$-structure now determines on each fiber E_p an alternating 2-form $\omega_p \in \Lambda^2 E_p^*$ that looks like ω_{std} in any $\mathrm{Sp}(2m)$-compatible local trivialization, and the map $p \mapsto \omega_p$ is then a smooth section ω of the vector bundle $\Lambda^2 E^* \to M$. The frames corresponding to $\mathrm{Sp}(2m)$-compatible trivializations consist of tuples of sections $(P_1, Q_1, \dots, P_m, Q_m)$ that satisfy the relations in (25.2); we call them **symplectic frames**. Conversely, for any choice of section $\omega \in \Gamma(\Lambda^2 E^*)$ that is nondegenerate on every fiber, one can use the procedure outlined above to construct frames that satisfy (25.2) over sufficiently small neighborhoods of any point in M. A covering of M by neighborhoods with such frames gives rise to an $\mathrm{Sp}(2m)$-structure on $E \to M$, and we then call $E \to M$ a **symplectic vector bundle**.

If (M, ω) is a $2n$-dimensional symplectic manifold, then ω makes $TM \to M$ into a symplectic vector bundle, for which any local coordinates $(p^1, q^1, \dots, p^n, q^n)$ in which $\omega = \sum_{j=1}^n dp^j \wedge dq^j$ give rise to a symplectic frame $\frac{\partial}{\partial p^1}, \frac{\partial}{\partial q^1}, \dots, \frac{\partial}{\partial p^n}, \frac{\partial}{\partial q^n}$. But not every $\mathrm{Sp}(2n)$-structure on the bundle $TM \to M$ arises in this way from a symplectic form on M. According to the previous paragraph, an $\mathrm{Sp}(2n)$-structure on $TM \to M$ is equivalent to a choice of smooth 2-form $\omega \in \Omega^2(M)$ that is nondegenerate on every fiber. In this situation, local symplectic frames can always be found, but can they always also be realized as coordinate vector fields for a chart $(p^1, q^1, \dots, p^n, q^n)$ in which $\omega = \sum_{j=1} dp^j \wedge dq^j$? There is an obvious necessary condition for this: ω cannot take that form in any local coordinates if it is not closed, and indeed, if $\dim M > 2$, there is no reason in general why a globally nondegenerate 2-form must also be closed. We can therefore view "$d\omega = 0$" as an integrability condition for a symplectic tangent bundle to be upgraded to a symplectic manifold. According to Darboux's theorem, this condition is also sufficient, i.e. every closed nondegenerate 2-form matches the standard symplectic form in some local coordinates. For more on both symplectic vector bundles and Darboux's theorem, see [MS17].

EXAMPLE 25.30. A volume form $\mu \in \Omega^n(M)$ on an n-manifold M is the same thing as an $\mathrm{SL}(n, \mathbb{R})$-structure on the bundle $TM \to M$, and the preferred class of frames consists of tuples of vector fields $X_1, \dots, X_n$ defined on open subsets $\mathcal{U} \subset M$ such that $\mu(X_1, \dots, X_n) \equiv 1$. The preferred class of charts $(x^1, \dots, x^n)$ can then be characterized by the condition that μ in any such chart looks like the *standard* volume form $dx^1 \wedge \dots \wedge dx^n$. It is very easy to turn any local frame into one that satisfies $\mu(X_1, \dots, X_n) \equiv 1$, but less obvious in general whether every volume form can be made to look standard near every point by choosing the right coordinates. However, it is true: the necessary and sufficient integrability condition for this is $d\mu = 0$, just as with symplectic forms, but with the important difference that it is *always* satisfied since μ is a top-dimensional form. One can prove this integrability result by a slight variation on one of the standard proofs of Darboux's theorem, using the "Moser deformation" trick.

EXAMPLE 25.31. A deep integrability theorem for almost complex structures $J \in \Gamma(\mathrm{End}(TM))$ on a $2n$-manifold M was mentioned in Exercise 8.5. An almost complex structure is equivalent to a $\mathrm{GL}(n, \mathbb{C})$-structure on TM, where $\mathrm{GL}(n, \mathbb{C})$ is identified with a subgroup of $\mathrm{GL}(2n, \mathbb{R})$ as in Example 18.10, and local frames in the preferred class take the form $(X_1, Y_1, \dots, X_n, Y_n)$ where $Y_j = JX_j$ and $X_j = -JY_j$ for each $j = 1, \dots, n$. A covering of M by charts that produce frames of this kind is equivalent to a covering by *complex* charts whose transition maps are holomorphic, thus making M into an n-dimensional complex manifold. An almost complex structure J is called *integrable* if M admits a covering by charts with this property, and according to the Newlander-Nirenberg theorem, the necessary and sufficient condition for J to be integrable is the vanishing of its associated Nijenhuis tensor $N \in \Gamma(T_2^1 M)$.

26. Curvature on a vector bundle

Like connections, curvature is one of those concepts that can be given several equivalent but cosmetically quite different definitions, each of which has distinct advantages in different situations.

In this lecture we give two definitions[70] of the curvature of a connection on a vector bundle $\pi : E \to M$, and prove that they are equivalent. It will be immediate from one of these definitions that a connection is flat if and only if its curvature vanishes, while the other definition answers the question of when covariant partial derivatives in different directions do or do not commute.

26.1. Prelude: bundle-valued forms. We have already had a few occasions to mention bundle-valued differential forms, but have not given any formal definition of this notion so far. The time for that is now: for any vector bundle $\pi : E \to M$ and each integer $k \geqslant 0$, we define

$$\Omega^k(M, E)$$

as the vector space of all smooth maps

$$\omega : \underbrace{TM \oplus \ldots \oplus TM}_{k} \to E$$

such that for every $p \in M$, the restriction of ω to the fiber over p is an antisymmetric k-fold multilinear map $\omega_p : T_pM \times \ldots \times T_pM \to E_p$. As with real-valued forms, the antisymmetry condition is vacuous in the cases $k = 0, 1$, and the convention is to define $\Omega^0(M, E) := \Gamma(E)$. Another way to formulate the definition would be that $\Omega^k(M, E)$ is the space of smooth sections of the vector bundle $(\Lambda^k T^*M) \otimes E$, whose fibers can be identified canonically with the spaces of antisymmetric multilinear maps described above. Note that if $E \to M$ is a *complex* vector bundle, then it is regarded as a real bundle for the purposes of these definitions, since the fibers of TM cannot be assumed to be equipped with any complex structure.

26.2. The curvature 2-form. In §25.4 we associated to any connection ∇ on a vector bundle $\pi : E \to M$ a bundle-valued 2-form $\widehat{\Omega}_K \in \Omega^2(E, VE)$ satisfying

$$\widehat{\Omega}_K(\eta, \xi) := -V\left([H(\eta), H(\xi)]\right)$$

for all $\eta, \xi \in \mathfrak{X}(E)$, where $V : TE \to VE$ denotes the fiberwise-linear projection along HE and $H : TE \to HE$ is the complementary projection. We saw that this formula for $\widehat{\Omega}_K$ is C^∞-linear in both variables, and that by the Frobenius theorem, it vanishes if and only if the distribution $HE \to TE$ is integrable, which means the connection ∇ is flat. All of this is true for *any* horizontal subbundle $HE \subset TE$, i.e. we have not yet actually used the additional requirement in Definition 19.4 that HE should be compatible with scalar multiplication,[71] i.e. the relation

$$(m_\lambda)_*(HE) = HE$$

for every $\lambda \in \mathbb{F}$, with $m_\lambda : E \to E$ denoting the smooth map $v \mapsto \lambda v$. This condition makes it possible to replace $\widehat{\Omega}_K \in \Omega^2(E, VE)$ with an object that is simpler, but equivalent. Recall from Definition 19.5 that the connection can also be characterized via a map $K : TE \to E$ that sends T_vE linearly to $E_{\pi(v)}$ and vanishes on the horizontal subspaces: K is actually just the composition of the fiberwise-linear projection $TE \to VE$ with the canonical isomorphisms

$$\mathrm{Vert}_v^{-1} : V_vE \to E_p \qquad \text{for } v \in E_p,\ p \in M.$$

The condition $(m_\lambda)_*(HE) = HE$ is then equivalent to the condition

$$K \circ Tm_\lambda = m_\lambda \circ K \tag{26.1}$$

[70] plus two more that will be implicit in the exercises at the end

[71] Without assuming the condition $(m_\lambda)_*(HE) = HE$ holds, one can still regard the horizontal subbundle $HE \subset TE$ as a connection on $\pi : E \to M$, but only after forgetting the fact that the fibers of this bundle are vector spaces and regarding them instead as smooth submanifolds, so that $\pi : E \to M$ is now an example of a *fiber bundle*.

for all $\lambda \in \mathbb{F}$. Writng $\mathrm{End}(E) := \mathrm{Hom}(E, E)$, we claim that the expression

$$\Omega_K(X,Y)v := \mathrm{Vert}_v^{-1}\left(\widehat{\Omega}_K(\mathrm{Hor}_v(X), \mathrm{Hor}_v(Y))\right) \in E_p \qquad \text{for } X, Y \in T_pM,\ v \in E_p,\ p \in M$$

defines a bundle-valued 2-form

$$\Omega_K \in \Omega^2(M, \mathrm{End}(E)).$$

We can already see that this expression is bilinear and antisymmetric in X and Y; the main thing to check is that for each fixed $X, Y \in T_pM$, the map $E_p \to E_p : v \mapsto \Omega_K(X,Y)v$ is linear. It is clearly smooth, so by Lemma 19.2, it will be sufficient to show that it is also compatible with scalar multiplication. To see this, let us associate to each vector field $X \in \mathfrak{X}(M)$ on M the "horizontal" vector field on E given by $X^h(v) := \mathrm{Hor}_v(X(p))$ for $v \in E_p$ and $p \in M$ as in §25.3. Since K is the composition of V with Vert_v^{-1}, we can rewrite Ω_K in terms of the definition of $\widehat{\Omega}_K$ as

$$\Omega_K(X,Y)v = -K\left([X^h, Y^h](v)\right) \tag{26.2}$$

for any $X, Y \in \mathfrak{X}(M)$. Now observe that since $(m_\lambda)_*(HE) = HE$, the horizontal vector field X^h (and similarly Y^h) satisfies the relation

$$X^h(\lambda v) = Tm_\lambda(X^h(v)).$$

If $\lambda \neq 0$, so that $m_\lambda : E \to E$ is a diffeomorphism, this relation can be stated more succinctly as the condition that X^h is its own pushforward under this diffeomorphism:

$$(m_\lambda)_* X^h = X^h \in \mathfrak{X}(E).$$

By Exercise 6.5, it follows that

$$(m_\lambda)_*[X^h, Y^h] = [(m_\lambda)_* X^h, (m_\lambda)_* Y^h] = [X^h, Y^h] \in \mathfrak{X}(E),$$

thus $[X^h, Y^h] \in \mathfrak{X}(E)$ also satisfies the relation

$$[X^h, Y^h](\lambda v) = Tm_\lambda([X^h, Y^h](v)).$$

Continuing under the assumption $\lambda \neq 0$, we can now use (26.1) to conclude

$$\begin{aligned}\Omega_K(X,Y)\lambda v &= -K\left([X^h, Y^h](\lambda v)\right) = -K\left(Tm_\lambda([X^h, Y^h](v))\right) = -m_\lambda \circ K\left([X^h, Y^h](v)\right)\\ &= \lambda \Omega_K(X,Y)v.\end{aligned}$$

If this holds for all nonzero $\lambda \in \mathbb{F}$, then by continuity it also holds for $\lambda = 0$, and the claim is thus proven: $v \mapsto \Omega_K(X,Y)v$ is linear, so Ω_K is a 2-form with values in the vector bundle $\mathrm{End}(E) \to M$.

EXERCISE 26.1. Show that $\Omega_K \in \Omega^2(M, \mathrm{End}(E))$ vanishes if and only if $\widehat{\Omega}_K \in \Omega^2(E, VE)$ vanishes.

DEFINITION 26.2. We call $\Omega_K \in \Omega^2(M, \mathrm{End}(E))$ the **curvature 2-form** of the connection ∇ on $\pi : E \to M$, and say that ∇ has **vanishing curvature** if $\Omega_K \equiv 0$.

Exercise 26.1 now combines with Corollary 25.24 to prove:

COROLLARY 26.3. *A connection on a vector bundle is flat if and only if its curvature vanishes.* □

26.3. The Riemann tensor. The definition of curvature given in the previous section does not easily lend itself to computations. In order to remedy this, let's go back to the third question in 25.1: do covariant partial derivatives in different coordinate directions commute? We've seen that the answer is yes for a flat connection, and one of the main results of the present lecture will be a converse to this: if they always commute, then the connection must be flat.

Let's first reframe the question in coordinate-invariant language. We could ask for instance whether the differential operators $\nabla_X, \nabla_Y : \Gamma(E) \to \Gamma(E)$ must commute for an arbitrary choice of two vector fields $X, Y \in \mathfrak{X}(M)$, but this is not even true in the simplest special case: for the trivial connection on the trivial real line bundle over M, $\Gamma(E)$ is identified with $C^\infty(M)$ and ∇_X and ∇_Y become the operators $\mathcal{L}_X$ and $\mathcal{L}_Y$ respectively, whose failure to commute is measured by the definition of the Lie bracket, which amounts to the formula

$$\mathcal{L}_X \mathcal{L}_Y - \mathcal{L}_Y \mathcal{L}_X = \mathcal{L}_{[X,Y]} \qquad \text{on } C^\infty(M).$$

One might extrapolate from this case and guess that the relation $\nabla_X \nabla_Y - \nabla_Y \nabla_X = \nabla_{[X,Y]}$ should hold for general connections. This turns out to be false in general, but the failure of this identity is measured by a tensor:

Definition 26.4. Given a connection ∇ on a vector bundle $E \to M$, the **Riemann curvature tensor** is the unique multilinear bundle map

$$R : TM \oplus TM \oplus E \to E : (X, Y, v) \mapsto R(X, Y)v$$

such that for all $X, Y \in \mathfrak{X}(M)$ and $v \in \Gamma(E)$,

$$R(X, Y)v = \nabla_X \nabla_Y v - \nabla_Y \nabla_X v - \nabla_{[X,Y]} v.$$

The exercise below shows that this is well defined, and in particular, if $E = TM$, R is a tensor field of type $(1, 3)$ on M.

Exercise 26.5. Show that $R(X, Y)v$ is C^∞-linear with respect to each of its three arguments.

Exercise 26.6. Choosing a chart $x = (x^1, \ldots, x^n) : \mathcal{U} \to \mathbb{R}^n$ and a frame $(e_1, \ldots, e_m)$ for E over some open subset $\mathcal{U} \subset M$, define the components $R^a{}_{jkb} : \mathcal{U} \to \mathbb{F}$ of the Riemann tensor R such that

$$R(\partial_j, \partial_k)e_b = R^a{}_{jkb} e_a,$$

hence $(R(X, Y)v)^a = R^a{}_{jkb} X^j Y^k v^b$ for any $X, Y \in T_pM$ and $v \in E_p$ at $p \in \mathcal{U}$. Show that these components are given in terms of the Christoffel symbols of the connection by

$$R^a{}_{jkb} = \partial_j \Gamma^a_{kb} - \partial_k \Gamma^a_{jb} + \Gamma^a_{jc}\Gamma^c_{kb} - \Gamma^a_{kc}\Gamma^c_{jb}.$$

Remark 26.7. Exercise 26.6 together with (22.6) shows that for the Levi-Cività connection on a pseudo-Riemannian manifold, the Riemann tensor is determined by the second derivatives of the components of the metric in any local coordinates.

Exercise 26.8. Suppose $\mathcal{V} \subset \mathbb{R}^2$ is an open subset with coordinates labelled (s, t), $f : \mathcal{V} \to M$ is a smooth map and $v \in \Gamma(f^*E)$ is a section of E along f. Prove the formula

$$\nabla_s \nabla_t v - \nabla_t \nabla_s v = R(\partial_s f, \partial_t f)v \qquad \text{on } \mathcal{V}.$$

Hint: On any neighborhood in $\mathcal{V}$ on which f is an embedding, you can derive this from the definition of the Riemann tensor after extending f to a diffeomorphism onto an open set in M and choosing a corresponding extension of v to a section of $E \to M$. If $\dim M \geqslant 2$, deduce the general case from this via continuity (cf. the proof of (21.2)), using the fact that any smooth map $\mathbb{R}^2 \supset \mathcal{V} \xrightarrow{f} M$ can be perturbed to become an embedding on some neighborhood of any given point. If $\dim M \leqslant 1$ then there is nothing to prove, because R vanishes (why?) and the connection ∇ is automatically flat, implying that its pullback to $f^*E \to \mathcal{V}$ is also flat (see Exercise 25.7).

It may be surprising at first sight that $R(X,Y)v$ doesn't depend on any derivatives of v: indeed, it seems to tell us less about v than about the connection itself. The main theorem in this lecture says that the Riemann tensor gives a complete characterization of the curvature of the connection—in particular, its vanishing gives yet another necessary and sufficient condition for the connection to be flat.

THEOREM 26.9. *For any vector bundle $E \to M$ with connection ∇, the Riemann tensor R and curvature 2-form Ω_K are related by*

$$R(X,Y)v = \Omega_K(X,Y)v.$$

COROLLARY 26.10. *The connection ∇ on $E \to M$ is flat if and only if for every chart $(x^1,\ldots,x^n)$, the covariant partial derivative operators ∇_i and ∇_j commute for all $i,j \in \{1,\ldots,n\}$.*

We will prove Theorem 26.9 in the next section.

If $\dim M = n$ and $\operatorname{rank}(E) = m$, then the Riemann tensor is described locally by n^2m^2 component functions $R^a{}_{jkb}$ for $j,k \in \{1,\ldots,n\}$ and $a,b \in \{1,\ldots,m\}$. This sounds like quite a lot of information, but it is often useful to notice that these components are not all independent of each other. One nontrivial relation is obvious already from the definition: since $R(X,Y)v$ is antisymmetric in X and Y, we have

$$R^a{}_{jkb} = -R^a{}_{kjb}.$$

One can say more if ∇ is compatible with a bundle metric, as is true in particular for the Levi-Cività connection on a tangent bundle:

EXERCISE 26.11. Show that if ∇ is compatible with a bundle metric $\langle\ ,\ \rangle$ on E, then the Riemann tensor satisfies

$$\langle R(X,Y)v,w\rangle + \langle v, R(X,Y)w\rangle = 0$$

for all $(X,Y,v,w) \in TM \oplus TM \oplus E \oplus E$.
Hint: Given $X,Y \in \mathfrak{X}(M)$ and $v,w \in \Gamma(E)$, apply the operator $\mathcal{L}_X\mathcal{L}_Y - \mathcal{L}_Y\mathcal{L}_X - \mathcal{L}_{[X,Y]}$ to the function $\langle v,w\rangle$.

Exercise 26.11 says that for each $X,Y \in T_pM$, the linear map $R(X,Y) : E_p \to E_p$ is antisymmetric with respect to the bundle metric on E. Let's see what this means in the case where E is the tangent bundle of an oriented Riemannian 2-manifold (Σ, g). The space of antisymmetric linear maps $T_p\Sigma \to T_p\Sigma$ in this case is 1-dimensional, and it has a canonical basis defined as follows. Let

$$J \in \Gamma(T^1_1\Sigma) = \Gamma(\operatorname{End}(T\Sigma))$$

denote the unique bundle map $T\Sigma \to T\Sigma$ such that for each $p \in \Sigma$, $J_p : T_p\Sigma \to T_p\Sigma$ is a 90-degree counterclockwise rotation; here "counterclockwise" means that (X, J_pX) is a positively-oriented basis for each $X \neq 0 \in T_p\Sigma$. Equivalently, J_p is represented by the matrix $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ in any positively-oriented orthonormal basis of $T_p\Sigma$. Since J_p is nontrivial and antisymmetric, every antisymmetric linear map $T_p\Sigma \to T_p\Sigma$ is a scalar multiple of it. Similarly, $\dim \Lambda^2 T^*_p\Sigma = 1$, thus every alternating 2-form on $T_p\Sigma$ is a scalar multiple of the Riemannian volume form (or "area form") $d\mathrm{vol}$ at that point. These two observations, together with Exercise 26.11, imply that for the Levi-Cività connection on the tangent bundle of (Σ, g), there is a unique real-valued function

$$K : \Sigma \to \mathbb{R}$$

such that the Riemann tensor is given by the formula

$$R(X,Y)Z = -K(p)\, d\mathrm{vol}(X,Y)\, JZ \qquad \text{for } p \in \Sigma,\ X,Y,Z \in T_p\Sigma. \tag{26.3}$$

This shows that despite the Riemann tensor being described on any coordinate neighborhood by a total of $2^4 = 16$ component functions, they are all determined by a single function $K : \Sigma \to \mathbb{R}$. This function is called the **Gaussian curvature** of (Σ, g), and we will have much more to say about it in the next two lectures.

REMARK 26.12. While it was convenient in the discussion above to assume Σ was oriented, the function $K : \Sigma \to \mathbb{R}$ in (26.3) is still well defined without this assumption. The reason is that reversing the chosen orientation of Σ causes sign changes in both $d\mathrm{vol}$ and J, and these two sign changes cancel each other so that (26.3) remains valid without any change in K. If Σ is not orientable, one can then define K in a small neighborhood of any point $p \in \Sigma$ by making an arbitrary choice of orientation on this neighborhood; since the result does not depend on this choice, $K : \Sigma \to \mathbb{R}$ is then well defined globally.

26.4. Covariant exterior derivatives. We will prove Theorem 26.9 by relating the bracket to an exterior derivative using a generalization of the formula

$$d\alpha(X, Y) = \mathcal{L}_X(\alpha(Y)) - \mathcal{L}_Y(\alpha(X)) - \alpha([X, Y])$$

for 1-forms $\alpha \in \Omega^1(M)$. This is possible because the definitions of Ω_K, K and R can all be expressed in terms of bundle-valued forms.

The covariant derivative gives a linear map

$$\nabla : \Gamma(E) = \Omega^0(M, E) \to \Omega^1(M, E) = \Gamma(\mathrm{Hom}(TM, E)),$$

and by analogy with the differential $d : \Omega^0(M) \to \Omega^1(M)$, it's natural to extend this to a **covariant exterior derivative**

$$d_\nabla : \Omega^k(M, E) \to \Omega^{k+1}(M, E),$$

defined as follows. Every $\omega \in \Omega^k(M, E)$ can be expressed in local coordinates $x = (x^1, \dots, x^n) : \mathcal{U} \to \mathbb{R}^n$ as

$$\omega = \sum_{1 \leqslant i_i < \ldots < i_k \leqslant n} \omega_{i_1 \ldots i_k} \, dx^{i_1} \wedge \ldots \wedge dx^{i_k}$$

for some *component sections* $\omega_{i_1 \ldots i_k} \in \Gamma(E|_{\mathcal{U}})$. Then $d_\nabla \omega$ is defined locally as

$$\begin{aligned} d_\nabla \omega &= \sum_{1 \leqslant i_i < \ldots < i_k \leqslant n} \nabla \omega_{i_1 \ldots i_k} \wedge dx^{i_1} \wedge \ldots \wedge dx^{i_k} \\ &= \sum_{1 \leqslant i_i < \ldots < i_k \leqslant n} \nabla_j \omega_{i_1 \ldots i_k} \, dx^j \wedge dx^{i_1} \wedge \ldots \wedge dx^{i_k}, \end{aligned}$$

where in the last expression, the Einstein summation convention applies to the index j but not to $i_1, \dots, i_k$. One can show by the same argument as for real-valued differential forms that this definition of d_∇ is independent of the choice of coordinates; see Exercise 26.13 below. Note that wedge products $\alpha \wedge \beta$ or $\beta \wedge \alpha \in \Omega^{k+\ell}(M, E)$ can naturally be defined for $\alpha \in \Omega^k(M)$ and $\beta \in \Omega^\ell(M, E)$, but it makes no sense if *both* forms are bundle-valued.

EXERCISE 26.13. Show that $d_\nabla : \Omega^k(M, E) \to \Omega^{k+1}(M, E)$ can be defined as the unique linear operator which matches ∇ on $\Omega^0(M, E)$ and satisfies the graded Leibnitz rule

$$d_\nabla(\alpha \wedge \beta) = d_\nabla \alpha \wedge \beta + (-1)^k \alpha \wedge d\beta$$

for all $\alpha \in \Omega^k(M, E)$ and $\beta \in \Omega^\ell(M)$.

EXERCISE 26.14. Show that for $\lambda \in \Omega^1(M, E)$ written in local coordinates over $\mathcal{U} \subset M$ as $\lambda = \lambda_i \, dx^i$ with $\lambda_1, \dots, \lambda_n \in \Gamma(E|_{\mathcal{U}})$, the component sections for $d_\nabla \lambda$ over $\mathcal{U}$ are given by

$$(d_\nabla \lambda)_{ij} = \nabla_i \lambda_j - \nabla_j \lambda_i.$$

Use this to prove the coordinate-free formula

$$d_\nabla \lambda(X,Y) = \nabla_X(\lambda(Y)) - \nabla_Y(\lambda(X)) - \lambda([X,Y]). \tag{26.4}$$

Hint: For the last step, the main task is to show that the right hand side of (26.4) gives a well-defined bundle-valued 2-form; the rest follows easily from the coordinate formula.

Proof of Theorem 26.9. We will show that both $R(X,Y)v$ and $\Omega_K(X,Y)v$ can be expressed in terms of a covariant exterior derivative of the map $K : TE \to E$. In this context, we regard K as a bundle-valued 1-form $K \in \Omega^1(E, \pi^*E)$ since it maps T_vE linearly to $E_{\pi(v)} = (\pi^*E)_v$. We use the connection ∇ on $\pi : E \to M$ to induce a natural connection on the pullback bundle $\pi^*E \to E$.

We claim first that for any $p \in M$, $v \in E_p$ and $X, Y \in T_pM$,

$$d_\nabla K(\mathrm{Hor}_v(X), \mathrm{Hor}_v(Y)) = \Omega_K(X,Y)v.$$

Indeed, extend X and Y to vector fields on M and use the corresponding horizontal lifts $X^h, Y^h \in \mathfrak{X}(E)$ as extensions of $\mathrm{Hor}_v(X)$ and $\mathrm{Hor}_v(Y) \in T_vE$ respectively. Then using (26.4), (26.2) and the fact that K vanishes on horizontal vectors,

$$d_\nabla K(X^h(v), Y^h(v)) = \nabla_{X^h(v)}\left(K(Y^h)\right) - \nabla_{Y^h(v)}\left(K(X^h)\right) - K([X^h, Y^h](v)) = \Omega_K(X,Y)v.$$

We now show that $R(X,Y)v$ can also be expressed in this way. Choose a smooth map $f(s,t) \in M$ for $(s,t) \in \mathbb{R}^2$ near $(0,0)$ such that $\partial_s f(0,0) = X$ and $\partial_t f(0,0) = Y$, and extend $v \in E_p$ to a section $v(s,t) \in E_{f(s,t)}$ along f such that $v(0,0) = v$ and $\nabla_s v(0,0) = \nabla_t v(0,0) = 0$. The latter can always be done e.g. by letting $v(0,0)$ determine the values $v(s,t)$ for all $(s,t) \in \mathbb{R}^2$ near $(0,0)$ via parallel transport along radial paths starting at the origin. (Note that this guarantees $\nabla v = 0$ at $(0,0)$ and also that ∇v vanishes in radial directions elsewhere, but each of $\nabla_s v$ and $\nabla_t v$ might still be nonzero for $(s,t) \neq (0,0)$; we cannot force both of these to vanish at every point unless we already know the connection is flat.) Expressing covariant derivatives via the map K (e.g. $\nabla_s v = K(\partial_s v)$) and applying (26.4) once more along with Exercise 26.8, we then find

$$\begin{aligned} R(X,Y)v &= \nabla_s \nabla_t v(0,0) - \nabla_t \nabla_s v(0,0) = \nabla_s\left(K(\partial_t v(s,t))\right) - \nabla_t\left(K(\partial_s v(s,t))\right)\big|_{(s,t)=(0,0)} \\ &= d_\nabla K(\partial_s v, \partial_t v)\big|_{(s,t)=(0,0)} = d_\nabla K(\mathrm{Hor}_v(X), \mathrm{Hor}_v(Y)), \end{aligned}$$

where in the last step we used the assumption that $v(s,t)$ has vanishing covariant derivatives at $(0,0)$, hence $\partial_s v(0,0)$ and $\partial_t v(0,0)$ are horizontal. □

The exercises below exhibit two further ways that curvature can be expressed in terms of exterior derivatives.

Exercise 26.15. For a connection ∇ on the bundle $\pi : E \to M$, prove:

(a) For any $v \in \Gamma(E) = \Omega^0(M,E)$ and $X, Y \in T_pM$ at a point $p \in M$, $d_\nabla^2 v := d_\nabla(d_\nabla v) \in \Omega^2(M,E)$ satisfies

$$(d_\nabla^2 v)(X,Y) = R(X,Y)v.$$

(b) The connection ∇ is flat if and only if the covariant exterior derivative operators $d_\nabla : \Omega^k(M,E) \to \Omega^{k+1}(M,E)$ for all $k \geqslant 0$ satisfy $d_\nabla \circ d_\nabla = 0$.

Exercise 26.16. Suppose $\pi : E \to M$ has structure group $G \subset \mathrm{GL}(m,\mathbb{F})$ with Lie algebra $\mathfrak{g} \subset \mathbb{F}^{m\times m}$ and ∇ is a G-compatible connection. Recall that ∇ then associates to every G-compatible local trivialization $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^m$ a connection 1-form $A_\alpha \in \Omega^1(\mathcal{U}_\alpha, \mathfrak{g})$, defined so that

$$(\nabla_X v)_\alpha = \mathcal{L}_X v_\alpha + A_\alpha(X) v_\alpha$$

for any $X \in \mathfrak{X}(\mathcal{U}_\alpha)$, where $v_\alpha : \mathcal{U}_\alpha \to \mathbb{F}^m$ expresses $v|_{\mathcal{U}_\alpha} \in \Gamma(E|_{\mathcal{U}_\alpha})$ with respect to the trivialization. The corresponding **local curvature 2-form** $F_\alpha \in \Omega^2(\mathcal{U}_\alpha, \mathbb{F}^{m\times m})$ is defined as the local representation of $\Omega_K \in \Omega^2(M, \mathrm{End}(E))$ with respect to this trivialization, meaning that for $X, Y \in \mathfrak{X}(\mathcal{U}_\alpha)$ and $v \in \Gamma(E|_{\mathcal{U}_\alpha})$,

$$(\Omega_K(X,Y)v)_\alpha = F_\alpha(X,Y)v_\alpha.$$

(a) Prove the formula

$$F_\alpha(X,Y) = dA_\alpha(X,Y) + [A_\alpha(X), A_\alpha(Y)],$$

where the bracket on the right hand side denotes the matrix commutator $[\mathbf{A}, \mathbf{B}] := \mathbf{AB} - \mathbf{BA}$ for $\mathbf{A}, \mathbf{B} \in \mathbb{F}^{m\times m}$.
Hint: Use the Riemann tensor as a stand-in for Ω_K.

(b) If $\Phi_\beta : E|_{\mathcal{U}_\beta} \to \mathcal{U}_\beta \times \mathbb{F}^m$ is a second trivialization related to Φ_α by the transition map $g = g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to G$, show that

$$F_\beta(X,Y) = gF_\alpha(X,Y)g^{-1}.$$

(c) Show that if G is abelian, then $F_\alpha = dA_\alpha$ and it is independent of the choice of trivialization, thus defining a global 2-form $F \in \Omega^2(M, \mathfrak{g})$. (It is sometimes also called the *curvature 2-form* of ∇.)

Remark: By a basic result in the theory of Lie groups, the commutator $[\mathbf{A}, \mathbf{B}]$ belongs to $\mathfrak{g}$ whenever $\mathbf{A}, \mathbf{B} \in \mathfrak{g}$; this is the reason why $\mathfrak{g}$ is called the "Lie algebra" of G. It thus follows from part (a) that $F_\alpha \in \Omega^2(\mathcal{U}_\alpha, \mathfrak{g})$. In the case $G = \mathrm{O}(k,\ell)$, this is a locally trivialized analogue of Exercise 26.11, which showed that Ω_K takes values in the bundle of antisymmetric *linear maps $E \to E$.*

27. Curvature in pseudo-Riemannian manifolds

For the remainder of the semester, we discuss properties and applications of the curvature of the Levi-Cività connection on the tangent bundle of a Riemannian (or occasionally pseudo-Riemannian) manifold.

27.1. The covariant Riemann tensor. When the bundle under consideration is the tangent bundle of a manifold M, the Riemann tensor defines a multilinear map $TM^{\oplus 3} \to TM : (X,Y,Z) \mapsto R(X,Y)Z$ that can be regarded as a type $(1,3)$ tensor field,

$$R \in \Gamma(T_3^1 M), \qquad R(X,Y)Z = \nabla_X\nabla_Y Z - \nabla_Y\nabla_X Z - \nabla_{[X,Y]}Z.$$

Assuming ∇ is the Levi-Cività connection for a metric g, we have observed two nontrivial relations so far that R must satisfy: one is the antisymmetry

$$R(X,Y)Z = -R(Y,X)Z$$

that is obvious from its definition, and the other (from Exercise 26.11) is

$$\langle V, R(X,Y)Z\rangle + \langle R(X,Y)V, Z\rangle = 0.$$

We saw in §26.3 that when (M,g) is a 2-dimensional Riemannian manifold, these two relations imply that R is determined by a real-valued function—we'll have more to say about that below. (A version of this is also true for indefinite metrics in dimension two; see Exercise 27.1 below.) For certain purposes, it is sometimes useful to repackage the Riemann tensor as a fully covariant tensor $\mathrm{Riem} \in \Gamma(T_4^0 M)$ defined by

$$\mathrm{Riem}(V,X,Y,Z) := \langle V, R(X,Y)Z\rangle.$$

This tensor contains all the same information, and R can be recovered from it; it is essentially the result of applying a musical isomorphism $\flat : T_3^1 M \to T_4^0 M$ that associates to any $S \in \Gamma(T_3^1 M)$ the tensor $S_\flat \in \Gamma(T^0_4 M)$ defined by $S_\flat(V, X, Y, Z) := S(V_\flat, X, Y, Z)$. The two antisymmetry relations mentioned above are now equivalent to

$$\operatorname{Riem}(V, X, Y, Z) = -\operatorname{Riem}(V, Y, X, Z) \qquad \text{and} \qquad \operatorname{Riem}(V, X, Y, Z) = -\operatorname{Riem}(Z, X, Y, V). \tag{27.1}$$

In local coordinates, writing $R^i{}_{jk\ell}\partial_i = R(\partial_j, \partial_k)\partial_\ell$ for the components of R, the components of Riem are traditionally written with the same symbol but a lowered index, hence

$$R_{ijk\ell} := \operatorname{Riem}(\partial_i, \partial_j, \partial_k, \partial_\ell) = \langle \partial_i, R(\partial_j, \partial_k)\partial_\ell \rangle = \langle \partial_i, R^m{}_{jk\ell}\partial_m \rangle = g_{im} R^m{}_{jk\ell}.$$

The Riemann tensor satisfies additional symmetry relations beyond (27.1) that we will talk about next semester, but we will not yet need them at present.

Exercise 27.1. Assuming (M, g) is a 2-dimensional pseudo-Riemannian manifold, use the antisymmetry relations (27.1) to show that in any local coordinate system on an open set $\mathcal{U} \subset M$, the Riemann tensor is determined on $\mathcal{U}$ by the single component function $R_{1122} : \mathcal{U} \to \mathbb{R}$.

Exercise 27.2. The **Ricci curvature** is a tensor $\operatorname{Ric} \in \Gamma(T^0_2 M)$ derived from the Riemann tensor that plays a vital role in more advanced topics in Riemannian geometry, and also in general relativity. If (M, g) is a Riemannian manifold, it can be defined at any point $p \in M$ by choosing an orthonormal basis $e_1, \ldots, e_n \in T_pM$ and writing

$$\operatorname{Ric}(Y, Z) := \sum_{j=1}^n \langle e_j, R(e_j, Y)Z \rangle = \sum_{j=1}^n \operatorname{Riem}(e_j, e_j, Y, Z) \in \mathbb{R}, \qquad \text{for } Y, Z \in T_pM. \tag{27.2}$$

You can convince yourself as follows that this is well defined:

(a) Use the Einstein summation convention to give a one-line proof that $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA})$ for all pairs of square matrices $\mathbf{A}$ and $\mathbf{B}$.
(b) Show that for linear maps $A : V \to V$ on a finite-dimensional vector space V, $\operatorname{tr}(A)$ can be defined as the trace of any matrix representing A in a basis, and it is independent of the choice of basis.
(c) Show that $\operatorname{Ric}(Y, Z)$ according to (27.2) is the trace of the linear map $T_pM \to T_pM : X \mapsto R(X, Y)Z$.
*Remark: This use of the trace demonstrates a general algebraic operation that can transform any tensor of type $(k+1, \ell+1)$ into a tensor of type (k, ℓ); it is known as a **contraction**. Notice that this also gives a definition of* Ric *that does not refer to the metric, and thus makes sense for an arbitrary connection on TM, including the Levi-Cività connection of an indefinite metric. (The formula (27.2) is not quite right in the indefinite case—can you see why not?)*
(d) Show that in local coordinates, the components $R_{k\ell}$ of Ric are given by $R_{k\ell} = R^i{}_{ik\ell}$.

A further simplification of the curvature tensor on a Riemannian manifold (M, g) can be obtained by contracting the Ricci tensor, giving rise to the **scalar curvature**

$$\operatorname{Scal}(p) := \sum_{j=1}^n \operatorname{Ric}(e_j, e_j) = \sum_{j,k=1}^n \operatorname{Riem}(e_j, e_j, e_k, e_k) \in \mathbb{R}, \tag{27.3}$$

where $e_1, \ldots, e_n \in T_pM$ again denotes an orthonormal basis. This defines a function $\operatorname{Scal} : M \to \mathbb{R}$.

(e) Show that (27.3) is independent of the choice of orthonormal basis $e_1, \ldots, e_n \in T_pM$ by reinterpreting it as the trace of the map $\operatorname{Ric}^\sharp : T_pM \to T_pM$ defined via the relation $\langle Y, \operatorname{Ric}^\sharp(Z) \rangle = \operatorname{Ric}(Y, Z)$ for $Y, Z \in T_pM$.
(f) Taking the trace in part (e) as a general definition of $\operatorname{Scal} : M \to \mathbb{R}$ for pseudo-Riemannian manifolds (M, g), rewrite (27.3) so that it is also valid when g is indefinite. (Note that

unlike Ric, Scal does depend explicitly on g and not just on the connection, as the definition of $\mathrm{Ric}^\sharp$ depends on g.)

(g) Prove that if $\dim M = 2$, then the entire Riemann tensor is determined at each point $p \in M$ by the number $\mathrm{Scal}(p)$.
Hint: Use Exercise 27.1 in coordinates chosen so that the coordinate vector fields are orthonormal at p.

(h) Show that in local coordinates, $\mathrm{Scal} = g^{k\ell} R^i{}_{ik\ell}$.

27.2. Locally flat metrics. A pseudo-Riemannian manifold (M, g) of dimension n is called **locally flat** if every point $p \in M$ admits a neighborhood $\mathcal{U} \subset M$ with a chart $(x^1, \dots, x^n) : \mathcal{U} \to \mathbb{R}^n$ in which the components $g_{ij} = \langle \partial_i, \partial_j \rangle$ of the metric are constant functions. Recall that for metrics of signature (k, ℓ), a metric with constant components is equivalent via a linear transformation to the standard *flat metric* η of the same signature, which has components

$$\eta_{ij} := \begin{cases} 1 & \text{if } i = j \leqslant k, \\ -1 & \text{if } i = j > k, \\ 0 & \text{if } i \neq j. \end{cases}$$

Thus (M, g) is locally flat if and only if it is *locally isometric* to the flat space $(\mathbb{R}^n, \eta)$, meaning every point has a neighborhood isometric to an open subset of $(\mathbb{R}^n, \eta)$.

We saw in §23.1 that it is always possible to find coordinates making g_{ij} match η_{ij} up to first order at a given point. Achieving $g_{ij} \equiv \eta_{ij}$ on an open neighborhood however is much more ambitious, and not usually possible. It requires an integrability condition, namely the vanishing of the curvature:

THEOREM 27.3. *A pseudo-Riemannian manifold (M, g) is locally flat if and only if its Riemann curvature tensor vanishes.*

PROOF. If $p \in M$ admits a neighborhood with a chart in which the components g_{ij} are constant, then the Christoffel symbols for the Levi-Cività connection vanish on this neighborhood, and by Exercise 26.6, so do the components $R^i{}_{jk\ell}$ of the Riemann tensor. Conversely, if $R \equiv 0$, then the Levi-Cività connection is flat, implying that any orthonormal basis $X_1, \dots, X_n$ of T_pM can be extended to a neighborhood $\mathcal{U} \subset M$ of p as a family of parallel vector fields that form a frame for TM over $\mathcal{U}$. Since the connection is compatible with the metric, this frame is also orthonormal at every point, meaning $g(X_i, X_j) \equiv \eta_{ij}$. By the symmetry of the connection, we also have

$$[X_i, X_j] = \nabla_{X_i} X_j - \nabla_{X_j} X_i \equiv 0$$

since the vector fields $X_1, \dots, X_n$ are all parallel. Theorem 25.11 now produces a chart near p in which $X_1, \dots, X_n$ are the coordinate vector fields, and the components of g in this chart are precisely the constants η_{ij}. □

EXERCISE 27.4. Prove that every Riemannian 1-manifold is locally flat. Give a direct proof, without mentioning the Riemann tensor. (You may have noticed that the latter vanishes for algebraic reasons whenever $\dim M = 1$.)

27.3. Gaussian curvature. The lowest dimension in which curvature is an interesting concept is 2. It was mentioned in §26.3 that for the Levi-Cività connection on a Riemannian 2-manifold (Σ, g), the Riemann curvature tensor is fully determined by a globally-defined real-valued function $K : \Sigma \to \mathbb{R}$. We would now like to clarify what geometric information this function carries, especially for surfaces embedded in $\mathbb{R}^3$.

We would also like to include the hyperbolic plane in this discussion, so in the following, we assume $\mathbb{R}^3$ with coordinates (x, y, z) is endowed with either the Euclidean or the Minkowski metric

$$g = \pm dx^2 + dy^2 + dz^2,$$

and $\Sigma \subset \mathbb{R}^3$ is a 2-dimensional Riemannian submanifold without boundary. For simplicity we also assume for now that Σ is orientable, though we will see that this assumption can be lifted. We will use the symbol

$$S^2_\pm := \left\{ X \in \mathbb{R}^3 \mid \langle X, X \rangle = \pm 1 \right\} \subset \mathbb{R}^3$$

to denote either the unit sphere $S^2_+ := S^2$ or the two-sheeted hyperboloid $S^2_- := \{x^2 - y^2 - z^2 = 1\}$, depending on whether $\langle\ ,\ \rangle$ is the Euclidean or the Minkowski metric. An orientation of Σ now determines a unit normal vector field,

$$\nu \in \Gamma(T\Sigma^\perp) \subset \Gamma(T\mathbb{R}^3|_\Sigma), \qquad \langle \nu, \nu \rangle = \pm 1,$$

which is unique if we require that for every $p \in \Sigma$ and every positively-oriented basis (X, Y) of $T_p\Sigma$, $(\nu(p), X, Y)$ is a positively-oriented basis of $T_p\mathbb{R}^3 = \mathbb{R}^3$. Note that the sign of $\langle \nu, \nu \rangle$ is determined by the signature of $(\mathbb{R}^3, g)$: since we have assumed $\langle\ ,\ \rangle$ is positive on $T\Sigma$, it must be positive on $T\Sigma^\perp$ if g is the Euclidean metric and negative for the Minkowski metric. This means that if we use the canonical isomorphisms $T_p\mathbb{R}^3 = \mathbb{R}^3$ to view ν as a map from Σ into $\mathbb{R}^3$, then it takes values in the submanifold $S^2_\pm$, giving a smooth map between surfaces

$$\nu : \Sigma \to S^2_\pm.$$

This is called the **Gauss map** of Σ. Its derivative at any point $p \in \Sigma$ has the following interesting property: $T_{\nu(p)}S^2_\pm \subset \mathbb{R}^3$ is the orthogonal complement of $\nu(p)$, which is by definition the same subspace as $T_p\Sigma$, so the tangent map $T_p\nu$ defines a linear map of $T_p\Sigma$ to *itself*,

$$T_p\nu : T_p\Sigma \to T_p\Sigma.$$

LEMMA 27.5. *The map $T_p\nu : T_p\Sigma \to T_p\Sigma$ is self-adjoint with respect to the inner product $\langle\ ,\ \rangle$.*

EXERCISE 27.6. Prove the lemma by showing that in some neighborhood $\mathcal{U} \subset \mathbb{R}^3$ of any point $p \in \Sigma$, ν can always be viewed as the restriction to Σ of the gradient of a function $f : \mathcal{U} \to \mathbb{R}$ for which $\Sigma \cap \mathcal{U} = f^{-1}(0)$. (Another proof of Lemma 27.5 will follow from more general considerations in the next section—see Remark 28.4.)

Applying the spectral theorem for self-adjoint operators, we conclude from Lemma 27.5 that $T_p\Sigma$ has an orthonormal basis X_1, X_2 consisting of eigenvectors of $T_p\nu$. The corresponding eigenspaces are called the **principal directions** of Σ at p, and their eigenvalues $\kappa_1, \kappa_2 \in \mathbb{R}$ with

$$T_p\nu(X_i) = \kappa_i X_i \qquad \text{for } i = 1, 2$$

are called the **principal curvatures** at p.

The principal curvatures at p can be interpreted in terms of the curvature of paths on Σ passing through p. In particular, fix a unit vector $X \in T_p\Sigma$ and choose a smooth path $\gamma : (-\epsilon, \epsilon) \to \Sigma$ with unit speed passing through $\gamma(0) = p$ such that $\dot\gamma(0) = X$. We can make some immediate observations about γ: first, since $\langle \dot\gamma(t), \dot\gamma(t) \rangle = 1$ is constant in t, differentiating it at $t = 0$ implies

$$\ddot\gamma(0) \in X^\perp \subset \mathbb{R}^3.$$

Second, $\langle \dot\gamma(t), \nu(\gamma(t)) \rangle = 0$ for all t since $\dot\gamma$ is tangent to Σ, and differentiating this at $t = 0$ then yields the relation

$$-\langle \ddot\gamma(0), \nu(p) \rangle = \langle X, T_p\nu(X) \rangle =: \kappa_n(X), \tag{27.4}$$

implying that the component of $\ddot{\gamma}(0)$ pointing orthogonally to Σ depends only on the unit vector X and not on the choice of path γ. The number $\kappa_n(X) \in \mathbb{R}$ is called the **normal curvature** of Σ at p in the direction X.

REMARK 27.7. One popular interpretation of the normal curvature $\kappa_n(X)$ is expressed in terms of *plane curves*. Suppose $P \subset \mathbb{R}^3$ is a plane and $C \subset P$ is a 1-dimensional submanifold with a choice of normal vector field $\mathbf{n} \in \Gamma(TP|_C)$ along C. At any point $q \in C$, choose a smooth curve $\gamma : (-\epsilon, \epsilon) \to P$ through $\gamma(0) = q$ with unit speed $|\dot{\gamma}| \equiv 1$ that parametrizes a neighborhood of q in C. Differentiating the relation $\langle \dot{\gamma}(t), \dot{\gamma}(t) \rangle = 1$ then reveals that $\ddot{\gamma}(t)$ is always orthogonal to $\dot{\gamma}(t)$, hence

$$\ddot{\gamma}(t) = \kappa(\gamma(t))\mathbf{n}(\gamma(t))$$

for a uniquely-determined function $\kappa : C \to \mathbb{R}$. This function is independent of the choice of path γ parametrizing C; in particular, reversing the direction of γ does not change its second derivative as it passes through the same point. In this context, $\kappa : C \to \mathbb{R}$ is called the **curvature** of the curve $C \subset P$.

For the surface Σ with unit vector $X \in T_p\Sigma$, define $P \subset \mathbb{R}^3$ as the unique plane that contains p such that T_pP is spanned by X and $\nu(p)$. The intersection $P \cap \Sigma$ is then a smooth 1-dimensional submanifold near p, and the path γ in (27.4) can be chosen to be a parametrization of this submanifold, in which case $\nu(p)$ spans the orthogonal complement of $\dot{\gamma}(0) = X$ in T_pP. The normal curvature $\kappa_n(X)$ is therefore the curvature of the curve $P \cap \Sigma$ in the plane P at p.

REMARK 27.8. Yet another interpretation of $\kappa_n(X)$ comes from comparing geodesics in Σ with geodesics in the ambient space $\mathbb{R}^3$, also known as straight lines. If we choose γ in (27.4) to be the unique geodesic in Σ with initial velocity X, then Corollary 24.12 tells us $\ddot{\gamma}(0)$ is a scalar multiple of $\nu(p)$, and thus vanishes if and only if $\kappa_n(X) = 0$. From this perspective, $\kappa_n(X)$ measures the extent to which the geodesic in Σ with $\dot{\gamma}(0) = X$ deviates from being a geodesic in $\mathbb{R}^3$.

Fixing the orthonormal eigenvectors $X_1, X_2 \in T_p\Sigma$ of $T_p\nu$, every other unit vector takes the form $X = aX_1 + bX_2$ with $a^2 + b^2 = 1$, thus by (27.4)

$$\kappa_N(X) = \langle X, T_p\nu(X) \rangle = \langle aX_1 + bX_2, a\kappa_1 X_1 + b\kappa_2 X_2 \rangle = a^2\kappa_1 + b^2\kappa_2.$$

The range of values this number can take is precisely the interval in $\mathbb{R}$ bounded by the numbers κ_1 and κ_2, so this proves:

PROPOSITION 27.9. *The principal curvatures of $\Sigma \subset \mathbb{R}^3$ at $p \in \Sigma$ are the maximum and minimum values of the normal curvatures $\kappa_n(X)$ for all unit vectors $X \in T_p\Sigma$.* □

Normal and principal curvatures are measurements of what is called the **extrinsic** curvature of Σ: they depend not just on the Riemannian metric of Σ but also on the way that Σ is embedded in $\mathbb{R}^3$. By contrast, the next object we will define is **intrinsic**, meaning it depends only on the metric and is thus an invariant of Riemannian 2-manifolds (Σ, g) up to isometry. This will not be obvious from the definition—proving that it is intrinsic will require a substantial effort.

DEFINITION 27.10. For a Riemannian hypersurface Σ in $\mathbb{R}^3$ with the Euclidean or Minkowski metric $g = \pm dx^2 + dy^2 + dz^2$, the **Gaussian curvature** of Σ at $p \in \Sigma$ is defined (up to a sign) as the product of its principal curvatures, that is,

$$K_G(p) := \pm\kappa_1\kappa_2 \in \mathbb{R}, \tag{27.5}$$

where the symbol $\pm$ means $+$ if g is the Euclidean metric and $-$ for the Minkowski metric. Equivalently, $K_G(p)$ is determined from the Gauss map $\nu : \Sigma \to S^2_\pm$ as

$$K_G(p) = \pm \det\left(T_p\Sigma \xrightarrow{T_p\nu} T_p\Sigma\right).$$

REMARK 27.11. For an arbitrary n-dimensional vector space V over the field $\mathbb{F}$, one can define the determinant of a linear map $A : V \to V$ as $\det(\mathbf{A}) \in \mathbb{F}$ where $\mathbf{A} \in \mathbb{F}^{n\times n}$ is the matrix representing A in any choice of basis. The result is independent of the choice of basis since for any $\mathbf{B} \in \mathrm{GL}(n, \mathbb{F})$, $\det(\mathbf{BAB}^{-1}) = \det(\mathbf{A})$.

REMARK 27.12. The normal and principal curvatures all depend on the choice of normal vector field ν, but the Gaussian curvature does not, because reversing ν causes a sign change in both κ_1 and κ_2, leaving $K_G(p)$ invariant. For this reason, the Gaussian curvature can be defined even if Σ is not orientable.

For surfaces in Euclidean space, the formula $K_G(p) = \det(T_p\nu)$ implies that the Gaussian curvature is positive in any region where the Gauss map is orientation preserving, and negative wherever it is orientation reversing. It vanishes at any point where $T_p\nu$ collapses $T_p\Sigma$ to a subspace of lower dimension.

EXAMPLE 27.13. For the unit sphere $S^2 \subset \mathbb{R}^3$ in Euclidean space, the Gauss map is simply the identity, so $K_G \equiv 1$.

EXAMPLE 27.14. Consider the cylinder $Z = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1\}$ in Euclidean space. The Gauss map on Z is independent of z, thus $T_p\nu$ only has rank 1 at every $p \in Z$, implying $K(p) = 0$. By Theorem 27.3 and Theorem 27.17 below, this result is equivalent to the observation that Z is locally flat: unlike a sphere, a small piece of a cylinder can easily be unfolded into a piece of a flat plane without changing lengths or angles on the surface. The same is true of the cone

$$C = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = z^2,\ z > 0\}.$$

It is easy to check that Z and C do have nontrivial normal and principal curvatures, showing that the latter are indeed extrinsic, i.e. they depend on the specific embeddings of these surfaces in $\mathbb{R}^3$ and are not isometry invariants.

EXAMPLE 27.15. The hyperboloid $H = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 - z^2 = 1\}$ in Euclidean space has everywhere negative curvature. (For a precise computation, see Exercise 27.20 below.) This is true of any surface that exhibits a "saddle" shape, for which the Gauss map is orientation reversing.

EXAMPLE 27.16. The hyperbolic plane H^2 was defined in §24.4.3 as the upper sheet of the hyperboloid S^2_- in $\mathbb{R}^3$ with the Minkowski metric. This would have positive curvature if it lived in Euclidean space, but in Minkowski space the extra sign in Definition 27.10 becomes relevant, so the curvature is negative. The situation is in fact very much analogous to the sphere in Example 27.13, because the Gauss map in this case is just the identity map on the upper sheet of S^2_-, giving $\det(T_p\nu) = 1$ at every point. We conclude $K_G \equiv -1$.

The next big result says that $K_G : \Sigma \to \mathbb{R}$ is determined by the Riemann curvature tensor, and therefore by the Riemannian metric on Σ. In fact, K_G turns out to be the same function that appeared in (26.3):

THEOREM 27.17. *Suppose Σ is an oriented Riemannian hypersurface embedded in Euclidean or Minkowski $\mathbb{R}^3$, $\mathrm{dvol} \in \Omega^2(\Sigma)$ denotes its Riemannian area form, $K_G : \Sigma \to \mathbb{R}$ is its Gaussian curvature, $R(X, Y)Z$ is its Riemann curvature tensor and $J : T\Sigma \to T\Sigma$ is the unique fiberwise linear map such that for any vector $X \in T_p\Sigma$ with $|X| = 1$, (X, JX) is a positively-oriented orthonormal basis. Then*

$$R(X, Y)Z = -K_G\,\mathrm{dvol}(X, Y)JZ.$$

We will prove this theorem in the next lecture. For arbitrary Riemannian 2-manifolds Σ, not embedded in $\mathbb{R}^3$, Theorem 27.17 can be taken as a definition of the Gaussian curvature $K_G : \Sigma \to \mathbb{R}$. Note that once again the result doesn't actually depend on an orientation (cf. Remark 27.12):

locally, if the orientation of Σ is flipped, this changes the sign of both J and $d\mathrm{vol}$, leaving the function K_G unchanged.

For surfaces in Euclidean $\mathbb{R}^3$, Theorem 27.17 implies the following famous result of Gauss, which has come to be known by the Latin term for "remarkable theorem":

THEOREMA EGREGIUM. *For a surface Σ embedded in Euclidean $\mathbb{R}^3$, the Gaussian curvature $K_G : \Sigma \to \mathbb{R}$ defined in (27.5) is an invariant of the induced Riemannian metric on Σ. To be precise, if $\Sigma_1, \Sigma_2 \subset \mathbb{R}^3$ are two surfaces embedded in $\mathbb{R}^3$ with induced metrics g_1, g_2 and Gaussian curvatures K_G^1, K_G^2 respectively, and $\varphi : (\Sigma_1, g_1) \to (\Sigma_2, g_2)$ is an isometry, then*

$$K_G^1 \equiv K_G^2 \circ \varphi.$$

Example 27.14 shows that nothing similar to the Theorema Egregium is true for the normal or principal curvatures of a surface. Here are a couple of sample applications:

- There are no isometries between any open subsets of the sphere $S^2 \subset \mathbb{R}^3$ (positive curvature) and the hyperboloid of Example 27.15 or the hyperbolic plane in Example 27.16 (negative curvature).
- A Riemannian 2-manifold Σ embedded in Euclidean or Minkowski $\mathbb{R}^3$ is locally flat if and only if at least one of its principal curvatures vanishes at every point.

EXERCISE 27.18. Given a constant $r > 0$, compute K_G for:

(a) The sphere $\{x^2 + y^2 + z^2 = r^2\}$ of radius r in Euclidean $\mathbb{R}^3$;
(b) The rescaled hyperbolic plane $\{x^2 - y^2 - z^2 = r^2,\ x > 0\}$ in Minkowski $\mathbb{R}^3$.

We can deduce from Theorem 27.17 a formula for K_G in terms of the Riemann tensor. We begin by observing that the metric $\langle\ ,\ \rangle$, area form $d\mathrm{vol} \in \Omega^2(\Sigma)$ and fiberwise-linear map $J \in \Gamma(\mathrm{End}(T\Sigma))$ satisfy the relation

$$d\mathrm{vol}(X, Y) = \langle JX, Y\rangle. \tag{27.6}$$

To see this, notice first that $(X, Y) \mapsto \langle JX, Y\rangle$ is an alternating 2-form, since J is an orthogonal transformation with $J^2 = -\mathbb{1}$, so

$$\langle JY, X\rangle = \langle J(JY), JX\rangle = \langle -Y, JX\rangle = -\langle JX, Y\rangle.$$

The 2-form $\langle J\cdot, \cdot\rangle$ is therefore a scalar multiple of $d\mathrm{vol}$ at every point, so it suffices to check that they match when evaluated on some particular basis at each point. This is true for instance for any basis of the form (X, JX) with $|X| = 1$, as this basis is positively oriented and orthonormal, so $d\mathrm{vol}(X, JX) = 1 = \langle JX, JX\rangle$, proving (27.6). Theorem 27.17 now implies

$$\langle R(X, Y)Y, X\rangle = -\langle K_G\ d\mathrm{vol}(X, Y)JY, X\rangle = -K_G\ d\mathrm{vol}(X, Y)\langle JY, X\rangle = K_G \cdot |d\mathrm{vol}(X, Y)|^2,$$

so we can write

$$K_G(p) = \frac{\langle R(X, Y)Y, X\rangle}{|d\mathrm{vol}(X, Y)|^2} = \frac{\mathrm{Riem}(X, X, Y, Y)}{|d\mathrm{vol}(X, Y)|^2} \qquad \text{for any basis } X, Y \in T_p\Sigma. \tag{27.7}$$

We can rewrite this as follows in terms of an oriented coordinate chart (x^1, x^2) defined near p. If the components of the metric are denoted by g_{ij} and we define the symmetric matrix-valued function

$$\mathbf{g} = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix},$$

we recall from Exercise 11.12 that $d\mathrm{vol}$ takes the form

$$d\mathrm{vol} = \sqrt{\det \mathbf{g}}\ dx^1 \wedge dx^2.$$

Then applying (27.7) to the coordinate vectors $X = \partial_1$ and $Y = \partial_2$, we obtain the formula

$$K_G = \frac{R_{1122}}{\det \mathbf{g}}. \tag{27.8}$$

If you did not already understand why the Theorema Egregium follows from Theorem 27.17, we can now prove it as follows:

PROOF OF THE THEOREMA EGREGIUM. If $\varphi : (\Sigma_1, g_1) \to (\Sigma_2, g_2)$ is an isometry and $p \in \Sigma_1$, then any chart $(\mathcal{U}, x)$ on a neighborhood $\mathcal{U} \subset \Sigma_2$ of $q := \varphi(p)$ gives rise to a chart $(\varphi^{-1}(\mathcal{U}), x \circ \varphi)$ on a neighborhood $\varphi^{-1}(\mathcal{U}) \subset \Sigma_1$ of p such that the components of the two metrics in these charts are related by $(g_1)_{ij} = (g_2)_{ij} \circ \varphi$. It follows that the components of their Riemann tensors and their Riemannian volume forms satisfy a similar relation, so by (27.8), so do their Gaussian curvatures. ☐

EXERCISE 27.19. A Riemannian manifold (M, g) is called **homogeneous** if for every pair of points $p, q \in M$, there exists an isometry $\varphi \in \mathrm{Isom}(M, g)$ such that $\varphi(p) = q$. Show that every homogeneous Riemannian 2-manifold has constant Gaussian curvature.
Remark: This partly explains why I claimed in §24.2 that one should not generally expect nontrivial isometries to exist. Constant curvature is a very delicate condition that is easy to destroy via small perturbations of the metric.

EXERCISE 27.20. Prove that for the hyperboloid $H \subset \mathbb{R}^3$ in Example 27.15,

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

Hint: This can be a horrible computation, but it doesn't have to be. For instance, there are some obvious isometries that make it sufficient to consider a point of the form $(r, 0, z) \in H$ with $r^2 - z^2 = 1$, which is the intersection of the smooth curves $\alpha(t) = (\cosh t, 0, \sinh t)$ and $\beta(t) = (r \cos t, r \sin t, z)$ in H. Since H is a level set of $f(x, y, z) = x^2 + y^2 - z^2$, there is a unit normal vector field of the form $\nu = g \cdot \nabla f$ for some function $g : H \to (0, \infty)$. Try to convince yourself without any calculations that the curves α and β are tangent to the principal directions. Then consider the following: if you know $\gamma(t) \in H$ satisfies $\frac{d}{dt}\nu(\gamma(t)) = \lambda\dot{\gamma}(t)$ for some $\lambda \in \mathbb{R}$, what happens if you take the inner product of both sides with $\dot{\gamma}(t)$? Write $\nu = g \cdot \nabla f$ and use this observation to compute the two principal curvatures at $(r, 0, z)$. You will need to write down the function g for this, but you should not need to differentiate it.
Final remark: It's also possible there's an easier way to do this that I haven't thought of.

EXERCISE 27.21. Show that for any Riemannian 2-manifold (Σ, g), the scalar curvature defined in Exercise 27.2 is related to the Gaussian curvature by $\mathrm{Scal} = 2K_G$.
Hint: Given a point $p \in \Sigma$, use coordinates for which ∂_1 and ∂_2 are orthonormal at p.

EXERCISE 27.22. Show that the Poincaré half-plane $(\mathbb{H}, h)$ from Exercise 22.8 has constant Gaussian curvature $K_G \equiv -1$.
Remark: You knew this already from Example 27.16 if you had already convinced yourself that $(\mathbb{H}, h)$ is isometric to the hyperbolic plane (see Exercise 24.16). But you can also compute this directly from (27.8) if you first work out the Christoffel symbols of the connection on $(\mathbb{H}, h)$ and then compute the Riemann tensor via Exercise 26.6.

REMARK 27.23. The hyperbolic plane is a funny animal. It is the most famous and most important example of a surface with constant negative curvature—in fact it is known to be the *only* one up to isometry and scaling that is both simply connected and geodesically complete—but you may have noticed that we've never mentioned any model of it that one can look at it and say, "yes, that looks like a surface with negative curvature!". The closest thing we have is the hyperboloid model in Minkowski space, which actually looks like a *positively* curved surface, but acquires an extra minus sign in Definition 27.10, which is difficult to justify intuitively. (The justification for it is that if the minus sign were not there, Theorem 27.17 would not be true.) What I'm getting at is this: it would be nice if we could view H^2 as an embedded hypersurface in *Euclidean* $\mathbb{R}^3$ whose "saddle" shape would make the negativity of its curvature obvious. There exist *local* models of this kind, e.g. the *pseudosphere* (also called the *tractricoid*)[72] is a surface in Euclidean $\mathbb{R}^3$ that is isometric to an open subset of H^2, but not the whole thing. The reason I have not explained any global model of H^2 in Euclidean 3-space is that according to a famous theorem of Hilbert, it is impossible: there exists no embedding (nor even an immersion!) of any geodesically complete surface with constant negative curvature into Euclidean $\mathbb{R}^3$. I'd conjecture that if this theorem were not true, it would have been recognized somewhat earlier in history that Euclid's first four postulates do not imply the fifth.

28. Properties of Gaussian curvature

I owe you a proof of Theorem 27.17, specifically the formula

$$R(X,Y)Z = -K_G \, d\mathrm{vol}(X,Y)JZ,$$

which relates the Gaussian curvature $K_G : \Sigma \to \mathbb{R}$ to the Riemann tensor $R \in \Gamma(T^1_3\Sigma)$ for a Riemannian hypersurface Σ in Euclidean or Minkowski 3-space $(\mathbb{R}^3, \pm dx^2 + dy^2 + dz^2)$. We'll take care of this in §28.1 by developing a general formula to compare the Riemann tensor of any pseudo-Riemannian manifold with that of a pseudo-Riemannian submanifold embedded in it. After that, we will restrict again to dimension 2 and examine some further properties of the Gaussian curvature, in preparation for proving the Gauss-Bonnet formula.

28.1. The second fundamental form. Assume (M,g) is a pseudo-Riemannian manifold with $\dim M > n \geqslant 2$ that contains

$$\Sigma \subset M$$

as an n-dimensional pseudo-Riemannian submanifold with inclusion map $j : \Sigma \hookrightarrow M$, so the induced metric j^*g on Σ is also nondegenerate. In this situation, Corollary 24.9 produces a direct sum decomposition

$$TM|_\Sigma = T\Sigma \oplus T\Sigma^\perp,$$

so that every $X \in T_pM$ for $p \in \Sigma$ is uniquely expressible as

$$X = X^\top + X^\perp, \qquad X^\top \in T_p\Sigma \text{ and } X^\perp \in (T_p\Sigma)^\perp \subset T_pM.$$

In this notation, the Levi-Cività connections ∇ and $\hat{\nabla}$ of (M,g) and (Σ, j^*g) are related according to Proposition 24.11 by

$$\hat{\nabla}_X Y = (\nabla_X Y)^\top.$$

A vector field $X(t) \in T_{\gamma(t)}\Sigma$ on Σ along a path $\gamma(t) \in \Sigma$ is thus parallel if and only if $\nabla_t X \in T\Sigma^\perp$, and since this allows $\nabla_t X$ to be nonzero, X may fail to be parallel when regarded as a vector field on M along γ. This failure can be measured by a tensor:

Lemma 28.1. *There exists a symmetric bilinear bundle map* $\mathrm{II} : T\Sigma \oplus T\Sigma \to T\Sigma^\perp$ *such that for any pair of vector fields* $X, Y \in \mathfrak{X}(\Sigma)$,

$$\mathrm{II}(X,Y) = (\nabla_Y X)^\perp.$$

In particular, the connections ∇ *on* M *and* $\hat{\nabla}$ *on* Σ *are then related to each other by*

$$\nabla_Y X = \hat{\nabla}_Y X + \mathrm{II}(X,Y).$$

[72]See https://en.wikipedia.org/wiki/Pseudosphere

Proof. We can define $\mathrm{II} : \mathfrak{X}(\Sigma) \times \mathfrak{X}(\Sigma) \to \Gamma(T\Sigma^\perp)$ by $\mathrm{II}(X,Y) := (\nabla_Y X)^\perp$. The symmetry of II then follows easily from the symmetry of the Levi-Cività connection: extending X and Y arbitrarily to vector fields on M, we have

$$\mathrm{II}(Y,X) - \mathrm{II}(X,Y) = (\nabla_X Y - \nabla_Y X)^\perp = [X,Y]^\perp = 0,$$

since X and Y taking values in $T\Sigma$ along Σ implies that the same is true for $[X,Y]$. Now since $\mathrm{II}(X,Y)$ is manifestly C^∞-linear in Y, the symmetry implies that it is also C^∞-linear in X, and therefore gives a well-defined bundle map $T\Sigma \oplus T\Sigma \to T\Sigma^\perp$. □

Definition 28.2. The symmetric bundle map $\mathrm{II} : T\Sigma \oplus T\Sigma \to T\Sigma^\perp$ in Lemma 28.1 is called the **second fundamental form**[73] of the submanifold $\Sigma \subset M$.

We can now associate to any normal section $\nu \in \Gamma(T\Sigma^\perp)$ the symmetric tensor field $\mathrm{II}_\nu \in \Gamma(T_2^0\Sigma)$ defined by

$$\mathrm{II}_\nu(X,Y) := \langle \nu, \mathrm{II}(X,Y)\rangle.$$

This is especially useful in the case where $\Sigma \subset M$ is a hypersurface with orientable normal bundle, as Σ then admits a *unit* normal vector field $\nu \in \Gamma(T\Sigma^\perp)$ that is unique up to a sign. The words "second fundamental form" are also sometimes used to refer to the symmetric tensor $\mathrm{II}_\nu \in \Gamma(T_2^0\Sigma)$ in this special case.

Remark 28.3. By now you may be wondering: what is the *first* fundamental form? This term was traditionally used for another symmetric $(0,2)$-tensor on Σ, namely the restricted metric $\langle\ ,\ \rangle|_{T\Sigma} = j^*g$. But the term has fallen somewhat out of fashion.

Since II_ν is a symmetric bilinear form on the tangent spaces of Σ for each normal section $\nu \in \Gamma(T\Sigma^\perp)$, it corresponds via the relation

$$\mathrm{II}_\nu(X,Y) = \langle X, W_\nu(Y)\rangle$$

to a unique bundle map $W_\nu : T\Sigma \to T\Sigma$ that is self-adjoint with respect to the bundle metric on $T\Sigma$. We call W_ν the **Weingarten map** associated to the normal section ν. One obtains a more revealing formula for it by differentiating the relation $\langle X,\nu\rangle \equiv 0$, which holds for any $X \in \mathfrak{X}(\Sigma)$ and $\nu \in \Gamma(T\Sigma^\perp)$: we find

$$\begin{aligned}0 = \mathcal{L}_Y\langle X,\nu\rangle &= \langle \nabla_Y X, \nu\rangle + \langle X, \nabla_Y \nu\rangle = \langle \hat{\nabla}_Y X + \mathrm{II}(X,Y), \nu\rangle + \langle X, (\nabla_Y\nu)^\top + (\nabla_Y\nu)^\perp\rangle \\ &= \mathrm{II}_\nu(X,Y) + \langle X, (\nabla_Y\nu)^\top\rangle = \langle X, W_\nu(Y) + (\nabla_Y\nu)^\top\rangle,\end{aligned}$$

having discarded terms that vanish due to orthogonality. The result is an interpretation of the Weingarten map as the tangential part of the covariant derivative of ν:

$$W_\nu(X) = -(\nabla_X\nu)^\top. \tag{28.1}$$

Remark 28.4. If $\Sigma \subset M$ is a hypersurface with an orientable normal bundle, then there are two canonical choices of $\nu \in \Gamma(T\Sigma^\perp)$ determined by the normalization condition $\langle \nu,\nu\rangle \equiv \pm 1$, where the sign depends on the signatures of (M,g) and (Σ, j^*g). Differentiating $\langle\nu,\nu\rangle$ now reveals that $\langle \nabla_X\nu, \nu\rangle \equiv 0$ for all $X \in \mathfrak{X}(\Sigma)$, and $\nabla_X\nu$ is therefore tangent to Σ, thus (28.1) simplifies to

$$W_\nu(X) = -\nabla_X \nu.$$

This particular form of the Weingarten map is sometimes called the **shape operator**. In the important special case where M is $\mathbb{R}^3$ with the Euclidean or Minkowski metric, ∇ is the trivial connection, so $-W_\nu : T\Sigma \to T\Sigma$ is now the derivative of the Gauss map introduced in §27.3. The

[73]Do not be misled by this use of the word "form"; II is not a differential form in any sense, as it is symmetric rather than antisymmetric.

self-adjointness of W_ν thus gives a second proof of Lemma 27.5, and the Gaussian curvature of (Σ, j^*g) in this situation is $\pm\det(W_\nu)$.

Like the Gauss map and the principal curvatures in §27.3, the Weingarten map and second fundamental form belong to the *extrinsic* rather than *intrinsic* geometry of (Σ, j^*g), meaning they depend on the way that Σ is embedded as a pseudo-Riemannian submanifold of (M, g), rather than intrinsically on the metric j^*g. They are not deeply meaningful objects, but they turn out to be useful tools for deriving the Riemann tensor of (Σ, j^*g) from that of (M, g). In the following, we denote by

$$R \in \Gamma(T_3^1 M), \qquad \hat{R} \in \Gamma(T_3^1 \Sigma)$$

the Riemann curvature tensors of (M, g) and (Σ, j^*g) respectively, along with their covariant versions

$$\mathrm{Riem} \in \Gamma(T_4^0 M), \qquad \widehat{\mathrm{Riem}} \in \Gamma(T_4^0 \Sigma)$$

as defined in §27.1.

Proposition 28.5 (Gauss equation). *The tensors* Riem *and* $\widehat{\mathrm{Riem}}$ *are related by*

$$\widehat{\mathrm{Riem}}(V, X, Y, Z) = \mathrm{Riem}(V, X, Y, Z) + \langle \mathrm{II}(V, X), \mathrm{II}(Y, Z)\rangle - \langle \mathrm{II}(V, Y), \mathrm{II}(X, Z)\rangle.$$

Proof. We observe first that for any tuple of vector fields $V, X, Y, Z \in \mathfrak{X}(\Sigma)$, differentiating the relation $\langle V, \mathrm{II}(Y, Z)\rangle \equiv 0$ with respect to X gives

$$\langle V, \nabla_X (\mathrm{II}(Y, Z))\rangle = -\langle \nabla_X V, \mathrm{II}(Y, Z)\rangle = -\langle \mathrm{II}(X, V), \mathrm{II}(Y, Z)\rangle, \tag{28.2}$$

where $\nabla_X V$ can be replaced by its normal part $\mathrm{II}(X, V)$ in the last expression because the inner product of its tangential part with $\mathrm{II}(Y, Z)$ necesarily vanishes. The same trick allows us in the following calculation to replace $\hat{\nabla}$ with ∇ in several places since we are taking the inner product with V; applying also (28.2) and the relation $\nabla_X Y = \hat{\nabla}_X Y + \mathrm{II}(X, Y)$, we find

$$\begin{aligned}
\widehat{\mathrm{Riem}}(V, X, Y, Z) &= \langle V, \hat{R}(X, Y)Z\rangle = \langle V, \hat{\nabla}_X \hat{\nabla}_Y Z - \hat{\nabla}_Y \hat{\nabla}_X Z - \hat{\nabla}_{[X,Y]} Z\rangle \\
&= \left\langle V, \nabla_X (\nabla_Y Z - \mathrm{II}(Y, Z)) - \nabla_Y (\nabla_X Z - \mathrm{II}(X, Z)) - \nabla_{[X,Y]} Z\right\rangle \\
&= \langle V, R(X, Y)Z\rangle + \langle \mathrm{II}(V, X), \mathrm{II}(Y, Z)\rangle - \langle \mathrm{II}(V, Y), \mathrm{II}(X, Z)\rangle.
\end{aligned}$$

□

Now let's specialize this to a situation closer to that of Theorem 27.17. We assume (M, g) is a locally flat pseudo-Riemannian 3-manifold, and $\Sigma \subset M$ is a Riemannian hypersurface. In this case $T\Sigma^\perp$ is a line bundle over Σ on which the bundle metric $\langle\ ,\ \rangle$ is nondegenerate, and it may be either positive or negative, depending on whether (M, g) has Riemannian signature $(3, 0)$ or Lorentz signature $(2, 1)$, which are the only two possibilites since we are assuming (Σ, g) is Riemannian. As usual it will also be convenient to assume that both Σ and its normal bundle $T\Sigma^\perp$ are orientable, though these assumptions will both be seen to be inessential in the end. Fixing an orientation of Σ determines the fiberwise-linear map

$$J : T\Sigma \to T\Sigma$$

that rotates each tangent space counterclockwise by 90 degrees. The orientability of $T\Sigma^\perp$ allows us in turn to choose a (unique up to a sign) unit normal vector field

$$\nu \in \Gamma(T\Sigma^\perp), \qquad \langle \nu, \nu\rangle \equiv \pm 1,$$

where the sign is positive if (M, g) is Riemannian and negative otherwise; in the following we will make consistent use of the symbol "$\pm$" for this sign, and write "$\mp$" whenever it gets reversed. For example, the symmetric tensors $\mathrm{II}(X, Y)$ and $\mathrm{II}_\nu(X, Y)$ are now related to each other by

$$\mathrm{II}(X, Y) = \pm\, \mathrm{II}_\nu(X, Y)\nu,$$

and in light of Remark 28.4, $\nabla\nu|_{T\Sigma}$ matches the shape operator $-W_\nu : T\Sigma \to T\Sigma$ and is thus related to the second fundamental form by

$$\mathrm{II}(X,Y) = \pm\,\mathrm{II}_\nu(X,Y)\nu = \pm\langle X, W_\nu(Y)\rangle\nu = \mp\langle X, \nabla_Y\nu\rangle\nu. \tag{28.3}$$

Fix a point $p \in \Sigma$ and write $\nabla\nu(p) : T_p\Sigma \to T_p\Sigma$ for the restriction of $\nabla\nu$ to the tangent space at p. The symmetry of II_ν implies that $\nabla\nu(p)$ is self-adjoint with respect to the inner product $\langle\ ,\ \rangle$ on $T_p\Sigma$, thus it has an orthonormal basis of eigenvectors $X_1, X_2 \in T_p\Sigma$, and we are free to order them so that

$$X_2 = JX_1 \qquad \text{and} \qquad X_1 = -JX_2,$$

in which case they are also a positively-oriented basis and thus satisfy

$$d\mathrm{vol}(X_1, X_2) = 1 \tag{28.4}$$

for the Riemannian volume form $d\mathrm{vol} \in \Omega^2(\Sigma)$. In the case where M is the Euclidean or Minkowski $\mathbb{R}^3$, the corresponding eigenvalues

$$\kappa_1, \kappa_2 \in \mathbb{R}$$

are the principal curvatures of Σ at p. We can now use this data to turn Proposition 28.5 into a more explicit formula for the Riemann tensor of (Σ, j^*g) at p: we assumed (M,g) is flat, so $R \equiv 0$, and thus for $V, X, Y, Z \in T_p\Sigma$, using (28.3) to replace various terms in the Gauss equation gives

$$\begin{aligned}
\langle V, \widehat{R}(X_1,X_2)Z\rangle &= \langle \mathrm{II}(V,X_1), \mathrm{II}(X_2,Z)\rangle - \langle \mathrm{II}(V,X_2), \mathrm{II}(X_1,Z)\rangle \\
&= \big\langle \langle V, \nabla\nu(p)X_1\rangle\,\nu(p), \langle Z, \nabla\nu(p)X_2\rangle\,\nu(p)\big\rangle - \big\langle \langle V, \nabla\nu(p)X_2\rangle\,\nu(p), \langle Z, \nabla\nu(p)X_1\rangle\,\nu(p)\big\rangle \\
&= \pm\kappa_1\kappa_2\big(\langle V, X_1\rangle \cdot \langle Z, X_2\rangle - \langle V, X_2\rangle \cdot \langle Z, X_1\rangle\big) \\
&= \pm\kappa_1\kappa_2\big\langle V, \langle Z, X_2\rangle X_1 - \langle Z, X_1\rangle X_2\big\rangle,
\end{aligned}$$

which implies

$$\widehat{R}(X_1,X_2)Z = \pm\kappa_1\kappa_2\left(\langle Z, X_2\rangle X_1 - \langle Z, X_1\rangle X_2\right).$$

Finally, we observe that since J maps $T_p\Sigma \to T_p\Sigma$ orthogonally and the vectors $X_2 = JX_1$ and $X_1 = -JX_2$ form an orthonormal basis,

$$\langle Z, X_2\rangle X_1 - \langle Z, X_1\rangle X_2 = \langle JZ, JX_2\rangle X_1 - \langle JZ, JX_1\rangle X_2 = -\langle X_1, JZ\rangle X_1 - \langle X_2, JZ\rangle X_2 = -JZ,$$

and combining this with (28.4) we thus obtain

$$\widehat{R}(X_1,X_2)Z = \mp\kappa_1\kappa_2 JZ = -\pm\,\kappa_1\kappa_2\, d\mathrm{vol}(X_1,X_2)JZ.$$

We already know there exists a unique function $K : \Sigma \to \mathbb{R}$ such that the relation $\widehat{R}(X_1,X_2)Z = -K(p)\, d\mathrm{vol}(X_1,X_2)JZ$ is satisfied, so the conclusion of this calculation is that $K(p) = \pm\kappa_1\kappa_2$, i.e. it is the Gaussian curvature K_G. This completes the proof of Theorem 27.17.

REMARK 28.6. While $\Sigma \subset M$ was allowed to be a pseudo-Riemannian submanifold of arbitrary signature in most of this section, the positivity of j^*g became essential as soon as we started talking about the orthonormal eigenvectors of the shape operator $\nabla\nu(p) : T_p\Sigma \to T_p\Sigma$. This is indeed a self-adjoint operator with respect to the bundle metric $\langle\ ,\ \rangle$ in every case, but the spectral theorem does not hold in general with indefinite inner products.

28.2. Local curvature 2-forms. We haven't mentioned it in a couple of lectures, but in addition to the Riemann tensor $R \in \Gamma(T^1_3\Sigma)$, the curvature of a Riemannian 2-manifold (Σ, g) can also be characterized via a differential 2-form, the curvature 2-form $\Omega_K \in \Omega^2(\Sigma, \mathrm{End}(T\Sigma))$. You might wonder: what happens if we integrate it? This question doesn't make much sense at first glance, as Ω_K is a bundle-valued 2-form, so it's not clear what $\int_\Sigma \Omega_K$ should mean. In order to clarify this, I'd like to expand on an exercise that was stated at the end of §26.4.

Assume $\pi : E \to M$ is a vector bundle with structure group $G \subset \mathrm{GL}(m, \mathbb{F})$, denote the Lie algebra of G by $\mathfrak{g} \subset \mathbb{F}^{m\times m}$, and suppose ∇ is a G-compatible connection. We recall that for every G-compatible local trivialization $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^m$, ∇ can be described over $\mathcal{U}_\alpha$ via a connection 1-form $A_\alpha \in \Omega^1(\mathcal{U}_\alpha, \mathfrak{g})$, defined so that

$$(\nabla_X v)_\alpha = \mathcal{L}_X v_\alpha + A_\alpha(X) v_\alpha \tag{28.5}$$

for any $X \in \mathfrak{X}(\mathcal{U}_\alpha)$. Here $v_\alpha : \mathcal{U}_\alpha \to \mathbb{F}^m$ expresses $v|_{\mathcal{U}_\alpha} \in \Gamma(E|_{\mathcal{U}_\alpha})$ with respect to the trivialization, meaning $\Phi_\alpha(v(p)) = (p, v_\alpha(p))$ for $p \in \mathcal{U}_\alpha$. The corresponding **local curvature 2-form** $F_\alpha \in \Omega^2(\mathcal{U}_\alpha, \mathbb{F}^{m\times m})$ is defined as the local representation of $\Omega_K \in \Omega^2(M, \mathrm{End}(E))$ with respect to this trivialization, meaning that for $X, Y \in \mathfrak{X}(\mathcal{U}_\alpha)$ and $v \in \Gamma(E|_{\mathcal{U}_\alpha})$,

$$(\Omega_K(X, Y)v)_\alpha = F_\alpha(X, Y)v_\alpha.$$

Let's compute $F_\alpha \in \Omega^2(\mathcal{U}_\alpha, \mathbb{F}^{m\times m})$ in terms of $A_\alpha \in \Omega^1(\mathcal{U}_\alpha, \mathfrak{g})$. By Theorem 26.9, we can use the Riemann tensor as a substitute for Ω_K, so plugging in the definition of $R(X, Y)v$ with a section $v \in \Gamma(E)$ and using (28.5), we find

$$\begin{aligned}
(\Omega_K(X, Y)v)_\alpha &= \left(\nabla_X \nabla_Y v - \nabla_Y \nabla_X v - \nabla_{[X,Y]} v\right)_\alpha \\
&= (\mathcal{L}_X + A_\alpha(X))\,(\mathcal{L}_Y + A_\alpha(Y))\,v_\alpha - (\mathcal{L}_Y + A_\alpha(Y))\,(\mathcal{L}_X + A_\alpha(X))\,v_\alpha \\
&\quad - \left(\mathcal{L}_{[X,Y]} + A_\alpha([X, Y])\right) v_\alpha \\
&= \left(\mathcal{L}_X \mathcal{L}_Y - \mathcal{L}_Y \mathcal{L}_X - \mathcal{L}_{[X,Y]}\right) v_\alpha \\
&\quad + A_\alpha(X)\mathcal{L}_Y v_\alpha + A_\alpha(Y)\mathcal{L}_X v_\alpha - A_\alpha(Y)\mathcal{L}_X v_\alpha - A_\alpha(X)\mathcal{L}_Y v_\alpha \\
&\quad + \left(\mathcal{L}_X\left(A_\alpha(Y)\right) - \mathcal{L}_Y\left(A_\alpha(X)\right) - A_\alpha([X, Y])\right) v_\alpha \\
&\quad + \left(A_\alpha(X)A_\alpha(Y) - A_\alpha(Y)A_\alpha(X)\right) v_\alpha \\
&= \left(dA_\alpha(X, Y) + [A_\alpha(X), A_\alpha(Y)]\right) v_\alpha,
\end{aligned}$$

where in the last line, we've introduced the **matrix commutator**

$$[\mathbf{A}, \mathbf{B}] := \mathbf{AB} - \mathbf{BA} \qquad \text{for } \mathbf{A}, \mathbf{B} \in \mathbb{F}^{m\times m}.$$

The formula for F_α is thus

$$F_\alpha(X, Y) = dA_\alpha(X, Y) + [A_\alpha(X), A_\beta(Y)] \in \mathbb{F}^{m\times m}. \tag{28.6}$$

A basic result in the theory of Lie groups implies that $[A_\alpha(X), A_\beta(Y)]$ always lies in the Lie algebra $\mathfrak{g} \subset \mathbb{F}^{m\times m}$, hence $F_\alpha \in \Omega^2(\mathcal{U}_\alpha, \mathfrak{g})$, but this will be obvious in the case we're interested in below, so there is no need right now for a digression on Lie groups.

The local curvature 2-form depends on a choice of trivialization, so we need to pay attention to the way that it transforms when trivializations are changed. Suppose $\Phi_\beta : E|_{\mathcal{U}_\beta} \to \mathcal{U}_\beta \times \mathbb{F}^m$ is a second trivialization, related to Φ_α by the transition map $g = g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to G$. Then the local representations of a section $v \in \Gamma(E)$ are related on the overlap $\mathcal{U}_\alpha \cap \mathcal{U}_\beta$ by $v_\beta = g v_\alpha$, thus $F_\beta(X, Y)(g v_\alpha) = F_\beta(X, Y)v_\beta = (\Omega_K(X, Y)v)_\beta = g\,(\Omega_K(X, Y)v)_\alpha = g F_\alpha(X, Y)v_\alpha$, implying the relation

$$F_\beta(X, Y) = g F_\alpha(X, Y) g^{-1} \qquad \text{on } \mathcal{U}_\alpha \cap \mathcal{U}_\beta. \tag{28.7}$$

The formulas (28.6) and (28.7) have an especially interesting consequence whenever the structure group G happens to be abelian.

EXERCISE 28.7. Show that if the Lie subgroup $G \subset \mathrm{GL}(m, \mathbb{F})$ is abelian, then all matrices in G also commute with all matrices in the Lie algebra $\mathfrak{g}$, and $[\mathbf{A}, \mathbf{B}] = 0$ for all pairs $\mathbf{A}, \mathbf{B} \in \mathfrak{g}$.

In the abelian case, it now follows from (28.6) that F_α is the exterior derivative of A_α, and is thus a $\mathfrak{g}$-valued 2-form; as mentioned above, it is true in general that F_α takes values in $\mathfrak{g}$, but this is especially obvious in the abelian case. With that in mind, the values of F_α can now be seen to commute with transition functions, so (28.7) implies that $F_\alpha = F_\beta$ on the domain where they overlap, meaning there exists a *globally*-defined $\mathfrak{g}$-valued 2-form

$$F \in \Omega^2(M, \mathfrak{g})$$

that matches $F_\alpha \in \Omega^2(\mathcal{U}_\alpha, \mathfrak{g})$ for every G-compatible trivialization $(\mathcal{U}_\alpha, \Phi_\alpha)$. This 2-form is exact on $\mathcal{U}_\alpha$, and therefore closed, though it might not be globally exact since the connection 1-forms A_α are generally not globally defined.

Let's apply these observations in the special case where E is the tangent bundle of an oriented Riemannian 2-manifold (Σ, g) and ∇ is its Levi-Cività connection. The orientation and bundle metric give $T\Sigma$ the structure group $\mathrm{SO}(2)$, the group of 2-by-2 rotation matrices, which is indeed abelian. For computational purposes, it will be more convenient to replace $\mathrm{SO}(2)$ with the unitary group $\mathrm{U}(1)$, to which it is isomorphic via the transformation

$$\mathrm{SO}(2) \to \mathrm{U}(1) : \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \mapsto e^{i\theta}. \tag{28.8}$$

The Lie algebra $\mathfrak{u}(1)$ of $\mathrm{U}(1)$ is the space of purely imaginary 1-by-1 matrices, so

$$\Omega^k(\Sigma, \mathfrak{u}(1)) = \Omega^k(\Sigma, i\mathbb{R})$$

consists of imaginary-valued forms. Identifying $\mathrm{SO}(2)$ with $\mathrm{U}(1)$ in this way is equivalent to identifying $\mathbb{R}^2$ with $\mathbb{C}$ via the bijection $(x, y) \leftrightarrow x + iy$, and real local trivializations $\Phi_\alpha : T\Sigma|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{R}^2$ are thus identified with complex trivializations $T\Sigma|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{C}$, related to each other by transition functions with values in $\mathrm{U}(1) \subset \mathbb{C}$. In this way, $T\Sigma$ can now be viewed as a complex line bundle, and according to (28.8), scalar multiplication by i on the fibers of $T\Sigma$ is represented in any $\mathrm{SO}(2)$-compatible real trivialization by the rotation matrix $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, so in other words, it is a 90-degree counterclockwise rotation on every fiber. This is precisely the bundle map that we have previously referred to as

$$J : T\Sigma \to T\Sigma.$$

The formula relating K_G and the Riemann tensor can now be seen in a slightly new light: for any $\mathrm{U}(1)$-compatible local trivialization $\Phi_\alpha : T\Sigma|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{C}$, we have

$$F(X, Y) Z_\alpha = (R(X, Y) Z)_\alpha = -\left(K_G \, d\mathrm{vol}(X, Y) J Z\right)_\alpha = -K_G \, d\mathrm{vol}(X, Y) i Z_\alpha,$$

implying:

PROPOSITION 28.8. *Under the identification* $\mathrm{SO}(2) \cong \mathrm{U}(1)$ *defined in* (28.8)*, the imaginary-valued* 2*-form* $F \in \Omega^2(\Sigma, \mathfrak{u}(1))$ *is related to the Gaussian curvature* $K_G : \Sigma \to \mathbb{R}$ *and the Riemannian volume form* $d\mathrm{vol}$ *on* (Σ, g) *by*

$$iF = K_G \, d\mathrm{vol} \in \Omega^2(\Sigma).$$

This formula strongly suggests that it might be interesting to compute integrals $\int_P K_G \, d\mathrm{vol}$ over regions $P \subset \Sigma$, especially if P is contained in the domain $\mathcal{U}_\alpha$ of a local trivialization, on which $iF = i\, dA_\alpha$, so that Stokes' theorem implies

$$\int_P K_G \, d\mathrm{vol} = i \int_P dA_\alpha = i \int_{\partial P} A_\alpha.$$

We will apply this in the next lecture to integrate K_G over disk-like regions with piecewise-smooth polygonal boundaries, e.g. triangles bounded by geodesic segments. The imaginary-valued integral $\int_{\partial P} A_\alpha$ turns out in this case to give a new perspective on one of Euclid's best-known propositions: the sum of angles in a triangle is π. As we will see, the only reason this is true on the Euclidean plane is that for that particular Riemannian 2-manifold, $K_G \equiv 0$. You can see from Figure 8 that it is not true on the positively-curved unit sphere S^2, and Exercise 24.18 shows that it is also not true on the negatively-curved hyperbolic plane.

29. The Gauss-Bonnet formula

In the previous lecture we observed that on any oriented Riemannian 2-manifold (Σ, g), the 2-form $K_G \, d\mathrm{vol} \in \Omega^2(\Sigma)$ is locally (up to multiplication by i) the exterior derivative of a connection 1-form, so that $\int_P K_G \, d\mathrm{vol}$ over sufficiently simple regions $P \subset \Sigma$ should be computable via Stokes' theorem. We shall now follow this idea to its logical conclusion.

29.1. Polygons and their angles. We assume throughout this section that (Σ, g) is an oriented Riemannian 2-manifold, possibly with boundary, ∇ is its Levi-Cività connection, and its Riemannian volume form is denoted by

$$d\mathrm{vol}_\Sigma \in \Omega^2(\Sigma).$$

Our goal is to compute $\int_P K_G \, d\mathrm{vol}_\Sigma$ for compact regions $P \subset \Sigma$ that have the topology of disks bounded by piecewise smooth polygons. In general, a **piecewise smooth** curve in a smooth manifold M is a continuous map $\gamma : [a, b] \to M$ for which there are finitely many points $a = t_0 < t_1 < \ldots < t_{N-1} < t_N = b$ such that the restrictions

$$\gamma|_{[t_{j-1}, t_j]} : [t_{j-1}, t_j] \to M$$

are smooth immersions for each $j = 1, \ldots, N$. The velocity $\dot\gamma(t)$ of such a curve is thus a smooth function of t except possibly at the finitely many points t_j for $j = 1, \ldots, N-1$, where the two one-sided limits

$$\lim_{t \to t_j^\pm} \dot\gamma(t) \in T_{\gamma(t_j)} M$$

are both defined and nonzero but need not be equal, i.e. there may be jump discontinuities. The curve is called a piecewise smooth **simple closed** curve if $\gamma(b) = \gamma(a)$ and there is no other self-intersection $\gamma(t) = \gamma(t')$ for $t \ne t'$. We do not require $\dot\gamma(a) = \dot\gamma(b)$, so if we view γ as a piecewise-smooth map $S^1 \to M$ by identifying S^1 with the quotient $[a, b]/\sim$ in which $a \sim b$, the velocity of $\gamma : S^1 \to M$ may also have a jump discontinuity at the point $[a] = [b]$.

Definition 29.1. A **smooth polygon in** $\mathbb{R}^2$ is the closure $P \subset \mathbb{R}^2$ of a region bounded by the image of a single piecewise-smooth simple closed curve $\gamma : [a, b] \to \mathbb{R}^2$. If we write $a = t_0 < \ldots < t_N = b$ so that $t_1, \ldots, t_{N-1}$ are the finitely-many points where γ is allowed to be nonsmooth, then the smooth curves $\gamma([t_{j-1}, t_j])$ will be called **edges**, and their boundary points are called **vertices**. The union of all the edges will be denoted by ∂P.

Remark 29.2. The point $\gamma(a) = \gamma(b)$ is always considered a vertex of the polygon in Definition 29.1, so there is always at least one edge and one vertex. There is also ambituity in the notion of edges and vertices since the definition requires the set $\{t_1, \ldots, t_{N-1}\}$ to contain all points where γ is not smooth, but not the converse, so there is always some freedom to add more vertices arbitrarily, even if γ is completely smooth. This is just a matter of bookkeeping, as it will never at any stage be important to *require* that $\dot{\gamma}$ is discontinuous at some point.

Observe that if the region P in this definition has a smooth boundary, then $\partial P \cong S^1$ inherits from the orientation of $\mathbb{R}^2$ a natural orientation as the boundary of P. This notion of orientation generalizes naturally to the piecewise-smooth case so that each edge of ∂P inherits a natural orientation, and is thus a compact oriented 1-manifold with boundary.

There are theorems in topology that give fairly strong restrictions on what a compact region bounded by a continuous simple closed curve can look like. In order to avoid too much of a digression into topology, let us single out the particular property of the curve $\gamma : [a, b] \to \mathbb{R}^2$ that we will need to know. Assuming $a < t_1 < \ldots < t_{N-1} < b$ denote the points where γ is allowed to be nonsmooth, we can define a piecewise-continuous function

$$\phi : [a, b] \backslash \{t_1, \ldots, t_{N-1}\} \to \mathbb{R}$$

that is smooth on each of the subintervals (t_{j-1}, t_j) and gives the angle between $\dot{\gamma}(t) \in \mathbb{R}^2$ and the first standard basis vector. There is some freedom in this definition, as any multiple of 2π can be added to ϕ on each of the subintervals (t_{j-1}, t_j), but we can reduce this freedom by restricting the jumps at $t = t_1, \ldots, t_{N-1}$ to a suitable interval, namely

$$\Delta\phi_j := \lim_{t \to t_j^+} \phi(t) - \lim_{t \to t_j^-} \phi(t) \in [-\pi, \pi], \qquad j = 1, \ldots, N-1. \tag{29.1}$$

Here the convention is that $\Delta\phi_j > 0$ if the curve makes a sudden counterclockwise turn at t_j and $\Delta\phi_j < 0$ if it turns clockwise; these notions are well defined even in the case of a full 180-degree turn since γ is not allowed to intersect itself, and in this way we see the difference between $\Delta\phi_j = \pi$ and $\Delta\phi_j = -\pi$. With this restriction in place, the function ϕ is uniquely defined modulo a constant multiple of 2π. There is also a possible angle change at the end point $\gamma(a) = \gamma(b)$ that we will need to keep track of, so let us define this by

$$\Delta\phi_N := \phi(a) - \phi(b) + 2\pi k \in [-\pi, \pi],$$

where there is a unique choice of $k \in \mathbb{Z}$ that makes this number lie in the correct interval and satisfy the convention regarding counterclockwise/clockwise turns. The main observation we need to make now is that the total change in ϕ as t traverses the interval from a to b, including the jump discontinuities, must be exactly 2π:

$$\int_0^1 \dot{\phi}(t)\, dt + \sum_{j=1}^{N} \Delta\phi_j = 2\pi. \tag{29.2}$$

This statement is obvious whenever P is e.g. a disk with smooth boundary or a convex polygon, and it will in fact be obviously true for every example we are likely to consider, thus you might as well regard it as an extra condition in Definition 29.1. It is true but not so straightforward to prove that it actually follows from the conditions already stated in that definition—if you want to know why, see §29.3 at the end of this lecture.

Definition 29.3. A **smooth polygon in** Σ is a compact subset $P \subset \Sigma$ admitting an open neighborhood $\mathcal{U} \subset \Sigma$ with a chart $x : \mathcal{U} \to \mathbb{R}^2$ that identifies P with a smooth polygon P_0 in $\mathbb{R}^2$. The points and smooth curves identified by this chart with the vertices and edges of P_0 are called the **vertices** and **edges** of P.

The orientation of Σ restricts to any smooth polygon $P \subset \Sigma$ and induces a natural orientation on its edges, whose union we again denote by ∂P. The metric also restricts to each edge $\ell \subset \partial P$

and defines a natural "volume form"
$$d\mathrm{vol}_{\partial P} \in \Omega^1(\ell).$$
Although ∂P is not generally a smooth manifold, it's easy to see that Stokes' theorem still holds:
$$\int_P d\lambda = \int_{\partial P} \lambda$$
for any $\lambda \in \Omega^1(\Sigma)$, where the integral over ∂P is defined by summing the integrals over the edges. One can prove this by an approximation argument, perturbing ∂P to a smooth loop that bounds a region P_ϵ on which $\int_{P_\epsilon} d\lambda$ is almost the same. (A similar argument was sketched in Example 12.14 for applying Stokes' theorem on the product of two manifolds with boundary, which is technically a manifold with boundary and corners.)

We can apply Stokes' theorem in particular to compute $\int_P K_G \, d\mathrm{vol}_\Sigma$ for any smooth polygon $P \subset \Sigma$. For this purpose, recall that since the bundle $T\Sigma$ is equipped with both an orientation and a positive bundle metric, it has structure group $\mathrm{SO}(2)$, which we can identify with $\mathrm{U}(1)$ as in §28.2, thus making $T\Sigma$ into a complex line bundle on which scalar multiplication by i is the counterclockwise 90-degree rotation map $J : T\Sigma \to T\Sigma$. From this perspective, a $\mathrm{U}(1)$-compatible frame for $T\Sigma$ over a region $\mathcal{U} \subset \Sigma$ is simply a vector field $X \in \mathfrak{X}(\mathcal{U})$ that has unit length everywhere; indeed, one obtains a *real* orthonormal frame from this by putting X together with JX. It is now easy to see that $T\Sigma$ always admits such a frame on some neighborhood of a smooth polygon $P \subset \Sigma$: simply choose a chart $(\mathcal{U}, (x^1, x^2))$ with $P \subset \mathcal{U}$ as in Definition 29.3 and define the vector field
$$X := \frac{\partial_1}{|\partial_1|} \in \mathfrak{X}(\mathcal{U}).$$
Let us denote the corresponding local trivialization by $\Phi : T\Sigma|_{\mathcal{U}} \to \mathcal{U} \times \mathbb{C}$ and the associated connection 1-form for the Levi-Cività connection by
$$A := i\lambda \in \Omega^1(\mathcal{U}, \mathfrak{u}(1)),$$
thus defining a real-valued 1-form $\lambda \in \Omega^1(\mathcal{U})$. The discussion in §28.2 then implies $K_G \, d\mathrm{vol}_\Sigma = iF = i\, dA = -d\lambda$, hence by Stokes' theorem,
$$\int_\Sigma K_G \, d\mathrm{vol}_\Sigma = -\int_{\partial P} \lambda.$$
Our remaining task is to compute $\int_{\partial P} \lambda$.

Let us assume the boundary ∂P has $N \in \mathbb{N}$ edges, and thus N vertices at which it is not required to be smooth, and denote the angles formed between neighboring edges at these vertices by
$$\alpha_1, \dots, \alpha_N \in [0, 2\pi].$$
Note that the definition of these angles requires the orientation: the convention is that $\alpha_j \in [0, \pi)$ if there is a counterclockwise turn and $\alpha_j \in (\pi, 2\pi]$ for a clockwise turn. The case $\alpha_j = \pi$ is allowed, and in this way we can also accommodate situations where ∂P is completely smooth.

Next, choose a parametrization of ∂P as a piecewise-smooth simple closed curve $\gamma : [0, T] \to \Sigma$, oriented so that the parametrization of each edge is orientation preserving. The length $T > 0$ of the the interval can be choosen so that $|\dot\gamma(t)| = 1$ for all t, except at the finitely-many parameter values
$$0 < t_1 < \ldots < t_{N-1} < T$$
where $\dot\gamma(t)$ may fail to exist, and we will assume $\alpha_j \in [0, 2\pi]$ is the angle formed by a vertex at time t_j for $j = 1, \dots, N-1$, or times 0 and T for $j = N$. One can now find a piecewise-continuous function $\theta : [0, T] \to \mathbb{R}$ such that
$$\dot\gamma(t) = e^{i\theta(t)} X(\gamma(t)) \qquad \text{for all } t \in [0, 1] \backslash \{t_1, \dots, t_{N-1}\},$$

where θ is smooth on the open intervals (t_{j-1}, t_j) and is allowed to have jump discontinuities

$$\Delta\theta_j := \lim_{t \to t_j^+} \theta(t) - \lim_{t \to t_j^-} \theta(t) = \pi - \alpha_j \in [-\pi, \pi], \qquad j = 1, \ldots, N-1,$$

in which the orientation of Σ can be used to distinguish between $\Delta\theta_j = \pi$ and $\Delta\theta_j = -\pi$ via the same counterclockwise/clockwise convention that we used to define $\Delta\phi_j$. These conditions determine the function $\theta(t)$ uniquely modulo a constant multiple of 2π. We can also keep track of the angle α_N at $\gamma(a) = \gamma(b)$ by writing

$$\pi - \alpha_N = \Delta\theta_N := \theta(a) - \theta(b) + 2\pi k \in [-\pi, \pi],$$

for the unique choice of $k \in \mathbb{Z}$ that puts this number in the right interval and distinguishes correctly between counterclockwise and clockwise turns. With these definitions in place, the jumps $\Delta\theta_j \in [-\pi, \pi]$ are related to the angles $\alpha_j \in [0, 2\pi]$ by

$$\alpha_j = \pi - \Delta\theta_j, \qquad j = 1, \ldots, N. \tag{29.3}$$

Lemma 29.4. $\displaystyle\int_0^T \dot{\theta}(t)\, dt + \sum_{j=1}^N \Delta\theta_j = 2\pi.$

Proof. It is clear from the definitions that this number is at least an integer multiple of 2π. Let $\gamma_0 := \varphi^{-1} \circ \gamma : [0, T] \to \mathbb{R}^2$, so γ_0 is a piecewise-smooth simple closed curve parametrizing ∂P_0, whose image under the embedding $\varphi : \mathcal{U}_0 \to \mathcal{U} \subset \Sigma$ is ∂P. If we equip $\mathcal{U}_0 \subset \mathbb{R}^2$ with the pullback metric $\varphi^* g$, then the way in which our frame $X \in \mathfrak{X}(\mathcal{U})$ was defined gives a new interpretation of $\theta(t)$: it is the angle of the tangent vector $\dot{\gamma}_0(t) \in \mathbb{R}^2$ relative to the standard basis vector $\mathbf{e}_1$, as measured using the metric $\varphi^* g$. If $\varphi^* g$ were the *standard* Euclidean metric on $\mathbb{R}^2$, the lemma would now just be a restatement of Equation 29.2. Unfortunately, we cannot assume $\varphi^* g$ is the standard Euclidean metric; this would be a very strong restriction, forcing (Σ, g) to be locally flat on the region $\mathcal{U}$. However, the space of all Riemannian metrics is convex, so we can define a smooth family of metrics on $\mathcal{U}_0 \subset \mathbb{R}^2$ by

$$g_s := s\varphi^* g + (1-s) g_E, \qquad s \in [0, 1],$$

where $g_E := dx^2 + dy^2$ denotes the Euclidean metric, so g_s interpolates between $g_1 = \varphi^* g$ and $g_0 = g_E$. For each $s \in [0, 1]$, we can now define a corresponding function $\theta^s(t)$ in the same manner as above, but using the metric $\varphi_* g_s$ on $\mathcal{U} \subset \Sigma$ to measure angles. The sum $\int_0^T \dot{\theta}^s(t)\, dt + \sum_{j=1}^N \Delta\theta_j^s$ depends continuously on the parameter s, and since it is always a multiple of 2π, we get the same answer for $s = 1$ and $s = 0$, so that the result in the case of the Euclidean metric is also valid in the general case. $\square$

Now let's compute $\int_{\ell_j} \lambda$ for a specific edge $\ell_j := \gamma([t_{j-1}, t_j]) \subset \partial P$. This requires computing $\lambda(\dot{\gamma}(t)) = -iA(\dot{\gamma}(t))$, which can be deduced by computing a covariant derivative in the direction of $\dot{\gamma}(t)$. In particular, $\dot{\gamma}(t)$ itself is expressed relative to our chosen frame X as the complex-valued function $e^{i\theta(t)}$, thus

$$\begin{aligned} \nabla_t \dot{\gamma}(t) &= \left(\partial_t e^{i\theta(t)} + A(\dot{\gamma}(t)) e^{i\theta(t)}\right) X(\gamma(t)) = \left(\dot{\theta}(t) + \lambda(\dot{\gamma}(t))\right) i e^{i\theta(t)} X(\gamma(t)) \\ &= \left(\dot{\theta}(t) + \lambda(\dot{\gamma}(t))\right) i\dot{\gamma}(t). \end{aligned} \tag{29.4}$$

This last expression has a useful geometric interpretation.

Definition 29.5. Suppose ℓ is a 1-dimensional submanifold of a Riemannian 2-manifold (Σ, g) and $\nu \in \Gamma(T\Sigma|_\ell)$ is unit normal vector field along ℓ. The (signed) **geodesic curvature** of ℓ is then defined as the unique function

$$\kappa_\ell : \ell \to \mathbb{R}$$

such that for any local parametrization $\gamma : (a, b) \to \ell$ of ℓ satisfying $|\dot\gamma| \equiv 1$,

$$\nabla_t \dot\gamma(t) = \kappa_\ell(\gamma(t))\nu(\gamma(t))$$

for all t.

This definition makes sense because if γ is parametrized with unit speed, differentiating the relation $\langle \dot\gamma(t), \dot\gamma(t)\rangle = 1$ reveals that $\nabla_t \dot\gamma(t)$ is always orthogonal to $\dot\gamma(t)$, and is therefore a real multiple of $\nu(\gamma(t))$. Moreover, one could change the local parametrization γ of ℓ, but all other parametrizations with unit speed take the form $t \mapsto \gamma(\pm t + c)$ for a constant c, so one obtains the same definition of κ_ℓ. It does depend on the choice of normal vector field: reversing ν changes κ_ℓ by a sign. It follows that κ_ℓ cannot be defined in this way if ℓ has non-orientable normal bundle, but this situation does not arise in the application that we have in mind. In the non-orientable case, one can still define an *unsigned* geodesic curvature $|\kappa_\ell| \geqslant 0$, which is actually just the norm of $\nabla_t \dot\gamma$, and the latter is given as a definition of the term "geodesic curvature" in many books. In either case, it should be emphasized that geodesic curvature is a purely *extrinsic* notion, as it depends on the embedding of the submanifold ℓ into the surface Σ. (Indeed, Exercise 27.4 shows that there is no interesting notion of intrinsic curvature for Riemannian 1-manifolds, as they are all locally flat.) The geodesic curvature is a measurement of the extent to which $\ell \subset \Sigma$ fails locally to the image of a geodesic in (Σ, g); in particular, $\kappa_\ell \equiv 0$ if and only if ℓ can be parametrized locally by geodesics.

With this definition understood, (29.4) can be reinterpreted using the observation that $i\dot\gamma(t)$ is a 90-degree counterclockwise rotation of $\dot\gamma(t)$, pointing inwards through the boundary of P. If we take this as a choice of normal vector field along ℓ_j, the relation now says:

Lemma 29.6. *For $t \in [t_{j-1}, t_j]$, $\dot\theta(t) + \lambda(\dot\gamma(t)) = \kappa_{\ell_j}(\gamma(t))$.* □

We now have enough ingredients in place to write down a revealing formula for $\int_{\partial P} \lambda$: combining Lemmas 29.6 and 29.4 with (29.3), we have:

$$\begin{aligned}\int_{\partial P} \lambda &= \sum_{j=1}^N \int_{\ell_j} \lambda = \sum_{j=1}^N \int_{t_{j-1}}^{t_j} \lambda(\dot\gamma(t))\, dt = \sum_{j=1}^N \int_{t_{j-1}}^{t_j} \kappa_{\ell_j}(\gamma(t))\, dt - \int_0^T \dot\theta(t)\, dt \\ &= \sum_{j=1}^N \int_{t_{j-1}}^{t_j} \kappa_{\ell_j}(\gamma(t))\, dt - \left(2\pi - \sum_{j=1}^N \Delta\theta_j\right) = \sum_{j=1}^N \int_{t_{j-1}}^{t_j} \kappa_{\ell_j}(\gamma(t))\, dt - 2\pi + \sum_{j=1}^N (\pi - \alpha_j) \\ &= \sum_{j=1}^N \int_{t_{j-1}}^{t_j} \kappa_{\ell_j}(\gamma(t))\, dt + (N-2)\pi - \sum_{j=1}^N \alpha_j.\end{aligned}$$

Since γ was parametrized to have unit speed on each edge, the integrals in this last expression are actually just the integrals of the 1-forms $\kappa_{\ell_j}\, d\mathrm{vol}_{\partial P}$ over the respective edges, and using Stokes' theorem to rewrite $\int_{\partial P} \lambda$ in terms of the Gaussian curvature, we obtain from this the first version of the Gauss-Bonnet formula:

Theorem 29.7 (Gauss-Bonnet formula, polygon version). *Suppose (Σ, g) is an oriented Riemannian 2-manifold with Gaussian curvature $K_G : \Sigma \to \mathbb{R}$, $P \subset \Sigma$ is a smooth polygon with N smooth edges $\ell_1, \dots, \ell_N \subset \partial P$ and angles $\alpha_1, \dots, \alpha_N \in [0, 2\pi]$ at its vertices, and the signed geodesic curvature κ_{ℓ_j} of each edge ℓ_j is defined with respect to a normal vector field pointing inwards. Then*

$$\sum_{j=1}^N \alpha_j = (N-2)\pi + \int_P K_G\, d\mathrm{vol}_\Sigma + \sum_{j=1}^N \kappa_{\ell_j}\, d\mathrm{vol}_{\partial P}.$$

□

We have arranged this formula to look like a generalization of the fact that triangles in the Euclidean plane have angles adding up to π; that is just the case where $N = 3$, $K_G \equiv 0$ and all edges are geodesic segments. More generally, the integral of the Gaussian curvature can now be viewed as a correction term that measures the failure of this relation to hold:

COROLLARY 29.8. *If $P \subset \Sigma$ is a smooth polygon with N edges that are all geodesic segments, then the angles $\alpha_1, \ldots, \alpha_N$ at the vertices satisfy*

$$\sum_{j=1}^{N} \alpha_j = (N-2)\pi + \int_P K_G \, d\mathrm{vol}_\Sigma.$$

$\square$

REMARK 29.9. The assumption that Σ carries an orientation is not actually necessary for Theorem 29.7, because even if Σ is not globally orientable, a neighborhood of the polygon $P \subset \Sigma$ is diffeomorphic to an open subset of $\mathbb{R}^2$, so an orientation can always be chosen on this neighborhood. If one reverses the orientation, none of the terms in the Gauss-Bonnet formula actually change: for the angles α_j and the term $(N-2)\pi$ this is obvious, though it takes a bit more thought for the two integrals. We already saw in the previous lecture that K_G does not change if the orientation is switched; the geodesic curvatures also do not change since they depend on the choice of normal vector field at the boundary and this was defined independently of all orientations. Changing the orientations of Σ and ℓ_j changes the volume forms $d\mathrm{vol}_\Sigma$ and $d\mathrm{vol}_{\ell_j}$ by a sign, but a cancelling sign is caused by the fact that $\int_{-M} \omega = -\int_M \omega$ for any oriented manifold M and any top-dimensional form ω.

EXAMPLE 29.10. Now is a good moment to look again at Figure 8 in Lecture 19, which shows a geodesic triangle in the unit sphere S^2 whose angles are all $\pi/2$. This triangle occupies exactly $1/8$ of the total area of S^2, so its area is $\pi/2$, and this is also $\int_\Sigma K_G \, d\mathrm{vol}_\Sigma$ since $K_G \equiv 1$ by Example 27.13. The formula in Corollary 29.8 thus becomes $3\pi/2 = \pi + \frac{\pi}{2}$ in this case.

EXERCISE 29.11. According to Exercise 27.22, the Poincaré half-plane $(\mathbb{H}, h)$ has constant curvature $K_G \equiv -1$.

(a) Write down the Riemannian volume form on $(\mathbb{H}, h)$, and show that any region of the form $[a, b] \times [c, \infty) \subset \mathbb{H}$ for $-\infty < a < b < \infty$ and $c > 0$ has finite area, while regions of the form $[a, b] \times (0, c] \subset \mathbb{H}$ have infinite area.
(b) Show that every compact region in $(\mathbb{H}, h)$ bounded by three geodesics has area strictly less than π, though its area can be arbitrarily close to π.
Hint: Use the result of Exercise 24.18.

29.2. Triangulation and the Euler characteristic. Next question: what happens if we integrate K_G over a region on which $T\Sigma$ is not trivializable? A nice way to approach this is by decomposing Σ into a union of polygons glued together along their edges.

DEFINITION 29.12. Let Σ be a 2-dimensional manifold, possibly with boundary. A **polygonal triangulation** of Σ is a collection of smooth polygons $\{P_\alpha \subset \Sigma\}_{\alpha \in I}$ with $\Sigma = \bigcup_{\alpha \in I} P_\alpha$, called the **faces** of the triangulation, while each edge or vertex of each of these polygons is called an **edge** or **vertex** of the triangulation respectively. They are required to satisfy the following conditions:

(1) Each edge ℓ is either contained in $\partial\Sigma$ or satisfies $\ell \cap \partial\Sigma \subset \partial\ell$, and in the latter case, it is an edge of exactly two faces.
(2) Two distinct faces are either disjoint or their intersection is a union of common edges.
(3) Every vertex is a vertex of at most finitely many faces.

The sets of vertices and edges of the triangulation are sometimes denoted by $\Sigma^0, \Sigma^1 \subset \Sigma$ and also called the 0-**skeleton** and 1-**skeleton** respectively. Note that if $\partial\Sigma \neq \varnothing$, then $\partial\Sigma \subset \Sigma^1$. We say the triangulation is **finite** if it has only finitely many faces (and therefore also finite-many vertices and edges).

Polygonal triangulations are somewhere in between two similar notions that are popular in topology: they are more general than what are normally just called *triangulations* (in which all the faces are required to be actual triangles), while also being special cases of the more general notion of *CW-complexes*. It is a general fact that all smooth surfaces admit polygonal triangulations, and one can even arrange without loss of generality for them to be triangulations in the stricter sense, in which every face has three edges. A similar result (based on *simplices*, a higher-dimensional generalization of triangles) also holds for smooth manifolds of all dimensions, though not generally for *topological* manifolds above dimension 3. In practice, we will not need to have such general existence results, because for our purposes it is more interesting to look at specific examples in which explicit triangulations are not hard to construct. But just out of interest, here is the most relevant special case of the general result:

PROPOSITION 29.13. *Every compact smooth surface Σ admits a finite polygonal triangulation consisting only of triangles.*

SKETCH OF THE PROOF. Every point in Σ admits a compact neighborhood that is a smooth polygon contained in the domain of a chart, and since Σ is compact, it can be covered with finitely many such polygons. After small perturbations, we can also arrange without loss of generality that no edge of any of these polygons intersects a vertex of another one, and that whenever two edges intersect, they do so transversely (and therefore only finitely-many times). Define Σ^0 to be the union of the set of vertices of all these polygons with the finite set of intersections between their edges; we should correspondingly redefine the word "edge" to mean any potentially shorter segment of one of the original edges that is bounded by two points of Σ^0. Each connected component in the complement of the set of edges is now an open region with compact closure contained in the domain of a chart, and bounded by some disjoint union of piecewise-smooth simple closed curves. It therefore remains only to show the following: any region $P \subset \mathbb{R}^2$ bounded by piecewise-smooth simple closed curves can be decomposed into a union of smooth polygons that each have three edges and intersect each other only along matching pairs of edges. This can be achieved by adding new edges, i.e. choosing new smooth paths through the interior of P that connect previously unconnected pairs of vertices. Once you've added enough of these, every component of the complement is bounded by a triangle. $\square$

REMARK 29.14. There's a subtlety in the construction of triangulations that should be mentioned. Most authors' definition of the term "smooth triangle" is stricter than ours: we are assuming a smooth polygon in $\mathbb{R}^2$ can be any compact region bounded by a piecewise-smooth simple closed curve, and we call it a triangle if that curve has three smooth edges, but in practice, such an object does not need to look very similar to what we typically imagine as a triangle. (Try drawing an example where the edges form gratuitously complicated spirals and vertices have angles $2\pi - \epsilon$.) Most authors add the condition that a "triangle" must actually be homeomorphic to a perfectly ordinary convex triangle with straight edges in $\mathbb{R}^2$. We are not assuming this, but it follows from our definition for somewhat nontrivial reasons, and you'll find this fact lurking in the background of Equation (29.2) if you look into the details as discussed in §29.3. Our argument in that appendix appeals to the classification of closed surfaces in order to show that every smooth polygon by our definition really is homeomorphic to a disk. One needs to be a bit careful about circularity here, because most popular proofs of the classification of surfaces are based on the fact that all surfaces can be triangulated. (There are ways to get around this, however, e.g. the proof via Morse theory in [**Hir94**] is quite illuminating and does not require triangulations.)

In practice, all useful constructions of triangulations on surfaces require some nontrivial topological input at some step to ensure that compact regions bounded by simple closed curves in $\mathbb{R}^2$ are always homeomorphic to disks. If the boundary curve is continuous but not smooth, then this fact requires a difficult classical result known as the *Schoenflies theorem* (see [**Moi77**]). In the smooth category there is a standard way to avoid this by using geodesics: the idea is to choose a Riemannian metric on Σ and carry out the proof of Proposition 29.13 so that every edge in the triangulation becomes the unique shortest geodesic between two nearby points. The final subdivision step is less obvious in this setting, but with a bit more care it can still be done, and in this way one obtains a triangulation whose edges are all geodesics. It is much easier than the Schoenflies theorem to see (e.g. by working in Riemann normal coordinates based at a vertex) that any region bounded by three short geodesics is homeomorphic to a disk.

DEFINITION 29.15. Given a finite polygonal triangulation of Σ with v vertices, e edges and f faces, the *Euler characteristic* of Σ is the integer

$$\chi(\Sigma) = v - e + f.$$

The Euler characteristic turns out to be a topological invariant of Σ, though our definition makes this far from obvious—*a priori* it appears to depend rather crucially on a choice of triangulation. It will follow from Theorem 29.17 below that this is not the case, that in fact $\chi(\Sigma)$ depends at most on the differentiable structure of Σ. Proving that it only depends on the topology of Σ requires methods from algebraic topology: the standard approach is to define $\chi(\Sigma)$ in terms of singular homology and use either cellular or simplicial homology to prove that the quantity above matches this definition for any triangulation. Details may be found in e.g. [**Hat02**, **Bre93**, **Wen18**].

EXERCISE 29.16. Taking it on faith for the moment that the Euler characteristic doesn't depend on a choice of triangulation, show that $\chi(S^2) = 2$, $\chi(\mathbb{D}^2) = 1$ and $\chi(\mathbb{T}^2) = 0$.

We shall now compute the integral of K_G over a compact surface using a finite polygonal triangulation with v vertices, e edges and f faces. Assume $e = e_0 + e_\partial$ where e_∂ is the number of edges contained in $\partial\Sigma$, and similarly $v = v_0 + v_\partial$. Observe that every vertex on $\partial\Sigma$ is a boundary point of exactly two edges on $\partial\Sigma$, and since every edge likewise has two boundary points, $e_\partial = v_\partial$.[74] By Theorem 29.7, $\int_\Sigma K_G \, d\mathrm{vol}_\Sigma$ contains a term of the form

$$-\sum_j \int_{\ell_j} \kappa_{\ell_j} \, d\mathrm{vol}_{\ell_j} + \sum_j \alpha_j - (N-2)\pi$$

for each face, assuming the face in question has N edges. Adding these up for all faces, we make the following observations:

(1) Every edge $\ell \subset \Sigma \backslash \partial\Sigma$ is an edge for two distinct faces and thus appears twice with two oppositely-oriented choices of normal vector field pointing toward different faces. The geodesic curvature terms for these edges cancel in the sum.
(2) The geodesic curvature terms for all edges $\ell \subset \partial\Sigma$ add up to
$$-\int_{\partial\Sigma} \kappa_{\partial\Sigma} \, d\mathrm{vol}_{\partial\Sigma}$$
if we define $\kappa_{\partial\Sigma}$ with respect to a normal vector field pointing inward at the boundary.

[74] There is a small loop-hole in this argument: our definition of smooth polygons allows the possibility that there is only one edge, whose two boundary points then coincide to form a single vertex, but if this happens, the claim that $e_\partial = v_\partial$ remains valid.

(3) The sum of all angles α_j at an interior vertex (for every face adjacent to that vertex) is 2π, and for boundary vertices the sum is π. Thus altogether these terms contribute $2\pi v_0 + \pi v_\partial = 2\pi v - \pi v_\partial$.

(4) Every interior edge is counted twice and boundary edges are counted once, so the $-(N-2)\pi$ terms add up to $-\pi(2e_0 + e_\partial - 2f) = 2\pi(f-e) + \pi e_\partial$.

Summing all these contributions, we have

$$-\int_{\partial\Sigma} \kappa_{\partial\Sigma}\, ds + 2\pi v + 2\pi(f-e) - \pi v_\partial + \pi e_\partial = -\int_{\partial\Sigma} \kappa_{\partial\Sigma}\, ds + 2\pi\chi(\Sigma).$$

This proves:

THEOREM 29.17 (Gauss-Bonnet formula, global version). *Assume (Σ, g) is a compact oriented 2-dimensional Riemannian manifold, possibly with boundary, and the signed geodesic curvature $\kappa_{\partial\Sigma} : \partial\Sigma \to \mathbb{R}$ is defined with respect to a normal vector field pointing inward at the boundary. Then*

$$\int_\Sigma K_G \, \mathrm{dvol}_\Sigma + \int_{\partial\Sigma} \kappa_{\partial\Sigma}\, \mathrm{dvol}_{\partial\Sigma} = 2\pi\chi(\Sigma).$$

REMARK 29.18. In keeping with Remark 29.9, Theorem 29.17 remains true if (Σ, g) is not oriented or orientable, though in this case the two integrals on the left hand side require some additional effort to interpret. The global volume form $d\mathrm{vol}_\Sigma \in \Omega^2(\Sigma)$ does not exist if Σ is not orientable, but recall from §11.4 that every Riemannian manifold, regardless of orientability, admits a canonical *volume element*, which is a density rather than a differential form. We can interpret both of the integrals in Theorem 29.17 as integrals of smooth real-valued functions with respect to measures defined via the canonical volume elements determined by the metric on Σ and $\partial\Sigma$. In practice, the volume element on (Σ, g) matches $|d\mathrm{vol}_\Sigma|$ on any region where an orientation can be chosen, so for instance $\int_\Sigma K_G \, d\mathrm{vol}_\Sigma$ can be computed as the sum of the terms $\int_{P_\alpha} K_G \, d\mathrm{vol}_\Sigma$ over all the faces P_α of a polygonal triangulation, where $d\mathrm{vol}_\Sigma$ and the integral are defined in each case by choosing an arbitrary orientation of $T\Sigma|_{P_\alpha}$, and Remark 29.9 shows that the result does not depend on this choice. Once this is understood, the proof of Theorem 29.17 also works in the non-oriented case.

Several wonderful things follow immediately from the global Gauss-Bonnet formula. Observe that the left hand side has nothing to do with the triangulation, while the right hand side makes no reference to the metric or curvature.

COROLLARY 29.19. *The Euler characteristic $\chi(\Sigma)$ does not depend on the choice of triangulation, and for any two diffeomorphic surfaces Σ_1 and Σ_2, $\chi(\Sigma_1) = \chi(\Sigma_2)$.* □

COROLLARY 29.20. *For a fixed compact surface Σ, the sum $\int_\Sigma K_G \, \mathrm{dvol}_\Sigma + \int_{\partial\Sigma} \kappa_{\partial\Sigma}\, \mathrm{dvol}_{\partial\Sigma}$ is an integer multiple of 2π, and is the same for any choice of Riemannian metric.* □

In particular, the latter statement imposes serious topological restrictions on the kinds of metrics that are allowed on any given surface: e.g. it is impossible to find a metric with everywhere positive Gaussian curvature on a surface with negative Euler characteristic. To get a handle on this, it helps to have some concrete examples in mind; these are provided by the following exercises.

EXERCISE 29.21. Suppose Σ is a compact oriented surface with boundary and $\ell_1, \ell_2 \subset \partial\Sigma$ are two distinct connected components of $\partial\Sigma$. We can *glue* these two components to produce a new surface Σ' as follows: since ℓ_1 and ℓ_2 are both circles, there is an orientation reversing diffeomorphism $\varphi : \ell_1 \to \ell_2$, which we use to define

$$\Sigma' = \Sigma / \sim$$

where the equivalence identifies $p \in \ell_1$ with $\varphi(p) \in \ell_2$, thus identifying ℓ_1 and ℓ_2 to a single circle, now in the interior of Σ'. Show that $\chi(\Sigma') = \chi(\Sigma)$. Note: Σ need not be a connected surface to start with, so this trick can be used to glue together two separate surfaces along components of their boundaries.
Hint: A given triangulation of Σ may have different numbers of vertices on ℓ_1 and ℓ_2, but one can always modify the triangulation by adding more vertices and edges so that these numbers become the same. The number of edges on each boundary component will also always match the number of vertices. (Why?)

EXERCISE 29.22. Let Σ be the closed unit disk in $\mathbb{R}^2$ with two smaller disjoint open disks removed: the resulting surface is called a *pair of pants*. Show that $\chi(\Sigma) = -1$.

Similarly, a *handle* is a surface Σ diffeomorphic to the torus $\mathbb{T}^2$ with one open disk removed. Show that $\chi(\Sigma) = -1$.

EXERCISE 29.23. Suppose Σ is a compact surface with boundary. The operation of *gluing a handle* to Σ is defined as follows: choose a smoothly embedded closed disk in the interior of Σ, remove its interior, and glue the resulting surface along its new boundary component to a handle (see Exercise 29.22). Show that this operation decreases the Euler characteristic of Σ by 2.

EXERCISE 29.24. A **closed oriented surface of genus** g is any compact surface Σ without boundary that is diffeomorphic to a surface obtained from S^2 by gluing g handles. Special cases include the sphere itself ($g = 0$) and the torus ($g = 1$). Show that

$$\chi(\Sigma) = 2 - 2g.$$

For Σ a compact surface with boundary, we say it has *genus* g if it is diffeomorphic to a closed surface of genus g with finitely many small open disks cut out. Show that if such a surface has m boundary components, then $\chi(\Sigma) = 2 - 2g - m$.

REMARK 29.25. In case you didn't already believe this, we now have a simple proof of the fact that two closed oriented surfaces with differing genera (that is the plural of "genus") are not diffeomorphic: if they were, then their Euler characteristics would have to match. The converse is also true, but harder to prove; it follows from the topological classification of surfaces (see e.g. [**Wen18**, Lecture 19] or [**Hir94**]).

The Gauss-Bonnet theorem enables us to make some sweeping statements regarding what kinds of metrics may exist on various compact surfaces. In general, we say that a surface Σ with a Riemannian metric has *positive* (or *zero* or *negative*) curvature if its Gaussian curvature is positive (or zero or negative) at every point.

THEOREM 29.26. *Let Σ be a closed oriented surface of genus g. Then Σ admits a Riemannian metric with positive curvature if and only if $\Sigma \cong S^2$, zero curvature if and only if $\Sigma \cong \mathbb{T}^2$, and negative curvature if and only if $g \geqslant 2$.*

PROOF. We shall not provide the entire proof, but by this point the result should at any rate seem believable, and in one direction the claim is clear: the stated conditions on the genus are necessary due to the Gauss-Bonnet theorem and the formula $\chi(\Sigma) = 2 - 2g$. It's easy to see that the sphere admits a metric with positive curvature: this is true for the induced metric coming from the standard embedding of S^2 in $\mathbb{R}^3$. Things are similarly simple for the torus, though the usual embedding of $\mathbb{T}^2$ into $\mathbb{R}^3$ (as a doughnut) is the wrong picture to look at. Instead take $\mathbb{R}^2$ with its standard flat metric and define $\mathbb{T}^2$ as $\mathbb{R}^2/\mathbb{Z}^2$: the translation invariance of the Euclidean metric implies that it gives a well defined metric on the quotient, and it is indeed locally flat.

The only part that is less obvious is that every surface of genus $g \geqslant 2$ admits a metric of negative curvature—in fact, by a famous result in the theory of surfaces, one can always find a metric that has *constant* curvature -1. One approach is to take the Poincaré half-plane $(\mathbb{H}, h)$ as a model (see Exercise 27.22) and show that every such surface can be constructed by drawing a smooth polygon in $(\mathbb{H}, h)$ and identifying certain edges appropriately. We refer to [**Spi99b**, Chapter 6, Addendum 1] for details. One can also prove this using geometric PDE methods, see for instance [**Tro92**]. $\square$

REMARK 29.27. For a surface Σ of genus $g \geqslant 2$, the standard way of embedding Σ into $\mathbb{R}^3$ as a surface with g handles is misleading in some respects: as a hypersurface in $\mathbb{R}^3$, its Gaussian curvature is sometimes positive and sometimes negative. Exercise 29.28 below shows that this will always be true, for *any* embedding of Σ in Euclidean $\mathbb{R}^3$, though the Gauss-Bonnet theorem guarantees at least that the part with negative curvature is the majority. Unfortunately (from the perspective of people who like to visualize things), there is no isometric embedding of any closed surface with everywhere negative curvature into $\mathbb{R}^3$. (This is a less deep observation than Hilbert's theorem about embeddings of the hyperbolic plane, mentioned in Remark 27.23. The exercise below does not say anything about the hyperbolic plane since it is not compact.)

EXERCISE 29.28. Prove: A closed surface Σ in Euclidean $\mathbb{R}^3$ cannot have $K_G \leqslant 0$ everywhere. *Hint: For some $R > 0$, Σ must lie inside the closed ball of radius R and touch its boundary tangentially at some point.*

29.3. Addendum: Polygons are disks. You should perhaps not bother to read this section unless you felt uncomfortable calling Equation (29.2) "obvious". Here is one way I can think of to prove it, using only the assumption that $\gamma : [a, b] \to \mathbb{R}^2$ is a piecewise-smooth simple closed curve bounding a compact region P. There may also be easier ways that I haven't thought of, but the basic idea of what I have in mind is to deform γ via a so-called **regular homotopy** to a smooth loop bounding a standard disk, for which (29.2) really is obvious. Let us call $\gamma : [a, b] \to \mathbb{R}^2$ a *smoothly immersed loop* if it is smooth and satisfies $\gamma(a) = \gamma(b)$, $\dot{\gamma}(a) = \dot{\gamma}(b)$ and $\dot{\gamma}(t) \neq 0$ for all t. One can associate to any smoothly immersed loop a smooth function $\phi : [a, b] \to \mathbb{R}$, unique modulo 2π, that measures the angle of $\dot{\gamma}(t) \in \mathbb{R}^2$ relative to a standard basis vector, and $\int_a^b \dot{\phi}(t)\, dt = \phi(b) - \phi(a)$ is then $2\pi k$ for some $k \in \mathbb{Z}$, called the **twisting number** of γ. A **regular homotopy** of loops is a smooth family of smoothly immersed loops $\{\gamma_s : [a, b] \to \mathbb{R}^2\}_{s \in [0,1]}$. Given such a family, the corresponding angle functions $\phi_s(t)$ can also be chosen to depend smoothly on both s and t, so that $\int_a^b \dot{\phi}_s(t)\, dt$ depends continuously on s, and therefore so does the twisting number. Since the latter is always in integer, this implies that it is the same for γ_0 and γ_1, i.e. the twisting number is invariant under regular homotopy. Our goal in the following is thus to show that, after smoothing the angles in order to make ∂P a smooth loop, it admits a regular homotopy to the boundary of a round disk, whose twisting number is clearly 1.

Step 1: Since $\gamma : [a, b] \to \mathbb{R}^2$ has only finitely-many nonsmooth points, each one is isolated, and it is therefore easy to modify γ by a C^0-small perturbation in small neighborhoods of these points to make it a smooth embedding with $\gamma(a) = \gamma(b)$ and $\dot{\gamma}(a) = \dot{\gamma}(b)$. This is an example of what topologists call "smoothing the corners", and the contribution to $\int_a^b \dot{\phi}(t)\, dt$ from the small neighborhoods of t_j where this modification is done then corresponds to $\Delta\phi_j$, so the left hand side of (29.2) now contains only the integral term, and computes 2π times the twisting number of γ. (Note: It is really important in this step to make sure that you're using the right convention about the distinction between $\Delta\phi_j$ being $+\pi$ or $-\pi$, i.e. counterclockwise vs. clockwise rotations!)

Step 2: The compact region $P \subset \mathbb{R}^2$ bounded by γ is now a compact oriented smooth 2-manifold with connected boundary, and we claim that it is diffeomorphic to a disk $\mathbb{D}^2$. Indeed, the classification of surfaces (see e.g. [**Wen18**, Lecture 19] or [**Hir94**]) implies that P must be diffeomorphic to the complement of an open disk in a closed orientable surface Σ_g of some genus $g \geqslant 0$, so our claim is equivalent to the assertion that $g = 0$. To see this, one can add a "point at infinity" to $\mathbb{R}^2$, making it diffeomorphic to the sphere S^2, so that the unbounded region of $\mathbb{R}^2$ lying outside of γ becomes another compact oriented smooth 2-manifold with connected boundary, embedded in S^2. Applying the classification of surfaces again, this region is diffeomorphic to the complement of an open disk in a closed orientable surface Σ_h of some genus $h \geqslant 0$. Gluing the two pieces together presents S^2 as the connected sum of Σ_g and Σ_h, which is Σ_{g+h}. But S^2 is not diffeomorphic to Σ_{g+h} unless $g + h = 0$, thus $g = h = 0$.

Step 3: We now know that P is diffeomorphic to $\mathbb{D}^2$, and by the tubular neighborhood theorem, ∂P has a neighborhood in $\mathbb{R}^2$ diffeomorphic to $(-1, 1) \times S^1$, where ∂P itself is identified with $\{0\} \times S^1$. This is enough information to construct an open neighborhood $\mathcal{U} \subset \mathbb{R}^2$ of P with a diffeomorphism $\psi : \mathcal{U} \to B^2_r$ onto the open ball $B^2_r \subset \mathbb{R}^2$ of some radius $r > 1$ such that $\psi(P)$ is the closed unit disk $\mathbb{D}^2$. Let us equip $\mathcal{U}$ with the Riemannian metric $g := \psi^*(dx^2 + dy^2)$. The point of this definition is that we can easily understand the geodesics for this metric: they are the images under ψ^{-1} of straight lines in B^2_r. As a consequence, we can now write

$$\gamma(t) = \exp^g_p(X(t)),$$

where $p := \psi^{-1}(0)$, $[a, b] \to T_p\mathcal{U} = \mathbb{R}^2 : t \mapsto X(t)$ is a parametrization of the unit circle in $T_p\mathcal{U}$ with respect to the metric g, and the superscript g is included to emphasize that we are using this metric (rather than the Euclidean metric) to define the exponential map.

Step 4: Since $(\mathcal{U}, g)$ is isometric via ψ to a standard ball with the Euclidean metric, $\exp^g_p$ defines a diffeomorphism from a ball in $T_p\mathcal{U}$ containing the loop $X(t)$ onto $\mathcal{U}$. The family

$$\gamma_s(t) := \exp^g_p(sX(t)), \qquad s \in [\epsilon, 1]$$

therefore defines a regular homotopy between γ and some loop $\gamma_\epsilon : [a, b] \to \mathcal{U}$ that can be assumed to lie in an arbitrarily small neighborhood of p by choosing $\epsilon > 0$ small.

Step 5: Consider the smooth family of Riemannian metrics $g_s := sg + (1 - s)g_E$ on $\mathcal{U}$ for $s \in [0, 1]$, where $g_E := dx^2 + dy^2$ is the Euclidean metric. For a sufficiently small neighborhood $\mathcal{O} \subset T_p\mathcal{U}$ of 0, we can assume that the corresponding exponential maps $\exp^{g_s}_p$ are embeddings of $\mathcal{O}$ onto open neighborhoods of p in $\mathcal{U}$. Now define

$$X_s(t) \in T_p\mathcal{U} = \mathbb{R}^2$$

for each $s \in [0, 1]$ and $t \in [a, b]$ as the unique positive rescaling of $X(t) \in \mathbb{R}^2$ that makes it a unit vector with respect to the metric g_s, and define another family of smooth loops by

$$\beta_s(t) := \exp^{g_s}_p(\epsilon X_s(t)) \qquad s \in [0, 1].$$

Taking $\epsilon > 0$ small enough so that $\epsilon X_s(t) \in \mathcal{O}$ for all s, t, these loops are all embeddings, and thus define a regular homotopy between $\beta_1 = \gamma_\epsilon$ and β_0, where the latter is a parametrization of the ϵ-disk about p with respect to the Euclidean metric. We conclude that γ_ϵ and therefore also γ have the same twisting number as β_0, which is 1.

30. The first Chern class

We are not yet done extracting mileage out of the formula

$$F = dA_\alpha.$$

Recall from §28.2: this relates a local connection 1-form $A_\alpha \in \Omega^1(\mathcal{U}_\alpha, \mathfrak{g})$ to a globally-defined Lie algebra-valued curvature 2-form $F \in \Omega^2(M, \mathfrak{g})$ on any vector bundle $E \to M$ with abelian structure group G carrying a compatible connection. The Gauss-Bonnet formula arose from the special case where E is the tangent bundle of an oriented Riemannian 2-manifold, so that the group G was $\mathrm{SO}(2) \cong \mathrm{U}(1)$, but this is not the only type of vector bundle with structure group $\mathrm{U}(1)$ one might want to consider. We will explore what else can be done with this in §30.1 and §30.2, giving a rudimentary introduction to the much larger subject of characteristic classes and Chern-Weil theory. We then apply this again to the case $E = T\Sigma$ in §30.3, and deduce yet another useful interpretation of the Euler characteristic, including some discussion of spheres with hair.

30.1. An invariant of complex line bundles. The basic object of study in this section is a smooth complex line bundle

$$\pi : E \to M$$

over a manifold M of some dimension $n \in \mathbb{N}$. We are going to construct an algebraic invariant that can detect whether two such bundles are isomorphic. There are many reasons one might want to do this. One easy one to name is that vector bundles arise naturally in the tubular neighborhood theorem (cf. Exercise 23.4), where they serve as local models for neighborhoods of submanifolds, so if one can classify the isomorphism classes of vector bundles of a given rank over a given manifold, one obtains a picture of all possible neighborhoods of embeddings of that manifold into a larger one up to diffeomorphism. Since $\mathrm{U}(1) \cong \mathrm{SO}(2)$, classifying complex line bundles is equivalent to classifying oriented real bundles of rank 2, which arise whenever one studies the embeddings of an oriented manifold into another oriented manifold two dimensions larger. The simplest case of the latter situation is knot theory, which studies embeddings of S^1 into 3-manifolds, and this is only one of many situations in topology and related areas where certain types of vector bundles need to be classified.

The construction of our invariant will depend on two choices of auxiliary data:

(1) A bundle metric $\langle\ ,\ \rangle$, thus making $E \to M$ a *Hermitian* line bundle and reducing its structure group from $\mathrm{GL}(1, \mathbb{C})$ to $\mathrm{U}(1)$;
(2) A metric connection ∇, represented in any $\mathrm{U}(1)$-compatible local trivialization $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{C}$ by an imaginary-valued connection 1-form $A_\alpha \in \Omega^1(\mathcal{U}_\alpha, \mathfrak{u}(1))$.

These choices are crucial for the definition of the invariant, but we will see that the invariant itself does not depend on them.

A clue about the right thing to do arises out of the observation in §28.2 that in this situation, there is a globally-defined imaginary-valued 2-form $F \in \Omega^2(M, \mathfrak{u}(1))$ that matches dA_α for every choice of local trivialization. In particular, it is obvious that F is closed, but it might not be exact since each of the individual 1-forms A_α is defined only on the domain $\mathcal{U}_\alpha$ and not necessarily on all of M. As we saw in Lecture 13, the distinction between closed and exact forms on a manifold is measured by its de Rham cohomology, so one wonders whether the cohomology class represented by F might carry interesting information. A further hint in this direction comes from the following:

LEMMA 30.1. *If $\widehat{\nabla}$ is another metric connection on $E \to M$ with curvature 2-form $\widehat{F} \in \Omega^2(M, \mathfrak{u}(1))$, then $\widehat{F} = F + i\, d\lambda$ for some $\lambda \in \Omega^1(M)$.*

PROOF. The difference between two connections is always a bundle map, i.e. there exists a smooth bilinear bundle map $B : TM \oplus E \to E$ such that $\widehat{\nabla}_X v = \nabla_X v + B(X, v)$ for all $X \in \mathfrak{X}(M)$ and $v \in \Gamma(E)$, and B can also be interpreted as a bundle-valued 1-form

$$\beta \in \Omega^1(M, \mathrm{End}(E)), \qquad \beta(X)v := B(X, v).$$

Since the fibers E_p for all $p \in M$ are 1-dimensional, all endomorphisms $E_p \to E_p$ come from scalar multiplication, giving a natural isomorphism $\mathbb{C} \to \mathrm{End}(E_p)$ so that β can be replaced with a *complex*-valued 1-form $\beta \in \Omega^1(M, \mathbb{C})$ such that $\widehat{\nabla}_X v = \nabla_X v + \beta(X)v$. Writing down this relation in the local trivialization $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{C}$ then gives the relation

$$\widehat{A}_\alpha(X) = A_\alpha(X) + \beta(X),$$

where $\widehat{A}_\alpha \in \Omega^1(\mathcal{U}_\alpha, \mathfrak{u}(1))$ is the local connection 1-form for $\widehat{\nabla}$. Since A_α and $\widehat{A}_\alpha$ are both purely imaginary-valued, the same is therefore true for β, giving $\beta = i\lambda$ for some real-valued 1-form

$\lambda \in \Omega^1(M)$. Taking the exterior derivative of $\widehat{A}_\alpha = A_\alpha + i\lambda$ then gives the stated relation between F and $\widehat{F}$. $\square$

Strictly speaking, we cannot talk about the de Rham cohomology class represented by $F \in \Omega^2(M, \mathfrak{u}(1))$ without slightly altering our previous definition of de Rham cohomology, because F is not a real-valued form. But that is easily fixed, and Lemma 30.1 then tells us that the following definition is independent of the choice of metric connection:

DEFINITION 30.2. The **first Chern class** of the complex line bundle $\pi : E \to M$ with bundle metric $\langle\ ,\ \rangle$ is the de Rham cohomology class

$$c_1(E) := \left[-\frac{1}{2\pi i} F \right] \in H^2_{\mathrm{dR}}(M),$$

where $F \in \Omega^2(M, \mathfrak{u}(1))$ is the curvature 2-form associated to any choice of metric connection on $E \to M$.

The reason for the factor of 2π and the minus sign in this definition will become clear when we discuss computations in §30.2.

THEOREM 30.3. *The first Chern class of complex line bundles has the following properties.*

(1) $c_1(E)$ is independent of the choice of bundle metric $\langle\ ,\ \rangle$ on $E \to M$.
(2) For the trivial line bundle $E^0 := M \times \mathbb{C} \to \mathbb{C}$, $c_1(E^0) = 0$.
(3) If $E, E' \to M$ are two complex line bundles admitting a bundle isomorphism $E \to E'$, then $c_1(E) = c_1(E')$.
(4) For any complex line bundle $E \to M$ and any smooth map $f : N \to M$, the pullback bundle $f^*E \to N$ has $c_1(f^*E) = f^*c_1(E) \in H^2_{\mathrm{dR}}(N)$.

PROOF. The easiest property to prove is (2), so we start with that: on the trivial bundle we can choose ∇ to be the trivial connection, and there is an obvious global trivialization in which the resulting connection 1-form vanishes identically, implying the same for the curvature 2-form and thus $[-F/2\pi i] = 0$. Alternatively, one can reach the same conclusion without assuming anything about the connection: it suffices to observe that since a trivialization can be defined on $\mathcal{U}_\alpha := M$, there is a globally-defined connection 1-form $A_\alpha \in \Omega^1(M, \mathfrak{u}(1))$, whose exterior derivative is F, hence $-F/2\pi i$ is exact.

Moving on, we will prove a slightly stricter version of property (3) that depends on a bundle metric, and then use this to prove (1). For two line bundles $E \to M$ and $E' \to N$ equipped with bundle metrics, a smooth linear bundle map $\Psi : E \to E'$ covering a smooth map $\psi : M \to N$ will be called a **bundle isometry** if for every $p \in M$, Ψ defines an isomorphism $E_p \to E'_{\psi(p)}$ that is unitary, meaning it preserves the inner products. If $E, E' \to M$ admit a bundle isomorphism $\Psi : E \to E'$, then for any choice of bundle metric on E, there is a unique one on E' that makes Ψ a bundle isometry. With this data in place, any metric connection ∇ on E can be "pushed forward" via Ψ to define a metric connection ∇' on E', namely by

$$\nabla'_X v := \Psi\left(\nabla_X(\Psi^{-1} v)\right).$$

It is an easy exercise to check that ∇'_X satisfies the Leibniz rule required to be a connection on E' and is also compatible with the bundle metric. We can also use Ψ to push forward local trivializations: given a trivialization $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{C}$, we can define a trivialization $\Phi'_\alpha : E'|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{C}$ by

$$\Phi'_\alpha := \Phi_\alpha \circ \Psi^{-1}.$$

For this choice, the section $v \in \Gamma(E)$ has the same local representation $v_\alpha : \mathcal{U}_\alpha \to \mathbb{C}$ as the section $\Psi v \in \Gamma(E')$, and the local connection 1-forms $A_\alpha, A'_\alpha \in \Omega^1(\mathcal{U}_\alpha, \mathfrak{u}(1))$ from our two connections via these two trivializations are exactly the same. It follows that the curvature 2-forms for ∇ and ∇' are identical over $\mathcal{U}_\alpha$, and since the same trick can be done on any region where E is trivializable, they are therefore identical everywhere, proving $c_1(E) = c_1(E')$.

We can now prove (1) as follows. The space of bundle metrics on $E \to M$ is convex, so any two bundle metrics $\langle \ , \ \rangle_0$ and $\langle \ , \ \rangle_1$ can be connected by a smooth family of bundle metrics

$$\langle \ , \ \rangle_s := s\langle \ , \ \rangle_1 + (1-s)\langle \ , \ \rangle_0, \qquad s \in [0,1].$$

Let $\widehat{E} \to [0,1] \times M$ denote the pullback of $E \to M$ via the obvious projection map $[0,1] \times M \to M$, and endow $\widehat{E}$ with a bundle metric such that the inner product at the point $(s,p) \in [0,1] \times M$ is $\langle \ , \ \rangle_s$ at the point p. Choosing a metric connection on $\widehat{E}$, parallel transport along the paths $s \mapsto (s,p)$ for each $p \in M$ now defines a bundle isometry $(E, \langle \ , \ \rangle_0) \to (E, \langle \ , \ \rangle_1)$, and the result of the previous paragraph thus implies that the two definitions of $c_1(E)$ via these two bundle metrics match.

Finally, we prove (4): assuming $E \to M$ is a line bundle with bundle metric $\langle \ , \ \rangle$ and $f : N \to M$ is a smooth map, equip $f^*E \to N$ with the unique bundle metric so that the canonical bundle map $f^*E \to E$ covering f is a bundle isometry. For any metric connection ∇ on E, the pullback connection on f^*E is then also compatible with the bundle metric, and the discussion following Equation (21.3) shows that for any local trivialization $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{C}$ with connection 1-form $A_\alpha \in \Omega^1(\mathcal{U}_\alpha, \mathfrak{u}(1))$, there is a pullback trivialization of f^*E over $f^{-1}(\mathcal{U}_\alpha) \subset N$ in which the connection 1-form for the pullback connection is f^*A_α. Taking the exterior derivative, it follows that the pullback $f^*F \in \Omega^2(N, \mathfrak{u}(1))$ of the curvature 2-form $F \in \Omega^2(M, \mathfrak{u}(1))$ for ∇ is the curvature 2-form for the pullback connection, thus $c_1(f^*E) = [-f^*F/2\pi i] = f^*[-F/2\pi i] = f^*c_1(E)$. $\square$

30.2. Computing the first Chern number. For $c_1(E) \in H^2_{\mathrm{dR}}(M)$ to be a truly useful invariant, we need a practical means of computing it. As a rule, the best way to understand a 2-dimensional cohomology class $[\omega] \in H^2_{\mathrm{dR}}(M)$ is by integrating it over closed oriented 2-dimensional submanifolds $\Sigma \subset M$: the result is independent of the 2-form $\omega \in \Omega^2(M)$ representing $[\omega]$ since, by Stokes' theorem, integrals of exact forms over closed manifolds always vanish. Integrating ω over $\Sigma \subset M$ is the same as computing $\int_\Sigma j^*\omega$ for the natural inclusion map $j : \Sigma \hookrightarrow M$. More generally, one can also consider integrals of the form $\int_\Sigma f^*\omega$ for arbitrary closed oriented surfaces Σ and smooth maps $f : \Sigma \to M$, which need not be embeddings; in this situation, if ω represents $c_1(E) \in H^2_{\mathrm{dR}}(M)$, then $f^*\omega$ represents $c_1(f^*E) \in H^2_{\mathrm{dR}}(\Sigma)$ according to Theorem 30.3. It can be deduced from de Rham's theorem and a result of Thom[75] that these integrals for all possible choices of surfaces Σ and maps $f : \Sigma \to M$ completely characterize $[\omega] \in H^2_{\mathrm{dR}}(M)$. I will not prove that here, but am mentioning it only as support for the following assertion: if you want to compute $c_1(E)$ in general, then it suffices in principle if you know how to compute the first Chern classes of bundles over *closed oriented surfaces*, as this is what the pullback bundles $f^*E \to \Sigma$ are. Moreover, the essential information about $c_1(E)$ for a bundle $E \to \Sigma$ is contained in the integral $\int_\Sigma \omega \in \mathbb{R}$ for any choice of 2-form ω representing $c_1(E) \in H^2_{\mathrm{dR}}(\Sigma)$. This number deserves a name.

Definition 30.4. For a complex line bundle E over a closed oriented surface Σ, the number

$$\int_\Sigma c_1(E) := \int_\Sigma \omega,$$

defined by choosing any representative $\omega \in \Omega^2(M)$ of the cohomology class $c_1(E) \in H^2_{\mathrm{dR}}(M)$, is called the **first Chern number** of $E \to \Sigma$.

[75]The result in question comes from the famous paper [Tho54], and states that every singular homology class $A \in H_k(M;\mathbb{Z})$ can be written as $A = \frac{1}{q} f_*[\Sigma]$ for some closed oriented k-manifold Σ, smooth map $f : \Sigma \to M$ and $q \in \mathbb{N}$.

According to the definition of $c_1(E)$, one can compute the first Chern number in principle by choosing a bundle metric and metric connection, which give rise to an imaginary-valued curvature 2-form $F \in \Omega^2(M, \mathfrak{u}(1))$, and then integrating

$$\int_\Sigma c_1(E) = -\frac{1}{2\pi i}\int_\Sigma F.$$

We already know how to do this in the case $E = T\Sigma$ with the Levi-Cività connection: the answer comes from the Gauss-Bonnet formula, which we'll come back to in §30.3 below. But without making any further assumptions about the line bundle $E \to \Sigma$ or the connection, it is possible to compute this integral in another way that has interesting applications. Let us make the following assumption, which we will see below is completely realistic: suppose Σ can be decomposed into two compact (but not necessarily connected) surfaces $\Sigma_\alpha, \Sigma_\beta \subset \Sigma$ with a common boundary consisting of a finite set of disjoint circles $C_1, \ldots, C_N \subset \Sigma$,

$$\partial\Sigma_\beta = \partial\Sigma_\alpha = \coprod_{j=1}^N C_j,$$

such that both subsets are contained in open neighborhoods $\mathcal{U}_\alpha, \mathcal{U}_\beta \subset \Sigma$ on which there exist U(1)-compatible trivializations Φ_α and Φ_β respectively. Denote the transition function relating these trivializations by

$$g := g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{U}(1),$$

and observe that it is defined in particular on each of the circles C_j. Both Σ_α and Σ_β inherit orientations from Σ such that the boundary orientations of $\partial\Sigma_\alpha$ and $\partial\Sigma_\beta$ are opposite; let us orient the individual circles C_j to match the boundary orientation of $\partial\Sigma_\beta$. Stokes' theorem then gives

$$(30.1)\quad \int_\Sigma F = \int_{\Sigma_\alpha} dA_\alpha + \int_{\Sigma_\beta} dA_\beta = \int_{\partial\Sigma_\alpha} A_\alpha + \int_{\partial\Sigma_\beta} A_\beta = \int_{\partial\Sigma_\beta}(A_\beta - A_\alpha) = \sum_{j=1}^N \int_{C_j}(A_\beta - A_\alpha).$$

A formula relating A_α and A_β to each other on $\mathcal{U}_\alpha \cap \mathcal{U}_\beta$ was worked out in Exercise 20.9, namely

$$A_\alpha(X) = g(p)^{-1}A_\beta(X)g(p) + g(p)^{-1}\,dg(X), \qquad \text{for } p \in \mathcal{U}_\alpha \cap \mathcal{U}_\beta,\ X \in T_pM,$$

and in the present case, the fact that U(1) is abelian simplifies it to

$$A_\alpha = A_\beta + g^{-1}\,dg \qquad \text{on } \mathcal{U}_\alpha \cap \mathcal{U}_\beta,$$

so that (30.1) becomes

$$\int_\Sigma F = -\sum_{j=1}^N \int_{C_j} g^{-1}\,dg.$$

This formula is further confirmation that $c_1(E)$ is an essentially topological quantity with no dependence on the choice of connection. Now observe that since g takes values in U(1), we can write it as $g = e^{i\theta}$ for a uniquely defined smooth function

$$\theta : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathbb{R}/2\pi\mathbb{Z}.$$

It should be emphasized that θ cannot necessarily be defined as a *real*-valued function on $\mathcal{U}_\alpha \cap \mathcal{U}_\beta$, at least not if we want it to be continuous, though a real-valued version could indeed be defined on a sufficiently small neighborhood of any given point in $\mathcal{U}_\alpha \cap \mathcal{U}_\beta$. Such a local function would be unique only up to the addition of constant multiples of 2π, but this means that its *differential* is uniquely defined as a perfectly ordinary closed (but not necessarily exact) real-valued 1-form

$$d\theta \in \Omega^1(\mathcal{U}_\alpha \cap \mathcal{U}_\beta).$$

We can therefore write $g^{-1}\,dg = e^{-i\theta}d(e^{i\theta}) = ie^{-i\theta}e^{i\theta}\,d\theta = i\,d\theta$, giving

$$\int_\Sigma c_1(E) = -\frac{1}{2\pi i}\int_\Sigma F = \frac{1}{2\pi}\sum_{j=1}^N \int_{C_j} d\theta. \tag{30.2}$$

This last integral looks at first like it should vanish, but remember that $d\theta$ is not necessarily an exact 1-form since θ is not a real-valued function. We encountered something similar in Example 13.12, and the following definition provides a useful topological interpretation for integrals of this type.

Definition 30.5. Suppose S is an oriented manifold diffeomorphic to S^1, and $f : S \to \mathbb{C}\backslash\{0\}$ is a smooth function. The **winding number**

$$\operatorname{wind}_S(f) \in \mathbb{Z}$$

of f is then the unique integer with the following property: for any smooth orientation-preserving map $\gamma : [0,1] \to S$ that satisfies $\gamma(0) = \gamma(1)$ and is an embedding on $(0,1)$, and any smooth functions $\rho, \phi : [0,1] \to \mathbb{R}$ such that $\rho > 0$ and $f(\gamma(t)) = \rho(t)e^{i\phi(t)}$ for all t,

$$\operatorname{wind}_S(f) = \frac{1}{2\pi}\left[\phi(1) - \phi(0)\right].$$

To see that $\operatorname{wind}_S(f)$ in Definition 30.5 is independent of the various choices involved, one can reinterpret it as the integral

$$\operatorname{wind}_S(f) = \frac{1}{2\pi}\int_S d\theta, \tag{30.3}$$

where the 1-form $d\theta \in \Omega^1(S)$ is defined from the unique smooth function $\theta : S \to \mathbb{R}/2\pi\mathbb{Z}$ satisfying $f(p) = r(p)e^{i\theta(p)}$ for some positive function $r : S \to \mathbb{R}$. Indeed, suppose $\gamma : [0,1] \to S$ is an orientation-preserving parametrization of S as in the definition: then we can write $f(\gamma(t)) = r(\gamma(t))e^{i\phi(t)}$ for some smooth function $\phi : [0,1] \to \mathbb{R}$ such that $\theta(\gamma(t))$ is the image of $\phi(t)$ under the quotient projection $\mathbb{R} \to \mathbb{R}/2\pi\mathbb{Z}$. Differentiating this relation between θ and ϕ gives $d\phi = d\theta \circ T\gamma = \gamma^*d\theta$. Now for any small $\epsilon > 0$, γ is an orientation-preserving diffeomorphism of $[\epsilon, 1-\epsilon]$ onto its image $S_\epsilon \subset S$, so the change-of-variables formula gives $\int_{S_\epsilon} d\theta = \int_{[\epsilon,1-\epsilon]} \gamma^*d\theta$. After taking $\epsilon \to 0$, this gives

$$\int_S d\theta = \int_{[0,1]} \gamma^*d\theta = \int_{[0,1]} d\phi = \phi(1) - \phi(0),$$

thus proving (30.3). (Caution: If we had manipulated the symbols in this last equation without thinking about their meaning, we might have said $\int_{[0,1]} \gamma^*d\theta = \int_{[0,1]} d(\gamma^*\theta) = \int_{[0,1]} d(\theta\circ\gamma) = \theta(\gamma(1)) - \theta(\gamma(0)) = 0$. This is where it is crucial to remember that θ is not a *real*-valued function on S, but instead takes values in the manifold $\mathbb{R}/2\pi\mathbb{Z}$. Thus $d(\gamma^*\theta)$ cannot be interpreted as an exact 1-form, and we cannot use Stokes' theorem to compute it.)

An easy corollary of (30.3) is that $\operatorname{wind}_S(f)$ is not only independent of the choice of parametrization $\gamma : [0,1] \to S$, but it is also *homotopy invariant* : if $f_0, f_1 : S \to \mathbb{C}\backslash\{0\}$ are two ends of a smooth family of nowhere-zero functions $\{f_s : S \to \mathbb{C}\backslash\{0\}\}_{s\in[0,1]}$, then $\operatorname{wind}_S(f_0) = \operatorname{wind}_S(f_1)$. One can see this from the fact that the integral $\int_S d\theta$ in this situation will depend continuously on the parameter s, and since it is also an integer multiple of 2π, this implies that it cannot change at all.

We can now rewrite (30.2) in terms of winding numbers:

PROPOSITION 30.6. *For the complex line bundle $E \to \Sigma = \Sigma_\alpha \cup \Sigma_\beta$ as described above with transition function $g = g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{U}(1)$,*

$$\int_\Sigma c_1(E) = \sum_{j=1}^N \mathrm{wind}_{C_j}(g).$$

□

It should now be obvious why the factor of $1/2\pi$ was included in the definition of $c_1(E)$: it makes the value of $\int_\Sigma c_1(E)$ an integer, a fact which was far from obvious in its definition.

Proposition 30.6 is of practical use for computing first Chern numbers. It can also be used to approach the following question, which you might not expect should have a well-defined answer at all:

QUESTION 30.7. *How many zeroes should a section $s \in \Gamma(E)$ be expected to have?*

If you think of sections of a complex line bundle over a surface as something analogous to complex-valued functions on an open domain in $\mathbb{R}^2$, then this does not at first seem like a sensible question, because the number of zeroes will generally depend on the choice of function, and one could always just choose a nonzero constant function for which the answer zero. But on a line bundle over a closed surface, nowhere-zero sections might not exist—indeed, a nowhere-zero section in this situation is equivalent to a frame, so such a thing exists if and only if the bundle is globally trivial. This observation hints that the issue in Question 30.7 is fundamentally topological, at least if we have the correct interpretation of the words "how many". Let us restrict our attention to smooth sections $s : \Sigma \to E$ such that the zero set

$$s^{-1}(0) := \left\{ p \in \Sigma \;\middle|\; s(p) = 0 \in E_p \right\} \subset \Sigma$$

is finite. One can show that all sections in an open and dense subset of $\Gamma(E)$ have this property (see Remark 30.12 below). One could now count the number of elements in $s^{-1}(0)$, but this notion of counting is too naive to give an answer independent of the choice of section. The right interpretation of the words "how many" turns out to be one that attaches to each individual zero an integer-valued weight, and this weight can be defined as a winding number:

DEFINITION 30.8. Suppose $p \in \Sigma$ is an isolated point in the zero set $s^{-1}(0)$ of a section $s \in \Gamma(E)$. The **index** of s at p (also sometimes called the **order** of the zero p) is defined as the integer

$$\mathrm{ind}(s; p) := \mathrm{wind}_{\partial\mathcal{D}}(s_\alpha),$$

where $\mathcal{D} \subset \Sigma$ is a small disk containing p in its interior such that $s^{-1}(0) \cap \mathcal{D} = \{p\}$, and $s_\alpha : \mathcal{D} \to \mathbb{C}$ is the local representative of s in some trivialization Φ_α of E defined on a neighborhood of $\mathcal{D}$.

REMARK 30.9. The winding number in Definition 30.8 requires $\partial\mathcal{D}$ to be oriented, so we assign it the boundary orientation, where $\mathcal{D}$ inherits an orientation from Σ. Reversing the orientation of Σ thus changes the sign of $\mathrm{ind}(s; p)$, and the index can only be defined up to a sign if Σ is not orientable.

EXERCISE 30.10. Use the homotopy-invariance of winding numbers to show that the index $\mathrm{ind}(s; p)$ in Definition 30.8 does not depend on the choices of disk $\mathcal{D} \subset \Sigma$ surrounding p and local trivialization Φ_α over $\mathcal{D}$.
Hint: The crucial detail is that s does not vanish on $\mathcal{D}$ except at the point p.

EXERCISE 30.11. Recall from Exercise 19.7 that at any point $p \in s^{-1}(0)$ in the zero-set of a section $s \in \Gamma(E)$, there is a well-defined *linearization* $Ds(p) : T_p\Sigma \to E_p$. For the following statement, we can regard E_p as an oriented 2-dimensional *real* vector space by defining any basis of the form (v, iv) for $v \neq 0 \in E_p$ to be positively oriented. Convince yourself that this orientation is well defined, and then prove the following: if $Ds(p)$ is invertible, then $\mathrm{ind}(s;p) = \pm 1$, positive if $Ds(p)$ is orientation preserving and negative if $Ds(p)$ is orientation reversing.

REMARK 30.12. We do not have time for a proper treatment of transversality theory in this course, but if you know the basic definitions, you may be able to convince yourself without much difficulty that the linearization $Ds(p) : T_p\Sigma \to E_p$ at a zero $p \in s^{-1}(0)$ is invertible if and only if the intersection at $0 \in E_p \subset E$ between the zero-section $Z := \bigcup_{p\in\Sigma} 0 \subset E$ and the submanifold $s(\Sigma) \subset E$ is transverse. General results in transversality theory (see e.g. [**Hir94**]) then imply that all zeroes satisfy this condition for all sections in some open and dense subset of $\Gamma(E)$. This is why we know there always exist sections whose zero-sets are finite.

EXERCISE 30.13. Suppose $s \in \Gamma(E)$ is an isolated zero $p \in s^{-1}(0)$ with $\mathrm{ind}(s;p) \neq 0$. Show that for any neighborhood $\mathcal{U} \subset \Sigma$ of p, any sufficiently C^0-small perturbation of s must also vanish somewhere in $\mathcal{U}$. In other words, zeroes with nonvanishing index cannot be perturbed away.
Hint: Consider only perturbations of s such that the winding number along some fixed circle around p does not change.

EXERCISE 30.14. On the trivial complex line bundle $E = \mathbb{R}^2 \times \mathbb{C} \to \mathbb{R}^2$, find an example of a section $s \in \Gamma(E)$ with an isolated zero at one point $p \in \mathbb{R}^2$ with $\mathrm{ind}(s;p) = 0$, such that s admits small perturbations with no zeroes at all.

DEFINITION 30.15. Suppose $s \in \Gamma(E)$ has a finite zero set. The **algebraic count of zeroes** of s is the integer

$$\#s^{-1}(0) := \sum_{p\in s^{-1}(0)} \mathrm{ind}(s;p) \in \mathbb{Z}.$$

THEOREM 30.16. *Suppose Σ is a closed oriented surface and $E \to \Sigma$ is a complex line bundle. Then for any section $s \in \Gamma(E)$ with at most finitely-many zeroes,*

$$\#s^{-1}(0) = \int_\Sigma c_1(E).$$

PROOF. Assuming $s^{-1}(0) \subset \Sigma$ is finite, choose for each $p \in s^{-1}(0)$ a small closed disk $\mathcal{D}_p \subset \Sigma$ whose boundary encircles p, and assume all of these disks are small enough so that they do not intersect each other and they are contained in a neighborhood on which E is trivializable. Set

$$\Sigma_\beta := \bigcup_{p\in s^{-1}(0)} \mathcal{D}_p, \qquad \Sigma_\alpha := \overline{\Sigma \setminus \Sigma_\beta},$$

and let $v \in \Gamma(E|_{\mathcal{U}_\beta})$ denote an arbitrary choice of section over some open neighborhood $\mathcal{U}_\beta \subset \Sigma$ of Σ_β such that $|v| \equiv 1$, hence v can be interpreted as a $\mathrm{U}(1)$-compatible frame over $\mathcal{U}_\beta$ and gives rise to a corresponding trivialization Φ_β. On $\mathcal{U}_\alpha := \Sigma \setminus s^{-1}(0)$, s itself determines a $\mathrm{U}(1)$-compatible trivialization, defined by interpreting the normalized section $s/|s|$ as a $\mathrm{U}(1)$-compatible frame, and we can denote the corresponding trivialization by Φ_α. This means that the local representation of s with respect to Φ_α is a positive real-valued function $s_\alpha > 0$; its representation with respect to Φ_β is related to this by $s_\beta = g s_\alpha$ for the transition map $g := g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{U}(1)$, and is therefore just a positive rescaling of g. This proves that for each zero $p \in s^{-1}(0)$,

$$\mathrm{wind}_{\partial\mathcal{D}_p}(s_\beta) = \mathrm{wind}_{\partial\mathcal{D}_p}(g),$$

so the equality of $\#s^{-1}(0)$ and $\int_\Sigma c_1(E)$ now follows from Proposition 30.6. □

EXERCISE 30.17. By counting zeroes of sections, show that for any pair of complex line bundles $E, E' \to \Sigma$ over a closed oriented surface Σ, $\int_\Sigma c_1(E \otimes E') = \int_\Sigma c_1(E) + \int_\Sigma c_1(E')$.

EXERCISE 30.18. For any vector bundle $E \to M$ over $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, with dual bundle $E^* \to M$, there is a canonical bilinear bundle map from $E \otimes E^*$ to the trivial line bundle $M \times \mathbb{F} \to M$, defined at each point $P \in M$ by

$$E_p \otimes E_p^* \to \mathbb{F} : v \otimes \lambda \mapsto \lambda(v).$$

(a) Show that if $\operatorname{rank}(E) = 1$, the bundle $E \otimes E^*$ is trivial.
(b) Show that for any complex line bundle E over a compact oriented surface Σ, $\int_\Sigma c_1(E^*) = -\int_\Sigma c_1(E)$.

Exercise 30.18 reveals an interesting difference between real and complex vector bundles. For any real bundle $E \to M$, choosing a bundle metric $\langle\ ,\ \rangle$ gives rise to a bundle isomorphism

$$E \to E^* : v \mapsto \langle v,\ \rangle.$$

This trick does not work in the complex case because bundle metrics are complex linear only in one argument and complex *antilinear* in the other, so the map $E \to E^*$ above can be defined, but it is complex antilinear on each fiber and thus not a complex bundle isomorphism. Exercise 30.18 shows that this is not just a defect in our method of finding isomorphisms: the bundles E and E^* really are not generally isomorphic in the complex case, as their first Chern classes will differ whenever they are nonzero. There do exist complex line bundles with $c_1(E) \neq 0$: we will see some explicit examples in the next section, and more generally, it is not hard to construct examples over surfaces that have arbitrary interger values for $\int_\Sigma c_1(E)$. The trick is to glue simpler pieces together in clever ways, e.g. if you present the sphere S^2 with its north and south poles $p_\pm \in S^2$ as the union of the two open subsets $\mathcal{U}_\pm := S^2 \backslash \{p_\pm\}$, then you can take two trivial line bundles $E_\pm := \mathcal{U}_\pm \times \mathbb{C} \to \mathcal{U}_\pm$, and glue these together to produce a bundle $E \to S^2$ with local trivializations over $\mathcal{U}_+$ and $\mathcal{U}_-$ having any desired transition function $g : \mathcal{U}_+ \cap \mathcal{U}_- \to \mathrm{U}(1)$.

30.3. The Poincaré-Hopf theorem on surfaces. If $E \to \Sigma$ is the tangent bundle $T\Sigma$ of a closed oriented surface with a Riemannian metric, one can choose ∇ to be the Levi-Cività connection, and by Proposition 28.8 and the Gauss-Bonnet formula, the first Chern number then becomes

$$\int_\Sigma c_1(T\Sigma) = -\frac{1}{2\pi i}\int_\Sigma F = \frac{1}{2\pi}\int_\Sigma iF = \frac{1}{2\pi}\int_\Sigma K_G \, d\mathrm{vol}_\Sigma = \chi(\Sigma).$$

This is the most famous explicit computation of a first Chern number, and is the main one that you should commit to memory if you don't have space for any others. Combining it with the results of the previous section now gives a new interpretation of the Euler characteristic:

THEOREM 30.19 (Poincaré-Hopf). *For any vector field $X \in \mathfrak{X}(\Sigma)$ with at most finitely-many zeroes on a closed oriented surface Σ, the algebraic count of zeroes is*

$$\#X^{-1}(0) = \chi(\Sigma).$$

□

I recommend taking a moment to think about what this implies for the most familiar surfaces. For the torus $\mathbb{T}^2$, whose Euler characteristic according to Exercise 29.16 is 0, it is consistent with the observation that nowhere-zero vector fields on $\mathbb{T}^2$ are easy to construct. The most famous consequence of the Poincaré-Hopf theorem applies to S^2, whose Euler characteristic is 2: it is often summarized by the colorful phrase, "you cannot comb the hair on a sphere".

COROLLARY 30.20. *There does not exist a nowhere-zero vector field on S^2.* □

EXERCISE 30.21. For a closed oriented surface Σ_g of genus $g \geqslant 0$, we can use the Poincaré-Hopf theorem to compute $\chi(\Sigma_g)$ without needing to choose triangulations. Recall from Exercise 29.22 the notion of a *pair of pants*.

(a) Show that a pair of pants admits a smooth vector field that is tangent to the boundary and nonzero there, and has only one zero in the interior, with index -1.
Hint: Just try to draw the flow lines. They should form the leaves of a foliation, with one singular point where two leaves intersect transversely.

(b) By gluing together pairs of pants, show that Σ_g admits a vector field with exactly $2-2g$ zeroes, all of index -1.

30.4. Addendum: counting zeroes in general. Having seen the definition of the first Chern class, it will surely not surprise you to learn that there is also a second, and a third and so forth: for every $k \in \mathbb{N}$ one can associate to every complex vector bundle $\pi : E \to M$ of arbitrary rank a *kth Chern class*

$$c_k(E) \in H^{2k}_{\mathrm{dR}}(M).$$

Its definition when either $k \geqslant 2$ or $\mathrm{rank}(E) > 1$ is more complicated than we have space to discuss here: this is the subject of a large sub-branch of differential topology known as *Chern-Weil theory*, which is one of the topics we might discuss near the end of next semester's followup course. One can also define analogous so-called *characteristic classes* for real vector bundles $E \to M$, such as the *Pontryagin classes*

$$p_k(E) \in H^{4k}_{\mathrm{dR}}(M)$$

for each $k \in \mathbb{N}$. In Chern-Weil theory, characteristic classes of a bundle $E \to M$ with structure group G are always constructed in terms of closed forms determined by the curvature of some chosen G-compatible connection, on which the cohomology class turns out not to depend. One can show as in §30.2 that the integrals of these classes over closed oriented submanifolds of suitable dimensions are always integers, despite this being highly nonobvious from their definition. This hints at the fact that all characteristic classes can also be constructed by completely different methods, using algebraic topology, where they live naturally in $\mathbb{Z}$-modules such as singular or Čech cohomology with integer coefficients, rather than the real vector space $H^*_{\mathrm{dR}}(M)$. (The major exceptions to this last statement are the Stiefel-Whitney classes, which can be defined for all real vector bundles and take values in cohomology with $\mathbb{Z}_2$ coefficients, thus there is no sensible way to define them in de Rham cohomology.) The fact that the Pontryagin numbers are integers played a major role e.g. in Milnor's discovery that the topological manifold S^7 admits smooth structures not diffeomorphic to its standard one. The fact that the widely differing constructions of characteristic classes via algebraic topology vs. Chern-Weil theory give equivalent results is also a deep theorem with many applications.

I'd like to add a word about one other characteristic class which places the discussion of §30.2 into a wider context. There is a certain perspective from which it is not at all surprising that the question "How many zeroes should a section $s : M \to E$ have?" might have a well-defined answer. The idea is roughly as follows: suppose $\pi : E \to M$ is an *oriented* real vector bundle of rank n over a closed oriented n-manifold M, and call a section $s \in \Gamma(E)$ *generic* if for every point $p \in s^{-1}(0)$ in its zero-set, the linearization

$$Ds(p) : T_pM \to E_p$$

is invertible. As mentioned in Remark 30.12, there is always an open and dense set of sections in $\Gamma(E)$ that satisfy this condition, and the inverse function theorem then implies that the zeroes of s are isolated; since M is assumed compact, this means there are only finitely many. Generalizing Exercise 30.11, one can now associate an index $\mathrm{ind}(s; p) = \pm 1$ to each zero by defining it to be $+1$ if $Ds(p)$ is orientation preserving and -1 if $Ds(p)$ is orientation reversing.

The key idea now is to regard the zero set $s^{-1}(0)$ as a compact oriented 0-dimensional submanifold of M, with the orientation of each point defined by the sign of $\mathrm{ind}(s; p)$. Now if $s_0, s_1 \in \Gamma(E)$

are two generic sections, we can find a smooth homotopy between them, i.e. a map

$$H : [0, 1] \times M \to E$$

such that $s_t := H(t, \cdot) \in \Gamma(E)$ for each t; such a map always exists, for instance the linear interpolation $H(t, p) := ts_1(p) + (1 - t)s_0(p)$. By a nontrivial bit of transversality theory, one can always make a small perturbation of H so as to assume without loss of generality that its image in E meets the zero-section transversely, in which case $H^{-1}(0) \subset [0, 1] \times M$ is a smooth oriented 1-dimensional submanifold with boundary. Then

$$\partial \left(H^{-1}(0)\right) = \left(\{1\} \times s_1^{-1}(0)\right) \cup \left(\{0\} \times (-s_0^{-1}(0))\right),$$

where the minus sign on the right hand side indicates reversal of orientation. The 1-manifold $H^{-1}(0)$ will generally have multiple connected components, which come in three flavors:

(1) Circles in the interior of $[0, 1] \times M$;
(2) Arcs with one boundary point in $\{1\} \times s_1^{-1}(0)$, and the other a point in $\{0\} \times s_0^{-1}(0)$ with the same orientation;
(3) Arcs with both boundary points in either $\{1\} \times s_1^{-1}(0)$ or $\{0\} \times s_0^{-1}(0)$, having opposite orientations.

The result is that the points in the disjoint union of $s_1^{-1}(0)$ with $s_0^{-1}(0)$ come in pairs: matching pairs of zeros of s_1 and s_0, or cancelling pairs of zeros of s_1 alone or s_0 alone. Thus the count of positive points in $s_1^{-1}(0)$ minus negative points in $s_1^{-1}(0)$ is the same as the corresponding count for s_0, and we conclude that for all *generic* sections $s \in \Gamma(E)$, the algebraic count

$$\#s^{-1}(0) = \sum_{p \in s^{-1}(0)} \operatorname{ind}(s; p) \in \mathbb{Z}$$

is the same. This number is called the **Euler number** of the bundle $E \to M$, and it corresponds to an **Euler class** $e(E) \in H^n_{\mathrm{dR}}(M)$ such that $\int_M e(E)$ is the Euler number. In this context, Theorem 30.16 can be rephrased as the statement that for any complex line bundle $E \to M$, if one regards it as an oriented real vector bundle of rank 2, its Euler class matches its first Chern class. The reason however for the terminology is that when E is TM for a closed oriented manifold M, its Euler number matches the Euler characteristic:

$$\int_M e(E) = \chi(M).$$

This is the general version of the Poincaré-Hopf theorem. It can be proved in various ways, depending on whether one prefers to define $e(E)$ via Chern-Weil theory or algebraic topology. Add this to our to-do list for next semester.

If you want to read the full details on why algebraic counts like $\#s^{-1}(0)$ do not depend on the choice of generic section, and how to generalize them without always assuming $\operatorname{ind}(s; p) = \pm 1$, I highly recommend Milnor's short book [**Mil97**]. It's something every mathematics student should read sooner or later.

Exercise 30.22. The argument sketched above for proving $\#s_0^{-1}(0) = \#s_1^{-1}(0)$ appealed to the classification of compact 1-manifolds with boundary, i.e. their connected components are each diffeomorphic to either a circle S^1 or a compact interval $[0, 1]$. This is a basic result in topology, but one doesn't really need to use it for this purpose: the main fact we actually needed was that whenever M is a compact oriented 1-manifold with boundary, the signed count of boundary points vanishes:

$$\sum_{p \in \partial M} \varepsilon(p) = 0,$$

where $\varepsilon : \partial M \to \{1, -1\}$ is the boundary orientation (see §12.1). Prove this without assuming anything about the topology of M.
Hint: Integrate an exact 1-form over M.

31. Sectional curvature

In this last lecture of the semester, we'll introduce a generalization of the Gaussian curvature that makes sense in all dimensions, not just on surfaces. Like the Gaussian curvature on surfaces, the sectional curvature of a Riemannian n-manifold contains all the same information as the Riemann tensor, but packaged in a more useful way: for instance, sectional curvature makes it possible to define the notions of positive/negative/zero curvature in arbitrary dimensions and observe qualitative distinctions between them. We will motivate the definition by investigating a natural question about geodesics: when can we conclude that a geodesic segment connecting two (not necessarily nearby) points actually is the *only* geodesic connecting them?

The following thought experiment shows that this question clearly has something to do with curvature: suppose $\ell_1, \ell_2 \subset \Sigma$ are two distinct embedded geodesic curves connecting distinct points p and q in a Riemannian 2-manifold (Σ, g). (We know from the example of S^2 that this scenario sometimes happens.) If ℓ_1 and ℓ_2 do not intersect at any other point between p and q, and if they are also "close" to each other in some sense, then they form the edges of a smooth polygon $P \subset \Sigma$ with vertices at p and q. Their angles $\alpha, \beta \in [0, 2\pi]$ at p and q must be positive in this situation, because any two geodesics that meet tangentially must be identical. The Gauss-Bonnet formula then gives

$$0 < \alpha + \beta = \int_P K_G \, d\mathrm{vol}_\Sigma,$$

which is a contradiction if the Gaussian curvature of (Σ, g) happens to be everywhere nonpositive. This proves a uniqueness result: if curvature is never positive, then any geodesic from p to q is the *only* geodesic among nearby curves connecting those two points. Our aim is to extend this observation to arbitrary dimensions.

31.1. The second variation of the energy functional. Throughout this lecture, (M, g) is a Riemannian manifold of arbitrary dimension. A more elaborate version of the question mentioned above can be stated as follows:

QUESTION 31.1. *On a Riemannian manifold (M, g), suppose $\gamma : [a, b] \to M$ is a geodesic segment with $\gamma(a) = p$ and $\gamma(b) = q$. Is there any other geodesic segment from p to q near γ? Is γ in fact the shortest path between p and q?*

The issue here is different from what we discussed in §23.2, because we are now allowing the points p and q to be arbitrarily far apart. It is clear from examples that non-short geodesics connecting two points need not be unique: consider for instance the geodesics in $S^2 \subset \mathbb{R}^3$ from the north pole to the south pole, which come in a whole 1-parameter, all equally long. We'll find that this is only allowed because S^2 has positive curvature; it *cannot* happen on a surface with negative or zero curvature.

To better understand the global relationship between geodesics and length, we can apply an infinite-dimensional version of the "second derivative test" to the length functional. As in §22.4, it turns out to be easier for this purpose to work with the energy functional instead of the length functional, because results about the latter can be derived from results about the former, but energy is easier to compute with. Recall the following notation: for two parameter values $a < b \in \mathbb{R}$ and points $p, q \in M$, we denote by

$$\mathcal{P} := C^\infty([a, b], M; p, q)$$

the space of smooth paths $\gamma : [a,b] \to M$ starting at p and ending at q. We think of this space intuitively as an infinite-dimensional smooth manifold, with tangent spaces

$$T_\gamma \mathcal{P} := \{\eta \in \Gamma(\gamma^* TM) \mid \eta(a) = 0 \text{ and } \eta(b) = 0\}.$$

We then have two functionals $\ell, E : \mathcal{P} \to \mathbb{R}$, the length

$$\ell(\gamma) = \int_a^b |\dot\gamma(t)|\, dt,$$

and the energy

$$E(\gamma) = \frac{1}{2}\int_a^b |\dot\gamma(t)|^2\, dt,$$

where an extra factor of $1/2$ has been inserted in front of the energy functional to make some of the expressions below look a bit nicer. For a smooth 1-parameter family of paths $\gamma_s \in \mathcal{P}$ with $\gamma_0 = \gamma$ and $\partial_s \gamma_s|_{s=0} = \eta \in T_\gamma\mathcal{P}$, we computed in §22.4 the *first variation* of the energy functional:

$$dE(\gamma)\eta := \left.\frac{d}{ds} E(\gamma_s)\right|_{s=0} = \int_a^b \langle -\nabla_t \dot\gamma(t), \eta(t)\rangle\, dt.$$

We can express this more succinctly by defining an inner product on the space of sections $\Gamma(\gamma^*TM)$ for each $\gamma \in \mathcal{P}$: for two such sections ξ and η, let[76]

$$\langle \xi, \eta\rangle_{L^2} = \int_a^b \langle \xi(t), \eta(t)\rangle\, dt.$$

Informally, we can think of $\langle\ ,\ \rangle_{L^2}$ as defining a Riemannian metric on $\mathcal{P}$. Now the first variation can be expressed as

$$dE(\gamma)\eta = \langle \nabla E(\gamma), \eta\rangle_{L^2},$$

where

$$\nabla E(\gamma) := -\nabla_t \dot\gamma \in \Gamma(\gamma^*TM)$$

is the so-called L^2*-gradient* of the energy functional. In this notation, γ is a geodesic if and only if $\nabla E(\gamma) = 0$.

Informally again, we think of ∇E as a vector field on $\mathcal{P}$ which represents the first derivative of E, though it would be more accurate to call it a section of a vector bundle $\mathcal{E} \to \mathcal{P}$ with fibers $\mathcal{E}_\gamma := \Gamma(\gamma^*TM)$, as $\nabla E(\gamma)$ need not take values in the subspace $T_\gamma\mathcal{P} \subset \Gamma(\gamma^*TM)$. In any case, we would like to compute the derivative of this section, i.e. it's *linearization* (cf. Exercise 19.7) at a point $\gamma \in \mathcal{P}$ where $\nabla E(\gamma) = 0$, and interpret it as the Hessian of E at the critical point γ. For $\eta \in T_\gamma\mathcal{P}$, we choose a 1-parameter family $\gamma_s \in \mathcal{P}$ with $\gamma_0 = \gamma$ and $\partial_s\gamma_s|_{s=0} = \eta$ and define a "covariant derivative" $\nabla_\eta \nabla E \in \Gamma(\gamma^*TM)$ by

$$(\nabla_\eta \nabla E)(t) := \nabla_s(\nabla E(\gamma_s)(t))|_{s=0}\,.$$

A quick computation using the definition of the Riemann tensor shows that this does indeed only depend on η rather than the 1-parameter family γ_s:

$$\nabla_s(\nabla E(\gamma_s))|_{s=0} = -\nabla_s\nabla_t\partial_t\gamma_s|_{s=0} = -\nabla_t\nabla_s\partial_t\gamma_s - R(\partial_s\gamma_s, \partial_t\gamma_s)\partial_t\gamma_s|_{s=0} = -\nabla_t^2\eta - R(\eta,\dot\gamma)\dot\gamma.$$

With this calculation as motivation, define for any $\gamma \in \mathcal{P}$ a linear operator

$$\nabla^2 E(\gamma) : \Gamma(\gamma^*TM) \to \Gamma(\gamma^*TM) : \eta \mapsto -\nabla_t^2\eta - R(\eta,\dot\gamma)\dot\gamma. \tag{31.1}$$

We can now state the *second variation formula*:

[76]The subscript L^2 refers to the standard notation for the Hilbert space completion of $\Gamma(\gamma^*TM)$ with respect to this inner product.

Proposition 31.2. *Suppose $\gamma \in \mathcal{P}$ is a geodesic and $\gamma_{\sigma,\tau} \in \mathcal{P}$ is a smooth 2-parameter family of paths with $\gamma_{0,0} = \gamma$, with variations $\xi, \eta \in T_\gamma\mathcal{P}$ defined by*

$$\xi = \partial_\sigma \gamma_{\sigma,\tau}|_{\sigma=\tau=0} \qquad \text{and} \qquad \eta = \partial_\tau \gamma_{\sigma,\tau}|_{\sigma=\tau=0}.$$

Then

$$\frac{\partial^2}{\partial\sigma\partial\tau} E(\gamma_{\sigma\tau})\bigg|_{\sigma=\tau=0} = \langle \nabla^2 E(\gamma)\xi, \eta\rangle_{L^2}.$$

Proof. Compute:

$$\begin{aligned}
\frac{\partial^2}{\partial\sigma\partial\tau} E(\gamma_{\sigma,\tau})\bigg|_{\sigma=\tau=0} &= \frac{\partial}{\partial\sigma}\left(\frac{\partial}{\partial\tau} E(\gamma_{\sigma,\tau})\bigg|_{\tau=0}\right)\bigg|_{\sigma=0} = \frac{\partial}{\partial\sigma}\left\langle \nabla E(\gamma_{\sigma,0}),\, \partial_\tau\gamma_{\sigma,\tau}|_{\tau=0}\right\rangle_{L^2}\bigg|_{\sigma=0} \\
&= \frac{\partial}{\partial\sigma}\int_a^b \left\langle \nabla E(\gamma_{\sigma,0})(t),\, \partial_\tau\gamma_{\sigma,\tau}(t)|_{\tau=0}\right\rangle\, dt\bigg|_{\sigma=0} \\
&= \int_a^b \left\langle \nabla_\sigma \nabla E(\gamma_{\sigma,0})(t)|_{\sigma=0}, \eta(t)\right\rangle\, dt + \int_a^b \left\langle \nabla E(\gamma)(t),\, \nabla_\sigma\partial_\tau\gamma_{\sigma,\tau}|_{\sigma=\tau=0}\right\rangle\, dt \\
&= \int_a^b \langle (\nabla_\xi \nabla E)(t), \eta(t)\rangle\, dt = \langle \nabla^2 E(\gamma)\xi, \eta\rangle_{L^2}.
\end{aligned}$$

Note that we used the assumption that γ is geodesic, so $\nabla E(\gamma) = 0$. □

For a 1-parameter family of paths $\{\gamma_s \in \mathcal{P}\}_{s\in(-\epsilon,\epsilon)}$, one can plug $\gamma_{\sigma,\tau} := \gamma_{\sigma+\tau}$ into the second variation formula and extract from it the first nontrivial term in the Taylor expansion of $E(\gamma_s)$ as a function of s: we have

$$\begin{aligned}
E(\gamma_s) &= E(\gamma) + \frac{1}{2}s^2 \cdot \langle \eta, \nabla^2 E(\gamma)\eta\rangle_{L^2} + O(|s|^3) \\
&= E(\gamma) - \frac{1}{2}s^2 \cdot \langle \eta, \nabla_t^2\eta + R(\eta,\dot\gamma)\dot\gamma\rangle_{L^2} + O(|s|^3)
\end{aligned}$$

If we want a criterion to guarantee that $E(\gamma_s)$ is locally minimized at $s = 0$, the first important thing to understand is whether the coefficient on the quadratic term is positive. This coefficient breaks up into two terms, and the sign of the first one can easily be ascertained after integrating by parts: since $\partial_t\langle\eta, \nabla_t\eta\rangle = |\nabla_t\eta|^2 + \langle\eta, \nabla_t^2\eta\rangle$ and η vanishes at the end points, we have

$$-\langle \eta, \nabla_t^2\eta\rangle_{L^2} = \langle \nabla_t\eta, \nabla_t\eta\rangle_{L^2} =: \|\nabla_t\eta\|_{L^2}^2 \geqslant 0,$$

with strict inequality unless η is parallel along γ, which would mean $\eta \equiv 0$ since $\eta(a) = 0$. The other term is

$$-\langle \eta, R(\eta,\dot\gamma)\dot\gamma\rangle_{L^2} = -\int_a^b \operatorname{Riem}(\eta(t),\eta(t),\dot\gamma(t),\dot\gamma(t))\, dt, \tag{31.2}$$

and evidently the question of whether this must be positive or not depends in some way on the curvature. This question merits a more thorough discussion.

31.2. Sectional curvature. At any given point, the vectors $\dot\gamma(t)$ and $\eta(t)$ appearing in the integrand of (31.2) can be completely arbitrary, so the real question here is whether any meaningful condition can be formulated that would determine the signs of real numbers of the form

$$\langle X, R(X,Y)Y\rangle = \operatorname{Riem}(X,X,Y,Y)$$

for arbitrary tangent vectors $X, Y \in T_pM$ at a point $p \in M$. We've seen products like this before: in the case $\dim M = 2$, if X and Y are taken to be a basis of T_pM, we saw in (27.7) that $\operatorname{Riem}(X, X, Y, Y)$ determines the Gaussian curvature by the formula

$$K_G(p) = \frac{\operatorname{Riem}(X, X, Y, Y)}{|d\mathrm{vol}_M(X, Y)|^2} = \frac{\operatorname{Riem}(X, X, Y, Y)}{\langle X, X\rangle\langle Y, Y\rangle - \langle X, Y\rangle^2}, \tag{31.3}$$

where in the second expression we have used Exercise 11.12 to write

$$d\mathrm{vol}_M(X, Y) = \sqrt{\det\begin{pmatrix}\langle X, X\rangle & \langle X, Y\rangle \\ \langle Y, X\rangle & \langle Y, Y\rangle\end{pmatrix}}.$$

The fact that this formula remains equally valid for any choice of the basis $X, Y \in T_pM$ is both non-obvious and useful, and the following definition will help us generalize it to higher dimensions.

Definition 31.3. Suppose (M, g) is a Riemannian manifold and $P \subset T_pM$ is a 2-dimensional subspace in the tangent space at some point $p \in M$. The **sectional curvature** $K_S(P) \in \mathbb{R}$ along P is defined as follows. Choose a sufficiently small neighborhood $0 \in \mathcal{O}_p \subset T_pM$ so that $\exp_p$ restricts to a diffeomorphism from $\mathcal{O}_p$ to a neighborhood of p in M. Then

$$\Sigma_P := \exp_p(\mathcal{O}_p \cap P) \subset M$$

is a 2-dimensional submanifold containing p, and we set

$$K_S(P) := K_G(p),$$

where $K_G : \Sigma_P \to \mathbb{R}$ is the Gaussian curvature of Σ_P with respect to the Riemannian metric induced by its embedding in (M, g).

Lemma 31.4. *For the embedded surface $\Sigma_P \subset M$ through $p \in M$ in Definition 31.3, the second fundamental form $\mathrm{II} : T\Sigma_P \oplus T\Sigma_P \to (T\Sigma_P)^\perp$ vanishes at p.*

Proof. Given any $X, Y \in T_p\Sigma_P$ and constants $a, b \in \mathbb{R}$, the geodesic $\gamma(t) = \exp_p(t(aX + bY))$ lies in Σ_P for t close to 0, thus $aX + bY$ extends to a parallel vector field on Σ_P along this curve, giving

$$\mathrm{II}(aX + bY, aX + bY) = (\nabla_{aX+bY}(aX + bY))^\perp = 0$$

at $p = \gamma(0)$. Since $a, b \in \mathbb{R}$ were arbitrary, this implies $\mathrm{II}(X, X) = \mathrm{II}(Y, Y) = 0$, and $\mathrm{II}(X, Y) = 0$ then follows from $\mathrm{II}(X + Y, X + Y) = 0$ using bilinearity and symmetry. □

Remark 31.5. If you remember the construction of Riemann normal coordinates in §23.1, you may have noticed that we used a trick from that construction in Lemma 31.4. We could alternatively have used normal coordinates to deduce the lemma: if we pick any orthonormal vectors $X_1, X_2 \in T_p\Sigma_P$ and extend them to an orthonormal basis $X_1, \ldots, X_n$ of T_pM, we obtain a normal coordinate system $(x^1, \ldots, x^n)$ near p in which Σ_P looks like the "flat" plane $\{x^3 = \ldots = x^n = 0\}$ and X_1 and X_2 match the coordinate vector fields ∂_1, ∂_2. The vanishing of II at p then follows from the fact that in normal coordinates centered at p, the Christoffel symbols (which are equivalent to the covariant derivatives $\nabla_j(\partial_k)$) vanish at p.

Proposition 31.6. *On a Riemannian manifold (M, g), the relation*

$$\operatorname{Riem}(X, X, Y, Y) = K_S(P) \cdot \operatorname{Area}(X, Y)^2 = K_S(P) \cdot \left(\langle X, X\rangle\langle Y, Y\rangle - \langle X, Y\rangle^2\right)$$

holds for any 2-dimensional subspace $P \subset T_pM$ and any $X, Y \in P$, where $\operatorname{Area}(X, Y) \geqslant 0$ denotes the area of the parallelogram in T_pM spanned by X and Y, as measured with respect to the metric $\langle\ ,\ \rangle$. In particular, whenever X, Y is a basis of P, this gives rise to the formula

$$K_S(P) = \frac{\operatorname{Riem}(X, X, Y, Y)}{\langle X, X\rangle\langle Y, Y\rangle - \langle X, Y\rangle^2}$$

for the sectional curvature of (M, g) along P.

PROOF. Both sides of the relation vanish if X and Y are linearly dependent: for the left hand side this follows from the symmetries of the Riemann tensor, and on the right hand side it follows because the parallelogram spanned by X and Y has no area. We are thus free to assume X and Y form a basis of $P = T_p\Sigma_P$. Lemma 31.4 and the Gauss equation (Prop. 28.5) now imply that the Riemann tensor of the submanifold Σ_P at p is just the restriction of the Riemann tensor $R \in \Gamma(T_3^1 M)$ of (M, g) to the subspace $T_p\Sigma_P \subset T_pM$. The result thus follows from the definition of sectional curvature and the formula (31.3) for the Gaussian curvature. □

Just as the Gaussian curvature determines the Riemann tensor in dimension 2, one can show that the Riemann tensor is determined in general by the sectional curvature. The proof of this requires a more thorough discussion of the symmetries of the Riemann tensor than we have given so far, thus we will save it for next semester, and merely record it here as a fact:

PROPOSITION 31.7. *The Riemann tensor $R \in \Gamma(T_3^1 M)$ at a point $p \in M$ is determined by the values of the sectional curvature $K_S(P) \in \mathbb{R}$ on all possible 2-dimensional subspaces $P \subset T_pM$. In particular, a Riemannian manifold is locally flat if and only if its sectional curvature vanishes identically.*

DEFINITION 31.8. A Riemannian manifold (M, g) is said to have **positive** (or **negative** or **zero**) **curvature** if $K_S(P)$ is positive (or negative or zero respectively) for every 2-dimensional subspace $P \subset T_pM$ at every point $p \in M$.

In §24.4, we discussed three emblematic examples of Riemannian n-manifolds that can be defined for each $n \geqslant 2$: Euclidean space $\mathbb{R}^n$, the sphere S^n and hyperbolic space H^n. The following exercises show that these are manifolds of constant sectional curvature, vanishing in the case of $\mathbb{R}^n$, positive for S^n and negative for H^n.

EXERCISE 31.9. Suppose (M, g) is a Riemannian manifold, $p, q \in M$ are points and $P \subset T_pM$ and $Q \subset T_qM$ are 2-dimensional subspaces such that there exists an isometry $\varphi \in \mathrm{Isom}(M, g)$ with $\varphi(p) = q$ and $\varphi_*P = Q$. Show that the sectional curvature of (M, g) satisfies $K_S(P) = K_S(Q)$.

EXERCISE 31.10. Show that H^n contains a point $p \in H^n$ with a 2-dimensional subspace $P \subset T_pH^n$ for which the surface $\Sigma_P \subset H^n$ appearing in the definition of sectional curvature is isometric to the hyperbolic plane H^2. Deduce from this and Exercise 31.9 that H^n has sectional curvature $K_S(P) = -1$ for all 2-dimensional subspaces $P \subset TH^n$.
Hint: What kind of submanifold is the intersection of $H^n \subset \mathbb{R}^{n+1}$ with the subspace $\mathbb{R}^3 \times \{0\} \subset \mathbb{R}^{n+1}$? Recall from §24.4 that the geodesics in H^n can be written down quite explicitly.

EXERCISE 31.11. Use a similar trick to Exercise 31.10 to prove that the sphere $S^n \subset \mathbb{R}^{n+1}$ has constant sectional curvature $+1$, and (this one is more obvious) Euclidean space $\mathbb{R}^n$ has vanishing sectionsl curvature.

An important result we will prove next semester states that up to rescaling by positive constants, these are the *only* geodesically complete and *simply connected* Riemannian n-manifolds with constant sectional curvature. This statement is no longer true if one removes the words "simply connected": there are many interesting examples of manifolds with constant sectional curvature, also called *Riemannian space forms.* But if you know a bit of covering space theory from algebraic topology, you will recognize that the specific examples S^n, $\mathbb{R}^n$ and H^n always remain relevant, because they appear as universal covers of space forms, implying that every Riemannian space form is isometric to a quotient of one of these three examples by a discrete group acting by isometries.

EXERCISE 31.12. The copy of H^2 embedded isometrically in H^n in Exercise 31.10 is an example of a totally geodesic submanifold: we say that a submanifold $N \subset M$ in a Riemannian manifold (M, g) is **totally geodesic** if all geodesics in N (with its induced metric) are also geodesics in (M, g). One nice way to recognize totally geodesic submanifolds is via isometries: suppose $N \subset M$ is a submanifold that is also the fixed-point set $N = \mathrm{Fix}(\varphi) := \{p \in M \mid \varphi(p) = p\}$ of an isometry $\varphi \in \mathrm{Isom}(M, g)$. Show that N is then totally geodesic.

EXERCISE 31.13. Find an example of an isometry on the hyperbolic n-space H^n whose fixed point set is a submanifold isometric to H^2.

31.3. Nonpositive curvature and geodesics. The notion of sectional curvature now provides a simply-stated condition that is sufficient to guarantee that geodesics are local minima of the length functional. We recall from §31.1 the notation $\mathcal{P} := C^\infty([a, b], M\,; p, q)$ and $T_\gamma \mathcal{P} := \{\eta \in \Gamma(\gamma^*TM) \mid \eta(a) = 0 \text{ and } \eta(b) = 0\}$.

LEMMA 31.14. *Suppose (M, g) is a Riemannian manifold with everywhere nonpositive sectional curvature $K_S \leqslant 0$, and $\gamma \in \mathcal{P}$ is a nonconstant geodesic. Then the second variation operator $\nabla^2 E(\gamma) : \Gamma(\gamma^*TM) \to \Gamma(\gamma^*TM)$ defined in (31.1) satisfies*

$$\langle \nabla^2 E(\gamma)\eta, \eta\rangle_{L^2} \geqslant 0$$

for all $\eta \in \Gamma(\gamma^*TM)$, and the inequality is strict for all nontrivial $\eta \in T_\gamma\mathcal{P}$.

PROOF. We have already observed that $\langle -\nabla_t^2\eta, \eta\rangle_{L^2} = \|\nabla_t\eta\|_{L^2}^2 \geqslant 0$, with strict inequality whenever η is a nontrivial section in $T_\gamma\mathcal{P}$. Using Proposition 31.6, the other term in $\langle \nabla^2 E(\gamma)\eta, \eta\rangle_{L^2}$ is the integral from a to b of

$$-\mathrm{Riem}(\eta(t), \eta(t), \dot\gamma(t), \dot\gamma(t)) = -K_S(P_t) \cdot \mathrm{Area}(\eta(t), \dot\gamma(t))^2 \geqslant 0,$$

where $P_t \subset T_{\gamma(t)}M$ can be taken to be any 2-dimensional subspace containing $\eta(t)$ and $\dot\gamma(t)$. □

THEOREM 31.15. *Suppose (M, g) is a Riemannian manifold with nonpositive sectional curvature and $\gamma : [a, b] \to M$ is a geodesic connecting $\gamma(a) = p$ to $\gamma(b) = q$. Then for any smooth 1-parameter family of paths $\gamma_s : [a, b] \to M$ with $\gamma_s(a) = p$, $\gamma_s(b) = q$ and $\gamma_0 \equiv \gamma$ such that $\partial_s\gamma_s|_{s=0}$ is not everywhere tangent to $\dot\gamma$, there is a number $\epsilon > 0$ such that:*

(1) γ is the only geodesic among the paths γ_s for $s \in (-\epsilon, \epsilon)$;
(2) For all paths γ_s with $s \in (-\epsilon, \epsilon)$ and $s \neq 0$,

$$\ell(\gamma_s) > \ell(\gamma).$$

PROOF. The result is already clear if $p = q$ and γ is a constant path, so let us assume γ is nonconstant, in which case $\dot\gamma$ is everywhere nonzero since it is a geodesic.

The second statement is easily proved if we replace length with energy, because by the second variation formula and Lemma 31.14,

$$\left.\frac{d^2}{ds^2}E(\gamma_s)\right|_{s=0} = \langle \nabla^2 E(\gamma)\eta, \eta\rangle_{L^2} > 0 \qquad \text{for} \qquad \eta := \partial_s\gamma_s|_{s=0} \neq 0.$$

In order to apply this result to the length functional, we can use the same trick as in the proof of Corollary 22.11 and reparametrize each of the paths $\gamma_s : [a, b] \to M$ so that they have constant speed; here we can restrict s if necessary to a neighborhood of 0 so that $\dot\gamma_s(t)$ can always be assumed nonzero, and after reparametrizing, its norm is a positive number indepenent of t, namely

$$|\dot\gamma_s(t)| =: v_s = \frac{\ell(\gamma_s)}{b - a}.$$

The assumption that $\eta = \partial_s \gamma_s|_{s=0}$ is not everywhere tangent to $\dot{\gamma}$ implies that after this reparametrization, η is still not everywhere zero. Since γ_0 is a geodesic, we know from Corollary 22.11 that $\frac{d}{ds}\ell(\gamma_s)\big|_{s=0} = 0$ and thus $\partial_s v_s|_{s=0} = 0$, so

$$\begin{aligned}
\frac{d^2}{ds^2}\ell(\gamma_s)\bigg|_{s=0} &= \frac{d}{ds}\int_a^b \partial_s\sqrt{|\dot{\gamma}_s(t)|^2}\,dt\bigg|_{s=0} = \frac{d}{ds}\int_a^b \frac{1}{2\sqrt{|\dot{\gamma}_s(t)|^2}}\partial_s|\dot{\gamma}_s(t)|^2\,dt\bigg|_{s=0} \\
&= \int_a^b \frac{\partial}{\partial s}\left(\frac{1}{2v_s}\partial_s|\dot{\gamma}_s(t)|^2\right)\bigg|_{s=0} dt = \frac{1}{2v_0}\int_a^b \partial_s^2|\dot{\gamma}_s(t)|^2\big|_{s=0}\,dt \\
&= \frac{1}{v_0}\frac{d^2}{ds^2}\frac{1}{2}\left(\int_a^b |\dot{\gamma}_s(t)|^2\,dt\right)\bigg|_{s=0} = \frac{1}{v_0}\frac{d^2}{ds^2}E(\gamma_s)\bigg|_{s=0} > 0.
\end{aligned}$$

This proves that $\ell(\gamma_s) > \ell(\gamma)$ for $s \ne 0$ close to 0.

To see that γ is the only geodesic among the family γ_s for s close to 0, we can differentiate the L^2-inner product of $\nabla E(\gamma_s) = \nabla_t\dot{\gamma}_s \in \Gamma(\gamma_s^*TM)$ with $\eta_s := \partial_s\gamma_s \in \Gamma(\gamma_s^*TM)$ at $s = 0$, using the fact that $\nabla E(\gamma) = 0$:

$$\frac{d}{ds}\langle\nabla E(\gamma_s), \eta_s\rangle_{L^2}\bigg|_{s=0} = \langle\nabla^2 E(\gamma)\eta, \eta\rangle_{L^2} + \langle\nabla E(\gamma), \nabla_s\eta_s|_{s=0}\rangle_{L^2} = \langle\nabla^2 E(\gamma)\eta, \eta\rangle_{L^2} > 0.$$

It follows that $\langle\nabla E(\gamma_s), \eta_s\rangle \ne 0$ for sufficiently small $|s| \ne 0$, implying that $\nabla E(\gamma_s)$ itself cannot be 0, so γ_s is not a geodesic. □

These results give a small hint of the larger story of geodesics on manifolds with nonpositive curvature. The general pattern is that nonpositive curvature implies uniqueness phenomena that clearly do not hold in simple examples with positive curvature, such as the sphere. Here is another example, which we will prove next semester as a consequence of the Hopf-Rinow and Cartan-Hadamard theorems:

THEOREM. *Suppose (M, g) is a connected and geodesically complete Riemannian manifold with nonpositive sectional curvature, and $p, q \in M$ are two points. Then every homotopy class (with fixed end points) of paths from p to q contains exactly one geodesic segment, up to parametrization.*

The intuition here, which you will find reasonable if you compare what you know about the 2-sphere and the hyperbolic plane, is that positive curvature typically causes two geodesics emerging from the same point to come back together at a later time, whereas negative curvature tries always to force them further apart. On a simply connected manifold, the latter guarantees that they never meet again, producing an absolutely *unique* shortest path between p and q. On surfaces, a slightly weaker version of the uniqueness statement in this theorem can be derived from the Gauss-Bonnet formula:

EXERCISE 31.16. Assume (Σ, g) is a Riemannian 2-manifold with $K_G \leqslant 0$.

(a) Show that (Σ, g) does not admit any *periodic* geodesic (i.e. a geodesic $\gamma : \mathbb{R} \to \Sigma$ satisfying $\gamma(t + T) = \gamma(t)$ for all $t \in \mathbb{R}$ and some fixed $T > 0$) whose image bounds an embedded disk.

(b) Given a pair of distinct points $p, q \in \Sigma$ and a pair of geodesic segments $\gamma_0, \gamma_1 : [0, 1] \to \Sigma$ with $\gamma_0(0) = \gamma_1(0) = p$ and $\gamma_0(1) = \gamma_1(1) = q$, show that there does not exist any smooth family of paths $\{\gamma_s : [0, 1] \to \Sigma\}_{s\in[0,1]}$ from p to q, matching the given geodesics for $s = 0, 1$, such that the map $[0, 1] \times (0, 1) \to \Sigma : (s, t) \mapsto \gamma_s(t)$ is an embedding.

(c) Find an example of a periodic geodesic on a surface with nonpositive Gaussian curvature. *(Note that by part (a), it had better not form the boundary of an embedded disk.)*

(d) Show that the phenomenon ruled out by part (b) can actually happen on S^2.

31.4. Addendum: Gauss and sectional curvature with indefinite metrics. I have thus far completely excluded metrics of indefinite signature in this lecture, and also in the discussion of Gaussian curvature in §27.3. For sectional curvature on a general pseudo-Riemannian manifold (M, g), the standard procedure is to take the formula in Proposition 31.6 as a definition, namely

$$K_S(P) = \frac{\operatorname{Riem}(X, X, Y, Y)}{\langle X, X\rangle\langle Y, Y\rangle - \langle X, Y\rangle^2}, \tag{31.4}$$

assuming X, Y to be any basis of the 2-dimensional subspace $P \subset T_pM$. There are two very large caveats here: first, the determinant in the denominator is nonzero if and only if the restriction of the metric g to the subspace $P \subset T_pM$ is nondegenerate, a condition that came for free when g was positive, but in the indefinite case it imposes a restriction on the set of subspaces $P \subset T_pM$ along which K_S can be defined. This should not be surprising however, as it is another symptom of the fact that arbitrary submanifolds are not always *pseudo-Riemannian* submanifolds, and if we want to interpret $K_S(P)$ as the Gaussian curvature at p of a submanifold $\Sigma_P \subset M$ tangent to P, we certainly want g to be nondegenerate on that submanifold. The second caveat is that it is not obvious at this stage why the right hand side of (31.4) should be independent of the choice of basis $X, Y \in P$. If $g|_P$ is positive, then $\Sigma_P \subset M$ is a Riemannian submanifold and we can again recognize (31.4) as a formula for the Gaussian curvature of Σ_P, which proves independence of the choices. If $g|_P$ is nondegenerate but not positive, then a similar argument will work, but we must first discuss how to define K_G on a surface with an indefinite metric.

For a pseudo-Riemannian surface (Σ, g), g is either positive or negative or has signature $(1, 1)$. The negative case is essentially the same as the positive case: it means simply that $(\Sigma, -g)$ is a Riemannian manifold, and since the Levi-Cività connections and volume forms with respect to g and $-g$ are the same, their Riemann tensors and Gaussian curvatures are also the same.

The more interesting case is where (Σ, g) is a *Lorentzian* 2-manifold, with signature $(1, 1)$. The formula $R(X, Y)Z = -K_G \, d\mathrm{vol}(X, Y)JZ$ in the Riemannian case was based on the fact that at each point $p \in \Sigma$, the space of antisymmetric bilinear forms on $T_p\Sigma$ is 1-dimensional, and so is the space of antisymmetric linear maps $T_p\Sigma \to T_p\Sigma$, for which we chose the 90-degree counterclockwise fiberwise rotation $J : T\Sigma \to T\Sigma$ as a canonical basis at each point. If we instead have an indefinite metric $\langle\ ,\ \rangle$ on $T\Sigma$, then antisymmetry of a map $T_p\Sigma \to T_p\Sigma$ with respect to this metric means something qualitatively different, and we will have to choose a new generator $J : T\Sigma \to T\Sigma$. In the Riemannian case, our generator was characterized by the property that for any unit vector $X \in T_p\Sigma$, (X, JX) forms a positively-oriented orthonormal basis. This turns out to be a reasonable condition to generalize.

Lemma 31.17. *Assume V is an oriented 2-dimensional vector space with a nondegenerate symmetric bilinear form $\langle\ ,\ \rangle$ of signature $(1, 1)$. Then there exists a unique linear map $J : V \to V$ with the property that for every positively-oriented basis (v, w) of V that is orthonormal with respect to $\langle\ ,\ \rangle$,*[77]

$$Jv = w \qquad \text{and} \qquad Jw = v.$$

Moreover, J is antisymmetric with respect to $\langle\ ,\ \rangle$, and also satisfies $J^2 = \mathbb{1}$ and $\langle Jv, Jw\rangle = -\langle v, w\rangle$ for all $v, w \in V$.

Proof. Choosing a positively-oriented orthonormal basis of V allows us to assume without loss of generality that $V = \mathbb{R}^2$ with the standard Minkowski inner product $\langle v, w\rangle = v^T\eta w$ determined by the matrix $\eta := \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$. The standard basis is then positively oriented and orthonormal, so

[77]Note that when we talk about an orthonormal basis (v, w) of V with respect to an indefinite inner product of signature $(1, 1)$, the order matters: our convention is that $\langle v, v\rangle = 1$ and $\langle w, w\rangle = -1$.

if a transformation $J : \mathbb{R}^2 \to \mathbb{R}^2$ with the stated properties exists, then it is clearly unique: it is given by the matrix

$$\mathbf{J}_0 := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{31.5}$$

which reflects $\mathbb{R}^2$ orthogonally across the diagonal subspace $\{(x,x)\} \subset \mathbb{R}^2$. To see that this does the trick, observe that every positively-oriented orthonormal basis of $(\mathbb{R}^2, \langle\ ,\ \rangle)$ is of the form (v,w) with

$$v = \begin{pmatrix} x \\ y \end{pmatrix}, \qquad w = \begin{pmatrix} y \\ x \end{pmatrix} \qquad \text{such that} \qquad x^2 - y^2 = 1,$$

so $J := \mathbf{J}_0$ does indeed send $v \mapsto w$ and $w \mapsto v$. This proves that $J : V \to V$ is defined independently of the choice of positively-oriented orthonormal basis on V. The relations $J^2 = \mathbb{1}$ and $\langle Jv, Jw\rangle = -\langle v, w\rangle$ are now quick computations, and antisymmetry follows: $\langle v, Jw\rangle = -\langle Jv, J^2 w\rangle = -\langle Jv, w\rangle$. □

REMARK 31.18. The fact that the reflection $J : V \to V$ in Lemma 31.17 is represented by the same matrix in any positive orthonormal basis reveals that the subspaces $\ell_\pm \subset V$ spanned by the vectors represented by $(1, \pm 1)$ in coordinates do not actually depend on the chosen coordinates. Their invariant description is as follows: $\ell_+ \cup \ell_-$ is the set of all *light-like* vectors $v \in V$, i.e. those which satisfy $\langle v, v\rangle = 0$. One can use the orientation on V to distinguish between ℓ_+ and ℓ_-, and then define J as the reflection along ℓ_- about ℓ_+.

REMARK 31.19. On $\mathbb{R}^2$ with the Minkowski inner product, the group of orientation-preserving isometries is the 1-dimensional abelian group

$$\mathrm{SO}(1,1) := \mathrm{O}(1,1) \cap \mathrm{SL}(2,\mathbb{R}) = \left\{ \pm e^{\theta \mathbf{J}_0} = \pm \begin{pmatrix} \cosh\theta & \sinh\theta \\ \sinh\theta & \cosh\theta \end{pmatrix} \;\middle|\; \theta \in \mathbb{R} \right\} \subset \mathrm{GL}(2,\mathbb{R}),$$

where $\mathbf{J}_0$ is the matrix defined in (31.5). From this perspective, $\mathbf{J}_0$ can be regarded as a canonical generator of the Lie algebra $\mathfrak{so}(1,1)$, and the fact that $\mathrm{SO}(1,1)$ is abelian is what makes it possible to define $J : V \to V$ in Lemma 31.17 via this matrix without it depending on the choice of positively-oriented orthonormal basis.

By Lemma 31.17, we can define on our oriented Lorentzian surface (Σ, g) a canonical antisymmetric bundle map

$$J : T\Sigma \to T\Sigma$$

such that whenever (X_1, X_2) is a positively-oriented orthonormal basis of some tangent space $T_p\Sigma$,

$$X_2 = JX_1 \qquad \text{and} \qquad X_1 = JX_2.$$

Every antisymmetric bundle map $T\Sigma \to T\Sigma$ is then of the form fJ for some function $f : \Sigma \to \mathbb{R}$, so the symmetries of the Riemann tensor imply

$$R(X,Y)Z = -K_G \, d\mathrm{vol}(X,Y) JZ \tag{31.6}$$

for a uniquely determined function $K_G : \Sigma \to \mathbb{R}$, which will be known henceforth as the Gaussian curvature of (Σ, g). We can compute K_G from the Riemann tensor in much the same way as in the Riemannian case: since J is antisymmetric, one checks by plugging in a positively-oriented orthonormal basis that

$$d\mathrm{vol}(X,Y) = \langle X, JY\rangle$$

for all $X, Y \in T_p\Sigma$, thus

$$\mathrm{Riem}(X,X,Y,Y) = \langle X, R(X,Y)Y\rangle = -K_G(p) \cdot \left| d\mathrm{vol}(X,Y)^2 \right|^2$$

and if X and Y are chosen to be linearly independent, we can write

$$K_G(p) = -\frac{\mathrm{Riem}(X, X, Y, Y)}{|d\mathrm{vol}(X, Y)|^2}. \tag{31.7}$$

If you look at (27.7), you'll notice that this formula has an extra minus sign compared with the Riemannian case, but this is more sensible than you might think. Recall from Exercise 18.30 how the canonical volume form is computed in the indefinite case: when $\langle\ ,\ \rangle$ has signature $(1, 1)$, the symmetric matrix

$$\mathbf{g} := \begin{pmatrix} \langle X, X\rangle & \langle X, Y\rangle \\ \langle Y, X\rangle & \langle Y, Y\rangle \end{pmatrix}$$

has one positive and one negative eigenvalue, so its determinant is *negative*, implying

$$d\mathrm{vol}(X, Y) = \sqrt{-\det \mathbf{g}} = \sqrt{-\langle X, X\rangle\langle Y, Y\rangle + \langle X, Y\rangle^2}.$$

The extra minus sign turns (31.7) into

$$K_G(p) = \frac{\mathrm{Riem}(X, X, Y, Y)}{\langle X, X\rangle\langle Y, Y\rangle - \langle X, Y\rangle^2}, \tag{31.8}$$

a formula that is equally valid in the Riemannian and Lorentzian cases. In local coordinates, (27.8) now generalizes in the form

$$K_G = \frac{R_{1122}}{g_{11}g_{22} - (g_{12})^2}.$$

With this definition of K_G in place, our original definition of the sectional curvature $K_S(P)$ also makes sense in pseudo-Riemannian manifolds: it is the Gaussian curvature at p of a particular 2-dimensional pseudo-Riemannian submanifold $\Sigma_P \subset M$ tangent to P at p, and (31.4) gives a correct formula for computing it, due to (31.8).

In order for K_G to be truly useful, we'd like to be able to generalize the results of §27.3 and have a convenient way of computing it for surfaces embedded as pseudo-Riemannian submanifolds in $\mathbb{R}^3$. In order for $\Sigma \subset \mathbb{R}^3$ to have signature $(1, 1)$, we need to assume $(\mathbb{R}^3, g)$ has signature either $(2, 1)$ or $(1, 2)$, so let's take the metric to be plus or minus the Minkowski metric,

$$g := \pm g_M, \qquad g_M = -dx^2 + dy^2 + dz^2.$$

If we take the plus sign in this definition, then $(\mathbb{R}^3, g)$ has signature $(2, 1)$, and g is thus positive on the normal bundle $T\Sigma^\perp \subset T\mathbb{R}^3|_\Sigma$ of any pseudo-Riemannian surface $\Sigma \subset \mathbb{R}^3$ with signature $(1, 1)$. Taking the minus sign makes the signature of $(\mathbb{R}^3, g)$ into $(1, 2)$, so g is then negative on $T\Sigma^\perp$. A unit normal vector field $\nu \in \Gamma(T\Sigma^\perp)$ can thus be said to satisfy

$$\langle \nu, \nu\rangle = \pm 1,$$

so that it takes values in the connected hyperboloid

$$H^2_L := \left\{X \in \mathbb{R}^3 \;\middle|\; \langle X, X\rangle = \pm 1\right\} = \{g_M(X, X) = 1\} = \left\{y^2 + z^2 - x^2 = 1\right\} \subset \mathbb{R}^3,$$

defining a Gauss map

$$\nu : \Sigma \to H^2_L,$$

whose derivative at any point $p \in \Sigma$ again defines a linear self-map of $T_p\Sigma$:

$$T_p\nu : T_p\Sigma \to T_{\nu(p)}H^2_L = \nu(p)^\perp = T_p\Sigma.$$

Proposition 31.20. *For a 2-dimensional pseudo-Riemannian submanifold Σ in $(\mathbb{R}^3, \pm g_M)$, the Gaussian curvature $K_G : \Sigma \to \mathbb{R}$ is related to the Gauss map $\nu : \Sigma \to \mathbb{R}$ by*

$$K_G(p) = \pm \det(T_p\nu).$$

Proof. The arguments of §28.1 apply equally well to this situation up to and including Equation 28.3. The rest of the proof requires modification, because while $\nabla\nu(p) : T_p\Sigma \to T_p\Sigma$ is self-adjoint with respect to the bundle metric $\langle\ ,\ \rangle$ on $T\Sigma$, that does not generally imply that it is diagonalizable if the metric is indefinite. The main danger here is that there could be a light-like eigenspace, which is its own orthogonal complement and thus does not imply the existence of any complementary eigenspace. (If you try writing down simple examples of symmetric operators on Minkowski $\mathbb{R}^2$, you will find that this really can happen.)

As luck would have it, the orthonormal basis of eigenvectors we used for proving Theorem 27.17 in §28.1 was convenient, but not truly necessary. We could have argued more generally as follows. Choose any positively-oriented orthonormal basis (X_1, X_2) of $T_p\Sigma$; recall that in the Lorentzian case, this means

$$\langle X_1, X_1\rangle = 1, \quad \langle X_2, X_2\rangle = -1, \quad \langle X_1, X_2\rangle = 0 \qquad \text{and} \qquad d\mathrm{vol}(X_1, X_2) = 1, \tag{31.9}$$

and the defining property of the reflection $J : T\Sigma \to T\Sigma$ then implies

$$X_2 = JX_1, \qquad \text{and} \qquad X_1 = JX_2. \tag{31.10}$$

In this basis, the transformation $\nabla\nu(p) : T_p\Sigma \to T_p\Sigma$ is represented by a matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$, so applying the Gauss equation as in §28.1 gives

$$\begin{aligned}
\langle V, \widehat{R}(X_1, X_2)Z\rangle &= \langle \mathrm{II}(V, X_1), \mathrm{II}(X_2, Z)\rangle - \langle \mathrm{II}(V, X_2), \mathrm{II}(X_1, Z)\rangle \\
&= \big\langle \langle V, \nabla\nu(p)X_1\rangle \nu(p), \langle Z, \nabla\nu(p)X_2\rangle \nu(p)\big\rangle - \big\langle \langle V, \nabla\nu(p)X_2\rangle \nu(p), \langle Z, \nabla\nu(p)X_1\rangle \nu(p)\big\rangle \\
&= \pm \left(\langle V, \nabla\nu(p)X_1\rangle\langle Z, \nabla\nu(p)X_2\rangle - \langle V, \nabla\nu(p)X_2\rangle\langle Z, \nabla\nu(p)X_1\rangle\right) \\
&= \pm \left(\langle V, aX_1 + cX_2\rangle\langle Z, bX_1 + dX_2\rangle - \langle V, bX_1 + dX_2\rangle\langle Z, aX_1 + cX_2\rangle\right) \\
&= \langle V, \pm \left(\langle Z, bX_1 + dX_2\rangle(aX_1 + cX_2) - \langle Z, aX_1 + cX_2\rangle(bX_1 + dX_2)\right)\rangle,
\end{aligned}$$

implying

$$\begin{aligned}
\widehat{R}(X_1, X_2)Z &= \pm\langle Z, bX_1 + dX_2\rangle(aX_1 + cX_2) \mp \langle Z, aX_1 + cX_2\rangle(bX_1 + dX_2) \\
&= \pm(ad - bc)\left(\langle Z, X_2\rangle X_1 - \langle Z, X_1\rangle X_2\right).
\end{aligned}$$

To simplify the last expression in parentheses, recall that J preserves $\langle\ ,\ \rangle$ with an extra sign, so using (31.9) and (31.10),

$$\langle Z, X_2\rangle X_1 - \langle Z, X_1\rangle X_2 = -\langle JZ, X_1\rangle X_1 + \langle JZ, X_2\rangle X_2 = -JZ,$$

and our computation thus becomes

$$\widehat{R}(X_1, X_2)Z = \mp \det(\nabla\nu(p))JZ = \mp \det(\nabla\nu(p))\, d\mathrm{vol}(X_1, X_2)JZ = -K_G\, d\mathrm{vol}(X_1, X_2)JZ.$$

□

Example 31.21. The hyperboloid $H_L^2 = \{y^2 + z^2 - x^2 = 1\}$ has the identity map $H_L^2 \to H_L^2$ as its Gauss map, so it has $K_G \equiv 1$ with the metric g_M, and $K_G \equiv -1$ if the metric is $-g_M$.

Exercise 31.22. In $(n+1)$-dimensional Minkowski space $(\mathbb{R}^{n+1}, g_M)$ with coordinates $(\tau, x^1, \ldots, x^n)$ and the metric $g_M := -d\tau^2 + (dx^1)^2 + \ldots + (dx^n)^2$, the connected n-dimensional hyperboloid

$$H_L^n := \left\{X \in \mathbb{R}^{n+1} \;\middle|\; \langle X, X\rangle = 1\right\} = \left\{(\tau, \mathbf{x}) \in \mathbb{R} \times \mathbb{R}^n \;\middle|\; |\mathbf{x}|^2 - \tau^2 = 1\right\}$$

is a Lorentzian submanifold, i.e. its signature is $(n-1, 1)$.

(a) Show that for $p \in H^n_L$ and $\mathbf{v} \in T_p H^n_L = p^\perp \subset \mathbb{R}^{n+1}$, the following curves $\gamma : \mathbb{R} \to \mathbb{R}^{n+1}$ are all examples of geodesics in H^N_L:

$$\gamma(t) = \begin{cases} (\cosh t)p + (\sinh t)\mathbf{v} & \text{if } \langle \mathbf{v}, \mathbf{v} \rangle = -1, \\ p + t\mathbf{v} & \text{if } \langle \mathbf{v}, \mathbf{v} \rangle = 0, \\ (\cos t)p + (\sin t)\mathbf{v} & \text{if } \langle \mathbf{v}, \mathbf{v} \rangle = 1. \end{cases}$$

Show moreover that every geodesic in H^n_L is a parametrization of one of these curves.

(b) Show that the isometry group of H^n_L is as large as possible (in the sense of Theorem 24.3). *Hint: Use the Lorentz group* $\mathrm{O}(n, 1)$.

(c) Find an isometrically embedded copy of H^2_L in H^n_L, and deduce via Example 31.21 and the abundance of isometries that H^n_L has constant sectional curvature 1.

Remark 31.23. If we instead regard the hyperboloid H^n_L in Exercise 31.22 as a submanifold of signature $(1, n-1)$ in $(\mathbb{R}^{n+1}, -g_M)$, then its constant sectional curvature becomes -1.

We can now adapt Theorem 31.15 to the study of time-like geodesics in Lorentzian manifolds. Here it is necessary to make a choice as to whether "Lorentzian" means the signature is $(n-1, 1)$ or $(1, n-1)$; I have used the latter convention in Remarks 22.12 and 23.8, and will thus stick with it here, even though the other convention seems to be slightly more popular in the literature on mathematical relativity. In a manifold (M, g) of signature $(1, n-1)$, we say a tangent vector $X \in TM$ is *time-like* if $\langle X, X \rangle$ is positive and *space-like* if it is negative. If we were instead using signature $(n-1, 1)$, we would have to define the terms "time-like" and "space-like" the other way around; the rule of thumb in order for these terms to be meaningful is that there should always be exactly one time dimension, though we can consider arbitrarily many spatial dimensions.

The following detail requires slightly more care than in the Riemannian case: if $\{\gamma_s \in \mathcal{P}\}_{s \in (-\epsilon, \epsilon)}$ is an arbitrary smooth family of time-like paths with $\gamma := \gamma_0$ a geodesic, then for the vector field $\eta := \partial_s \gamma_s|_{s=0} \in \Gamma(\gamma^* TM)$ along γ, the integral

$$\int_a^b \langle \eta(t), \eta(t) \rangle \, dt$$

can no longer be viewed as the square of an L^2-norm, and its integrand might be sometimes positive and sometimes negative. This is where it becomes important to assume there is only one time dimension, because in this case, reparametrizing to achieve constant speed removes the uncertainty:

Lemma 31.24. *On a pseudo-Riemannian manifold (M, g), suppose $\{\gamma_s \in \mathcal{P}\}_{s \in (-\epsilon, \epsilon)}$ is a smooth family of paths with fixed end points $\gamma_s(a) = p$, $\gamma_s(b) = q$ such that $\gamma := \gamma_0$ is a geodesic segment and for each s, the "speed squared" $\langle \dot{\gamma}_s(t), \dot{\gamma}_s(t) \rangle \in \mathbb{R}$ is nonzero and independent of t. Then the vector field $\eta := \partial_s \gamma_s|_{s=0}$ along γ and its covariant derivatives $\nabla^k_t \eta \in \Gamma(\gamma^* TM)$ of arbitrary orders $k \geqslant 0$ are everywhere orthogonal to $\dot{\gamma}$.*

Proof. By assumption $\partial_t \langle \dot{\gamma}_s(t), \dot{\gamma}_s(t) \rangle$ vanishes for all s and t, thus

$$\begin{aligned} 0 &= \partial_s \partial_t \langle \dot{\gamma}_s(t), \dot{\gamma}_s(t) \rangle|_{s=0} = \partial_t \partial_s \langle \dot{\gamma}_s(t), \dot{\gamma}_s(t) \rangle|_{s=0} = 2\partial_t \langle \nabla_s \partial_t \gamma_s(t)|_{s=0}, \dot{\gamma}(t) \rangle \\ &= 2\partial_t \langle \nabla_t \eta(t), \dot{\gamma}(t) \rangle = 2 \langle \nabla^2_t \eta(t), \dot{\gamma}(t) \rangle, \end{aligned}$$

where in the last line the term involving $\nabla_t \dot{\gamma}(t)$ does not appear since γ is a geodesic. This proves that $\nabla^2_t \eta \in \Gamma(\gamma^* TM)$ is everywhere orthogonal to $\dot{\gamma}$. Since $\langle \dot{\gamma}, \dot{\gamma} \rangle \neq 0$ by assumption, the image of γ is a pseudo-Riemannian submanifold of (M, g), and it follows that $T_{\gamma(t)} M = \mathbb{R}\dot{\gamma}(t) \oplus \dot{\gamma}(t)^\perp$ for each t, so that $\eta \in \Gamma(\gamma^* TM)$ splits uniquely into a sum

$$\eta(t) = f(t)\dot{\gamma}(t) + \eta^\perp(t)$$

with a function $f : [a,b] \to \mathbb{R}$ and a section $\eta^\perp \in \Gamma(\gamma^* TM)$ that is everywhere orthogonal to $\dot\gamma$. Since $\dot\gamma$ is parallel along γ, differentiating the relation $\langle \dot\gamma, \eta^\perp \rangle \equiv 0$ gives $\langle \dot\gamma, \nabla_t \eta^\perp \rangle \equiv 0$, and applying the Leibniz rule to $f\dot\gamma$, we obtain

$$\nabla_t \eta(t) = \dot f(t)\dot\gamma(t) + \nabla_t \eta^\perp(t)$$

with $\nabla_t \eta^\perp$ also everywhere orthogonal to $\dot\gamma$. Repeating this same argument one more time gives

$$\nabla_t^2 \eta(t) = \ddot f(t)\dot\gamma(t) + \nabla_t^2 \eta^\perp(t)$$

with $\nabla_t^2 \eta^\perp$ also everywhere orthogonal to $\dot\gamma$, and since we already know the same holds for $\nabla_t^2 \eta$, it follows that $\ddot f \equiv 0$. The function $f : [a,b] \to \mathbb{R}$ is therefore affine, and since $\eta(a) = 0$ and $\eta(b) = 0$, the condition $f(a) = f(b) = 0$ then implies $f \equiv 0$, so η is everywhere orthogonal to $\dot\gamma$. The same now follows for all derivatives $\nabla_t^k \eta$ by repeatedly differentiating the relation $\langle \eta, \dot\gamma \rangle \equiv 0$. □

Let us now assume we are in the setting of Theorem 31.15, but with the following modifications: (M, g) has signature $(1, n-1)$ and the geodesic $\gamma : [a,b] \to M$ is time-like. The paths in the family $\gamma_s : [a,b] \to M$ with $\gamma_0 = \gamma$ and $\eta := \partial_s \gamma_s|_{s=0}$ can then be assumed to be all time-like after restricting s to a small enough neighborhood of 0, and we are also free to reparametrize them so that they all have constant speed. Lemma 31.24 then implies $\langle \eta(t), \dot\gamma(t) \rangle = 0$ for all t, making η a section of the orthogonal complement bundle along the image of γ. Since γ satisfies $\langle \dot\gamma, \dot\gamma \rangle > 0$ and the signature of (M, g) is $(1, n-1)$, the restriction of g to this complementary subbundle must be strictly *negative*, implying

$$-\langle \nabla_t^2 \eta, \eta \rangle_{L^2} = \langle \nabla_t \eta, \nabla_t \eta \rangle_{L^2} \leqslant 0,$$

with strict inequality unless η vanishes. This is a significant change compared with the proof of Lemma 31.14, but the second term in $\langle \nabla^2 E(\gamma)\eta, \eta \rangle_{L^2}$ undergoes a similar change: whenever $\eta(t)$ is nonzero, the restriction of the metric to the space spanned by $X := \eta(t)$ and $Y := \dot\gamma(t)$ has signature $(1, 1)$, making the determinant

$$\det \begin{pmatrix} \langle X, X \rangle & \langle X, Y \rangle \\ \langle Y, X \rangle & \langle Y, Y \rangle \end{pmatrix}$$

negative, so that this determinant is not $\mathrm{Area}(X, Y)^2$, but instead $-\mathrm{Area}(X, Y)^2 \leqslant 0$. Using (31.4), we thus have

$$-\mathrm{Riem}(\eta(t), \eta(t), \dot\gamma(t), \dot\gamma(t)) = K_S(P_t) \cdot \mathrm{Area}(\eta(t), \dot\gamma(t))^2 \leqslant 0,$$

assuming $K_S \leqslant 0$. The result is that in contrast to Lemma 31.14, the operator $\nabla^2 E(\gamma)$ is now *negative*-definite, and Theorem 31.15 becomes the following statement about the *proper time* (cf. Remark 23.8) of a time-like geodesic:

THEOREM 31.25. *Suppose (M, g) is a pseudo-Riemannian manifold of signature $(1, n-1)$ with nonpositive sectional curvature and $\gamma : [a,b] \to M$ is a time-like geodesic connecting $\gamma(a) = p$ to $\gamma(b) = q$. Then for any smooth 1-parameter family of paths $\gamma_s : [a,b] \to M$ with $\gamma_s(a) = p$, $\gamma_s(b) = q$ and $\gamma_0 \equiv \gamma$ such that $\partial_s \gamma_s|_{s=0}$ is not everywhere tangent to $\dot\gamma$, there is a number $\epsilon > 0$ such that:*

(1) γ is the only geodesic among the paths γ_s for $s \in (-\epsilon, \epsilon)$;
(2) For all paths γ_s with $s \in (-\epsilon, \epsilon)$ and $s \neq 0$, the proper time $\tau(\gamma_s)$ satisfies

$$\tau(\gamma_s) < \tau(\gamma).$$

□

Remark 31.26. In the literature, results similar to Theorem 31.25 are often stated with signature $(n-1,1)$ instead of $(1,n-1)$, in which case the definitions of the terms "time-like" and "space-like" must be interchanged. The argument via Lemma 31.24 above then leads to the conclusion that $\langle\nabla_t\eta,\nabla_t\eta\rangle_{L^2}\geqslant 0$, so that the hypothesis necessary for the proof to work becomes

$$K_S\geqslant 0$$

instead of $K_S\leqslant 0$. To get an idea of why this makes sense, take another look at the 2-dimensional case of Exercise 31.22. The hyperboloid H^2_L has constant positive curvature in $(\mathbb{R}^3,g_M)$, and since signature $(1,1)$ can be interpreted as either $(n-1,1)$ or $(1,n-1)$, we have some freedom here as to which curves we choose to call time-like vs. space-like. If we view the signature as $(n-1,1)$, then a tangent vector X is considered time-like if $\langle X,X\rangle<0$, so the time-like geodesics have the form $\gamma(t)=(\cosh t)p+(\sinh t)\mathbf{v}$, and there is indeed only one connecting any given pair of points, consistent with the conclusions of Theorem 31.25. But if the signature is instead viewed as $(1,n-1)$, then time-like means $\langle X,X\rangle>0$, and the time-like geodesics are all *periodic* curves of the form $(\cos t)p+(\sin t)\mathbf{v}$. Note that we have not changed the metric, so the curvature is still $+1$, making this a situation in which Theorem 31.25 simply does not apply, and we can see this explicitly since the geodesic segments from a point p to itself are clearly not isolated.[78] In order to apply the literal statement of Theorem 31.25 to this example, one should follow the advice of Remark 31.23 and view it as a submanifold of $(\mathbb{R}^3,-g_M)$ with signature $(1,n-1)$, so that its curvature becomes -1, and the time-like geodesics are again the paths $\gamma(t)=(\cosh t)p+(\sinh t)\mathbf{v}$.

I have one final remark about sectional curvature in the indefinite case. On any manifold M, the set of all 2-dimensional subspaces $P\subset T_pM$ at points $p\in M$ can be given the structure of a smooth manifold,

$$\mathrm{Gr}_2(TM):=\left\{P\subset T_pM\;\middle|\;p\in M,\ \dim P=2\right\},$$

known as a *Grassmannian.* This manifold is compact if M is, so if (M,g) is a compact Riemannian manifold, its sectional curvature can be viewed as a smooth function on a compact manifold,

$$K_S:\mathrm{Gr}_2(TM)\to\mathbb{R},$$

and is therefore necessarily bounded. In the indefinite case, this is no longer true, because K_S is not defined on all of $\mathrm{Gr}_2(TM)$, but only on the open subset

$$\mathrm{Gr}_2^*(TM,g):=\left\{P\in\mathrm{Gr}_2(TM)\;\middle|\;g|_P\text{ is nondegenerate}\right\},$$

which is always noncompact when $\dim M\geqslant 3$ unless g is positive or negative. It's not just that K_S *can* be unbounded in this case: by a theorem of Kulkarni and Nomizu, it *must* be unbounded, both above and below, outside of exceptional cases like the hyperboloid H^n_L in Exercise 31.22 for which K_S is constant. More details on this theorem can be founded e.g. in [**Bau06**].

[78] Let it be said that this would in any case be an extremely strange spacetime manifold to consider: all time-like geodesics being periodic means a universe in which time travel is not just possible, but mandatory!

Second semester (Differentialgeometrie II)

Prologue: Some terminology and notation

This is the second semester of a year-long course, but it should be possible to follow it without having learned everything that was covered in the first semester. Several things that appeared in the first course will be reviewed in this one, albeit quickly, and sometimes in more general or abstract settings. It is of course important to be clear on what will *not* be reviewed: I will assume in this course that you are familiar with the basic theory of smooth n-dimensional manifolds, including the notions of vector fields and the Lie bracket, tensors, differential forms, integration and Stokes' theorem. Aside from the obvious prerequisites in analysis, linear algebra and basic topology, all the knowledge that I will assume in the near future is contained in the notes from the first semester of this course. Later on, when we talk about Hodge theory, some knowledge of functional analysis will be useful, but it will be possible to take the relevant results as black boxes without losing the thread.

Let's make sure we are all clear on the meaning of certain terms and symbols. For smooth manifolds M and N, we write

$$C^\infty(M,N) := \{ f : M \to N \mid f \text{ is smooth} \},$$

for the space of smooth maps, and abbreviate the space of real-valued functions on M by

$$C^\infty(M) := C^\infty(M,\mathbb{R}).$$

The tangent space to M at a point $p \in M$ is denoted by T_pM, so the tangent bundle is $TM = \bigcup_{p\in M} T_pM$. Differentiating a smooth map $f : M \to N$ defines its **tangent map**

$$Tf : TM \to TN, \qquad T_pf := Tf|_{T_pM} \in \operatorname{Hom}(T_pM, T_{f(p)}N) \text{ for } p \in M,$$

which is also sometimes denoted by

$$f_* : TM \to TN$$

and called the **pushforward** operation on tangent vectors. If the target of $f : M \to V$ is a vector space V, then the tangent map can be expressed more succinctly via the **differential**

$$df : TM \to V, \qquad d_pf := df|_{T_pM} \in \operatorname{Hom}(T_pM, V) \text{ for } p \in M.$$

When $V = \mathbb{R}$, $d_pf : T_pM \to \mathbb{R}$ defines an element of the **cotangent space** $T_p^*M = \operatorname{Hom}(T_pM,\mathbb{R})$, and the union $T^*M = \bigcup_{p\in M} T_p^*M$ is the **cotangent bundle** of M.

For integers $k, \ell \geqslant 0$, the space of smooth **tensor fields of type** (k,ℓ), also known as tensors that are **contravariant** of rank k and **covariant** of rank ℓ, is denoted by

$$\Gamma(T_\ell^k M) := \{\text{smooth tensor fields of type } (k,\ell) \text{ on } M\}.$$

An element $S \in \Gamma(T_\ell^k M)$ associates to each point $p \in M$ a multilinear map

$$S_p : \underbrace{T_p^*M \times \ldots \times T_p^*M}_{k} \times \underbrace{T_pM \times \ldots \times T_pM}_{\ell} \to \mathbb{R},$$

where by convention one defines S_p in the case $k = \ell = 0$ to be simply a real number, so that $\Gamma(T^0_0 M) = \mathbb{R}$. Note that for $S \in \Gamma(T^1_\ell M)$, the canonical isomorphism of $T_p M$ with the dual of its dual space implies that S_p can be identified canonically with a multilinear map

$$S_p : \underbrace{T_p M \times \ldots \times T_p M}_{\ell} \to T_p M,$$

and it is usually more convenient to think of type $(1, \ell)$ tensor fields in this way. In particular, this identifies $\Gamma(T^1_0 M)$ with the space of **smooth vector fields**

$$\mathfrak{X}(M) = \Gamma(T^0_1 M) = \left\{ X \in C^\infty(M, TM) \;\middle|\; X(p) \in T_p M \text{ for all } p \in M \right\}.$$

We recall that $\mathfrak{X}(M)$ also has a natural isomorphism with the space of all **derivations** on the ring $C^\infty(M)$, identifying each vector field $X \in \mathfrak{X}(M)$ with the **Lie derivative** operator

$$\mathcal{L}_X : C^\infty(M) \to C^\infty(M) : f \mapsto \mathcal{L}_X f := df(X).$$

Many books use these two perspectives on vector fields interchangeably and thus write $Xf := \mathcal{L}_X f \in C^\infty(M)$ for $X \in \mathfrak{X}(M)$ and $f \in C^\infty(M)$; my personal preference however is to keep the distinction between vector fields and derivations by always writing $\mathcal{L}_X$ for the latter. The major exception is the **coordinate vector fields** associated to a chart $x = (x^1, \ldots, x^n) : \mathcal{U} \to \mathbb{R}^n$ on some open subset $\mathcal{U} \subset M$, for which it is standard to use the derivations defined by the resulting partial derivative operators as notation for vector fields,

$$\partial_j := \frac{\partial}{\partial x^j} \in \mathfrak{X}(\mathcal{U}), \qquad j = 1, \ldots, n.$$

We will typically denote the flow of a vector field $X \in \mathfrak{X}(M)$ over some time $t \in \mathbb{R}$ by φ^t_X, hence

$$\varphi^{t_0}_X(p) := \gamma(t_0) \quad \text{ where } \gamma(t) \in M \text{ satisfies } \dot{\gamma}(t) = X(\gamma(t)) \text{ and } \gamma(0) = p.$$

The domain of $(t, p) \mapsto \varphi^t_X(p)$ is in general an open subset of $\mathbb{R} \times M$ containing $\{0\} \times M$, though various conditions will sometimes imply that it is all of $\mathbb{R} \times M$, in which case we say that X is a **complete vector field**, or that it has a **global flow**. This is true in particular whenever M is a closed manifold, i.e. compact and with empty boundary.

Given a diffeomorphism $\varphi : M \to N$, there are natural **pushforward** maps φ_* and **pullback** maps φ^* between the respective spaces of tensor fields,

$$\varphi_* : \Gamma(T^k_\ell M) \to \Gamma(T^k_\ell N), \qquad \varphi^* : \Gamma(T^k_\ell N) \to \Gamma(T^k_\ell M),$$

all of which can be derived in natural ways from the pushforward $\varphi_* : T_p M \to T_{\varphi(p)} N$ of tangent vectors mentioned above, together with its dualization, the pullback of cotangent vectors

$$\varphi^* : T^*_{\varphi(p)} N \to T^*_p M, \qquad (\varphi^* \lambda)(X) := \lambda(\varphi_* X).$$

So for example, the pullback $\varphi^* X \in \mathfrak{X}(M)$ of a vector field $\mathfrak{X}(N)$ is determined by the relation

$$\varphi_*((\varphi^* X)(p)) = X(\varphi(p)), \qquad \text{for } p \in M,$$

and the pushforward $\varphi_* J \in \Gamma(T^1_1 N)$ of a type $(1,1)$ tensor field $J \in \Gamma(T^1_1 M)$ by

$$(\varphi_* J)_{\varphi(p)}(\lambda, \varphi_* X) = J_p(\varphi^* \lambda, X), \qquad \text{for } p \in M,\ X \in T_p M,\ \lambda \in T^*_{\varphi(p)} N.$$

If we view J_p instead as a linear map $T_p M \to T_p M$, the latter relation becomes

$$(\varphi_* J)_{\varphi(p)}(\varphi_* X) = \varphi_*(J_p(X)), \qquad \text{for } p \in M,\ X \in T_p M.$$

The **Lie derivative** of a tensor field $S \in \Gamma(T^k_\ell M)$ with respect to a vector field $X \in \mathfrak{X}(M)$ is the tensor field $\mathcal{L}_X S \in \Gamma(T^k_\ell M)$ defined by

$$\mathcal{L}_X S := \left. \frac{d}{dt} (\varphi^t_X)^* S \right|_{t=0}.$$

Recall that for fully covariant tensor fields $S \in \Gamma(T^0_\ell N)$, pullbacks $\varphi^* S \in \Gamma(T^0_\ell M)$ are well defined for all smooth maps $\varphi : M \to N$ and not just for diffeomorphisms, as the formula

$$(\varphi^* S)_p(X_1, \dots, X_\ell) = S_{\varphi(p)}(\varphi_* X_1, \dots, \varphi_* X_\ell), \qquad \text{for } p \in M,\ X_1, \dots, X_\ell \in T_p M$$

makes sense without any need to assume φ is invertible.

Finally, we denote by

$$\Omega^k(M) \subset \Gamma(T^0_k M)$$

the space of **differential k-forms** for each integer $k \geqslant 0$, i.e. the antisymmetric tensor fields of type $(0, k)$.

Wherever appropriate, we will use the Einstein summation convention, meaning that when a product contains a matching pair of upper and lower indices, then a summation over the possible values of that index is implied. For example, if $A^i{}_j$ denote the entries of an m-by-n matrix and $B^i{}_j$ are the entries of an n-by-p matrix, the standard definition of matrix multiplication can be written as

$$(AB)^i{}_k = A^i{}_j B^j{}_k := \sum_{j=1}^n A^i{}_j B^j{}_k.$$

Similarly, a very quick proof of the relation $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ takes the form

$$\operatorname{tr}(AB) = (AB)^i{}_i = A^i{}_j B^j{}_i = B^j{}_i A^i{}_j = B^i{}_j A^j{}_i = (BA)^i{}_i = \operatorname{tr}(BA).$$

Whenever two indices i, j take values in the same set, we can define the symbols

$$\delta^{ij} = \delta^i_j = \delta_{ij} := \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases}$$

known collectively as the **Kronecker delta**. So for instance, the statement that $A^i{}_j$ and $B^i{}_j$ are inverse matrices now becomes

$$A^i{}_j B^j{}_k = \delta^i_k.$$

Note that for the matrices A and B, the notation is designed to keep track of not only the distinction between upper and lower indices but also the order in which they appear, whereas for the Kronecker delta the order does not matter.

For a tensor field $S \in \Gamma(T^k_\ell M)$ in a local chart $x = (x^1, \dots, x^n) : \mathcal{U} \to \mathbb{R}^n$ on some open subset $\mathcal{U} \subset M$, we can use the coordinate vector fields $\partial_j \in \mathfrak{X}(\mathcal{U})$ and coordinate differentials $dx^j \in \Omega^1(\mathcal{U})$ for $j = 1, \dots, n$ to define the **component** functions $S^{i_1 \dots i_k}{}_{j_1 \dots j_\ell} : \mathcal{U} \to \mathbb{R}$ by

$$S^{i_1 \dots i_k}{}_{j_1 \dots j_\ell} := S(dx^{i_1}, \dots, dx^{i_k}, \partial_{j_1}, \dots, \partial_{j_\ell}),$$

and thus write S locally as

$$S = S^{i_1 \dots i_k}{}_{j_1 \dots j_\ell}\, \partial_{i_1} \otimes \dots \otimes \partial_{i_k} \otimes dx^{j_1} \otimes \dots \otimes dx^{j_\ell}, \qquad \text{on } \mathcal{U}.$$

32. Vector bundles and connections

In this lecture we give a quick review of the essentials on vector bundles and connections, with emphasis on aspects that will be needed in the next few lectures for proving results in Riemannian geometry. Some deeper aspects of the subject that were discussed last semester will be glossed over for now, as we plan to cover them later in the more general context of fiber bundles.

We will consider both real and complex vector bundles, so in order to allow both possibilities whenever possible, we will use the symbol

$$\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$$

to denote either of the two fields $\mathbb{R}$ or $\mathbb{C}$. We will generally not specify one or the other unless it is necessary in the given context. For each integer $m \geqslant 0$, the vector space $\mathbb{F}^m$ will be endowed with its standard topology and smooth structure, making it a smooth manifold of dimension m in the case $\mathbb{F} = \mathbb{R}$, or $2m$ in the case $\mathbb{F} = \mathbb{C}$.[79]

32.1. Topological vector bundles. In the topological category, a **vector bundle** (*Vektorbündel*) $\pi : E \to B$ of **rank** (*Rang*) $m \geqslant 0$ over the field $\mathbb{F}$ consists of two topological spaces E, B related by a continuous surjective map π, such that for each $p \in B$ the **fiber** (*Faser*)

$$E_p := \pi^{-1}(p)$$

is an m-dimensional vector space, and moreover, p has a neighborhood $\mathcal{U}_\alpha \subset B$ on which there exists a **local trivialization** (*lokale Trivialisierung*) $(\mathcal{U}_\alpha, \Phi_\alpha)$, meaning a homeomorphism

$$\Phi_\alpha : E|_{\mathcal{U}_\alpha} := \pi^{-1}(\mathcal{U}_\alpha) \to \mathcal{U}_\alpha \times \mathbb{F}^m$$

that sends E_q for each $q \in \mathcal{U}_\alpha$ to $\{q\} \times \mathbb{F}^m$ as a vector space isomorphism. The topological spaces B and E are called the **base** (*Basis*) and **total space** (*Totalraum*) respectively of the bundle. We call $\pi : E \to B$ a **real** vector bundle if $\mathbb{F} = \mathbb{R}$ and a **complex** bundle if $\mathbb{F} = \mathbb{C}$, and in either case, we call it a **line bundle** (*Geradenbündel*) if $\operatorname{rank}(E) = 1$.

A **vector bundle isomorphism** between two vector bundles $E, F \to B$ is a homeomorphism $\Psi : E \to F$ such that for every $p \in B$, Ψ maps E_p to F_p as a vector space isomorphism. We say that two vector bundles are **isomorphic** if there exists a bundle isomorphism between them, and the bundle $E \to B$ is called (globally) **trivial** if it is isomorphic to the product bundle

$$B \times \mathbb{F}^m \to B : (p, v) \mapsto p,$$

also known as the **trivial m-plane bundle**.

A **section** (*Schnitt*) of $\pi : E \to B$ is a continuous map

$$s : B \to E \qquad \text{such that} \qquad \pi \circ s = \mathrm{Id}_B,$$

meaning $s(p)$ belongs to the specific fiber E_p for each $p \in B$. The set of all sections forms a vector space (infinite dimensional unless $\operatorname{rank}(E) = 0$), which is often denoted by

$$\Gamma(E) := \{\text{sections of } \pi : E \to B\},$$

though we will modify the meaning of this symbol in a moment when we discuss smoothness. Observe that for any subset $\mathcal{U} \subset B$, the **restriction** (*Einschränkung*)

$$E|_{\mathcal{U}} := \pi^{-1}(\mathcal{U}) \xrightarrow{\pi} \mathcal{U}$$

is also a (real or complex) vector bundle of rank m over the base $\mathcal{U}$, and every section $s \in \Gamma(E)$ has a restriction $s|_{\mathcal{U}} \in \Gamma(E|_{\mathcal{U}})$; a section of $E|_{\mathcal{U}}$ is also sometimes called a **section of E over $\mathcal{U}$**. A local trivialization $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^m$ provides a natural isomorphism of $\Gamma(E|_{\mathcal{U}_\alpha})$ with the space of continuous functions $\mathcal{U}_\alpha \to \mathbb{F}^m$: namely, we associate to any section $s \in \Gamma(E)$ its **local representative**

$$s_\alpha : \mathcal{U}_\alpha \to \mathbb{F}^m, \qquad \text{such that} \qquad \Phi_\alpha(s(p)) = (p, s_\alpha(p)) \text{ for } p \in \mathcal{U}_\alpha.$$

Equivalently, it is often convenient to describe a local trivialization in terms of a **frame** (*Rahmen*) for E over $\mathcal{U}_\alpha$, meaning a collection of sections $e_1, \dots, e_m \in \Gamma(E|_{\mathcal{U}_\alpha})$ whose values form a basis of E_p at every point $p \in \mathcal{U}_\alpha$. The natural frame corresponding to the trivialization Φ_α consists of the sections that are identified by the bijection $\Gamma(E|_{\mathcal{U}_\alpha}) \to C^\infty(\mathcal{U}_\alpha, \mathbb{F}^m) : s \mapsto s_\alpha$ with the constant functions whose values are the standard basis vetors $\mathbf{e}_1, \dots, \mathbf{e}_m \in \mathbb{F}^m$, i.e. we set

$$e_b(p) := \Phi_\alpha^{-1}(p, \mathbf{e}_b), \qquad \text{for} \qquad b = 1, \dots, m.$$

[79] This remark about the smooth structure of $\mathbb{F}^m$ is the reason why we are not allowing $\mathbb{F}$ to be a more general field such as $\mathbb{Q}$ or $\mathbb{Z}_p$, as would be allowed for instance in algebraic geometry. It is crucial for our purposes that we have a standard notion of differentiability for $\mathbb{F}$-valued functions on open subsets of $\mathbb{F}^m$.

We can then write $s \in \Gamma(E)$ over $\mathcal{U}_\alpha$ in the form[80]

$$s(p) = s^b(p) e_b(p) \tag{32.1}$$

for uniquely-determined **component** functions

$$s^1, \dots, s^m : \mathcal{U}_\alpha \to \mathbb{F},$$

which are actually just the individual coordinates of $s_\alpha = (s^1, \dots, s^m)$. We will sometimes also use this notation for individual vectors $v \in E_p$ in fibers over a single point $p \in \mathcal{U}_\alpha$, thus

$$\Phi_\alpha(v) = (p, v_\alpha), \qquad v = v^b e_b, \qquad v_\alpha = (v^1, \dots, v^m) \in \mathbb{F}^m.$$

Whenever two local trivializations $(\mathcal{U}_\alpha, \Phi_\alpha)$ and $(\mathcal{U}_\beta, \Phi_\beta)$ have overlapping domains $\mathcal{U}_\alpha \cap \mathcal{U}_\beta \neq \emptyset$, they can be related to each other by uniquely-defined **transition functions** (*Übergangsfunktionen*)

$$g_{\alpha\beta}, g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{GL}(m, \mathbb{F})$$

such that for $p \in \mathcal{U}_\alpha \cap \mathcal{U}_\beta$ and $v \in \mathbb{F}^m$,

$$\Phi_\beta \circ \Phi_\alpha^{-1}(p, v) = (p, g_{\beta\alpha}(p)v), \qquad \Phi_\alpha \circ \Phi_\beta^{-1}(p, v) = (p, g_{\alpha\beta}(p)v).$$

These functions are continuous, and the corresponding local representatives of a section $s \in \Gamma(E)$ are now related to each other on $\mathcal{U}_\alpha \cap \mathcal{U}_\beta$ by

$$s_\beta = g_{\beta\alpha} s_\alpha, \qquad s_\alpha = g_{\alpha\beta} s_\beta.$$

A collection of local trivializations $\{(\mathcal{U}_\alpha, \Phi_\alpha)\}_{\alpha \in I}$ such that $B = \bigcup_{\alpha \in I} \mathcal{U}_\alpha$ is sometimes called a **bundle atlas** for $\pi : E \to B$. Here I is an arbitrary set; we call it an "index set" since it is used only for bookkeeping purposes. Any bundle atlas $\{(\mathcal{U}_\alpha, \Phi_\alpha)\}_{\alpha \in I}$ determines a set of transition functions $g_{\alpha\beta}$ as described above, one for every pair $(\alpha, \beta) \in I \times I$.

The following exercise shows that the topology on the total space of a vector bundle does not really need to be specified so long as one has a bundle atlas with continuous transition functions (cf. Proposition 2.12 from last semester).

EXERCISE 32.1. Suppose B is a topological space, $\{E_p\}_{p \in B}$ is a collection of vector spaces of rank m over $\mathbb{F}$, and E denotes the set defined as the disjoint union of the vector spaces E_p for all $p \in B$, with $\pi : E \to B$ denoting the projection map that collapses each space E_p to the corresponding point p. Define the terms *local trivialization* $\Phi_\alpha : \pi^{-1}(\mathcal{U}_\alpha) \to \mathcal{U}_\alpha \times \mathbb{F}^m$, *transition function* and *bundle atlas* to mean the same thing as above, except without assuming that Φ_α is continuous (since we have not equipped E with a topology). Show that if $\{(\mathcal{U}_\alpha, \Phi_\alpha)\}_{\alpha \in I}$ is a bundle atlas whose transition functions are all continuous, then the set E admits a unique topology for which $\pi : E \to B$ is a vector bundle and the maps Φ_α are all homeomorphisms.

Many types of structure can be added to the fibers of a vector bundle by selecting a bundle atlas whose transition functions take values in a specified subgroup $G \subset \mathrm{GL}(m, \mathbb{F})$. A list of the most interesting possibilities was given in Lecture 18 last semester, and we will revisit this subject from a more abstract perspective when we discuss principal fiber bundles later in this course. For now, it will suffice to have the following two examples at our disposal:

EXAMPLE 32.2. An **orientation** (*Orientierung*) of a real vector bundle $E \to B$ associates to each point $p \in B$ an orientation of the vector space E_p such that these orientations depend continuously on p. Recall that an orientation of a real finite-dimensional vector space is by definition an equivalence class of ordered bases (which we call **positively oriented**) such that two ordered

[80]The right hand side of (32.1) is our first use this semester of the Einstein summation convention: the symbol "$\sum_{b=1}^m$" is implied but not written. We will use this convention routinely from now on without further comment, so you should always be on the lookout for matching pairs of upper and lower indices.

bases are considered equivalent whenever one can be deformed to the other through a continuous family of ordered bases. (Note that this notion does not make sense on complex vector spaces, since all ordered bases can be deformed to each other in the complex case.) We leave it as an exercise for the reader to come up with a sensible definition of the words "depend continuously on p", but the right notion can be deduced from the following statement: an orientation of $E \to B$ determines a distinguished class of so-called *oriented* local trivializations, for which the corresponding frames define positively-oriented bases of the relevant fibers. Any bundle atlas can then be reduced to one that consists only of oriented frames, and for any bundle atlas with this property, the transition functions all take values in the subgroup

$$\mathrm{GL}_+(m, \mathbb{R}) := \left\{ \mathbf{A} \in \mathrm{GL}(m, \mathbb{R}) \;\middle|\; \det(\mathbf{A}) > 0 \right\},$$

i.e. the group of orientation-preserving isomorphisms on $\mathbb{R}^m$. Conversely, any bundle atlas whose transition functions all take values in $\mathrm{GL}_+(m, \mathbb{R})$ determines an orientation of $E \to B$.

Example 32.3. A (positive) **bundle metric** on $E \to B$ associates to each point $p \in B$ an inner product $\langle\ ,\ \rangle$ on the vector space E_p such that these inner products depend continuously on p. By the Gram-Schmidt procedure, any local frame for E over a subset $\mathcal{U} \subset B$ can then be modified so that it gives an orthonormal basis of every fiber E_p for $p \in \mathcal{U}$, producing a so-called **orthonormal frame** over $\mathcal{U}$. For any bundle atlas consisting only of trivializations corresponding to orthonormal frames, the transition functions all take values in the orthogonal group $\mathrm{O}(m)$ or (in the complex case) the unitary group $\mathrm{U}(m)$. Conversely, any bundle atlas whose transition functions have this property determines a bundle metric that looks like the standard inner product of $\mathbb{F}^m$ in any local trivialization belonging to the bundle atlas.

Remark 32.4. More generally, one can consider bundle metrics that are not positive but have **indefinite** signature, meaning that instead of requiring $\langle v, v\rangle > 0$ for every nonzero $v \in E_p$, we require

for all nonzero $v \in E_p$, $\langle v, w\rangle \neq 0$ for some $w \in E_p$.

Symmetric bilinear pairings that satisfy this condition are called **nondegenerate**, and they are characterized algebraically via their **signature** (k, ℓ), where k and ℓ are the maximal dimensions of subspaces on which the pairing is positive- or negative-definite respectively. Indefinite bundle metrics of signature $(1, m-1)$ or $(m-1, 1)$ play an essential role in General Relativity, and a few of the important theorems about Riemannian manifolds are also valid in the indefinite case, though in this course we will more often restrict our attention to positive bundle metrics.

If the topology of the base B is sufficiently nice (e.g. it must be paracompact), then one can glue locally-constructed bundle metrics together via continuous partitions of unity to show that every vector bundle $E \to B$ admits a positive bundle metric. This works mainly due to the fact that the set of all inner products on a vector space is a convex set, making linear interpolation between different choices possible. It does not work in general for indefinite bundle metrics, and it also does not work for orientations, which indeed do not always exist. A real vector bundle is called **orientable** (*orientierbar*) if it admits an orientation.

Exercise 32.5. Show that for any bundle atlas $\{(\mathcal{U}_\alpha, \Phi_\alpha)\}_{\alpha \in I}$, the associated transition functions satisfy the following relations for any $\alpha, \beta, \gamma \in I$:

(1) $g_{\alpha\alpha} \equiv \mathbb{1}$;
(2) $g_{\alpha\beta} g_{\beta\gamma} \equiv g_{\alpha\gamma}$ on $\mathcal{U}_\alpha \cap \mathcal{U}_\beta \cap \mathcal{U}_\gamma$.

In particular, $g_{\alpha\beta} = g_{\beta\alpha}^{-1}$.

Remark 32.6. One can show that for any open cover $\{\mathcal{U}_\alpha\}_{\alpha \in I}$ of B together with a collection of functions $\{g_{\alpha\beta} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{GL}(m, \mathbb{F})\}_{(\alpha,\beta) \in I \times I}$ satisfying the relations in Exercise 32.5, there exists a vector bundle $\pi : E \to B$ with a bundle atlas having these as its transition functions. The relation $g_{\alpha\beta} g_{\beta\gamma} \equiv g_{\alpha\gamma}$ is known as the **cocycle condition**, and the following digression on the origin of this terminology is only for readers who know a bit about Čech cohomology (see e.g. [**Wen18**, Lecture 46]). Assume for simplicity that $\mathbb{F} = \mathbb{R}$ and the overlaps $\mathcal{U}_\alpha \cap \mathcal{U}_\beta$ are all connected sets, in which case the determinant of each of the functions $g_{\alpha\beta} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{GL}(m, \mathbb{R})$ is everywhere either positive or negative. Associate to each $(\alpha, \beta) \in I \times I$ the number $\sigma_{\alpha\beta} \in \mathbb{Z}_2$ defined by

$$\sigma_{\alpha\beta} := \begin{cases} 0 & \text{if } \det(g_{\alpha\beta}) > 0, \\ 1 & \text{if } \det(g_{\alpha\beta}) < 0, \end{cases}$$

so the cocycle condition implies

$$\sigma_{\alpha\beta} - \sigma_{\alpha\gamma} + \sigma_{\beta\gamma} = 0 \in \mathbb{Z}_2. \tag{32.2}$$

This relation can be interpreted literally as saying that the collection of numbers $\{\sigma_{\alpha\beta} \in \mathbb{Z}_2\}$ determine a 1-dimensional cocycle in the Čech cohomology $\check{H}^*(B; \mathbb{Z}_2)$ of B with $\mathbb{Z}_2$ coefficients. More precisely, $\check{H}^*(B; \mathbb{Z}_2)$ can be defined as the direct limit of a family of cohomologies of cochain complexes $\check{C}^*(\mathfrak{U}; \mathbb{Z}_2)$, each depending on a choice of open cover $\mathfrak{U} = \{\mathcal{U}_\alpha\}_{\alpha \in I}$ for B. For each integer $k \geqslant 0$, elements of the kth cochain group $\check{C}^k(\mathfrak{U}; \mathbb{Z}_2)$ are $\mathbb{Z}_2$-valued functions

$$(\alpha_1, \ldots, \alpha_{k+1}) \mapsto f(\alpha_1, \ldots, \alpha_{k+1}) \in \mathbb{Z}_2,$$

defined on the set of all ordered tuples $(\alpha_1, \ldots, \alpha_{k+1}) \in I^{k+1}$ such that $\mathcal{U}_{\alpha_1} \cap \ldots \cap \mathcal{U}_{\alpha_{k+1}} \neq \varnothing$, and the coboundary operator $\delta : \check{C}^k(\mathfrak{U}; \mathbb{Z}_2) \to \check{C}^{k+1}(\mathfrak{U}; \mathbb{Z}_2)$ is defined by

$$(\delta f)(\alpha_0, \ldots, \alpha_{k+1}) = \sum_{j=0}^{k+1} (-1)^j f(\alpha_0, \ldots, \widehat{\alpha}_j, \ldots, \alpha_{k+1}),$$

where the hat indicates that the corresponding term does not appear. The function $(\alpha, \beta) \mapsto \sigma_{\alpha\beta}$ can thus be understood as a 1-cochain $\sigma \in \check{C}^1(\mathfrak{U}; \mathbb{Z}_2)$, and (32.2) then says $\delta\sigma = 0$, i.e. σ is a cocycle. The cohomology class it represents in $\check{H}^1(B; \mathbb{Z}_2)$ turns out to be an invariant of the vector bundle $\pi : E \to B$ up to isomorphism, and is called the **first Stiefel-Whitney class** $w_1(E) \in \check{H}^1(B; \mathbb{Z}_2)$; this is the simplest of the standard characteristic classes for real vector bundles. It vanishes if and only if the bundle is orientable (see Exercise 32.7 below).

Exercise 32.7. The cocycle $\sigma \in \check{C}^1(\mathfrak{U}; \mathbb{Z}_2)$ described in Remark 32.6 is a coboundary if and only if there exists a function $I \to \mathbb{Z}_2 : \alpha \mapsto o_\alpha$ such that $\sigma_{\alpha\beta} = o_\alpha - o_\beta$ for every $\alpha, \beta \in I$ with $\mathcal{U}_\alpha \cap \mathcal{U}_\beta \neq \varnothing$. Show that this is true if and only if $E \to B$ is orientable.

32.2. Smoothness. If the base of our vector bundle is not just a topological space B but also a smooth manifold M, then we can speak of *smooth* vector bundles over M.

Definition 32.8. A **smooth structure** (*glatte Struktur*) on a vector bundle $\pi : E \to M$ over a smooth manifold M is a maximal bundle atlas with the property that all transition functions are smooth. When equipped with this extra data, $E \to M$ is called a **smooth vector bundle**, and the trivializations in its chosen bundle atlas are called **smooth trivializations**.

Since this is a course on differential geometry rather than topology, we will typically omit the word "smooth" when we talk about vector bundles and trivializations: *all vector bundles in this course will be assumed smooth unless stated otherwise.*

In the same manner as Exercise 32.1, the total space E of a smooth vector bundle $\pi : E \to M$ naturally inherits from its smooth local trivializations and the smooth structure of M the structure of a smooth manifold, such that the projection π and all smooth trivializations Φ_α become smooth maps. Having a smooth structure allows us also to speak of **smooth sections**: a section $s : M \to E$ is called smooth if its local representatives $s_\alpha : \mathcal{U}_\alpha \to \mathbb{F}^m$ with respect to smooth trivializations $(\mathcal{U}_\alpha, \Phi_\alpha)$ are all smooth functions. This definition makes sense due to the formula $s_\beta = g_{\beta\alpha} s_\alpha$ that relates two local representatives on an overlap domain $\mathcal{U}_\alpha \cap \mathcal{U}_\beta$: since the transition function $g_{\beta\alpha}$ is smooth, s_β is smooth if and only if s_α is smooth. One can alternatively appeal to the smooth structure on the total space E and consider a section $s : M \to E$ smooth if it defines a smooth map between these two smooth manifolds; this is equivalent to the definition in terms of local representatives. Since we only intend to consider smooth bundles henceforth, we shall modify the definition of the notation $\Gamma(E)$ so as to exclude continuous sections that are not smooth: from now on,

$$\Gamma(E) := \{\text{smooth sections of } \pi : E \to M\}.$$

Similarly, we will typically only consider local frames $e_1, \ldots, e_m$ that consist only of smooth sections, which is true if and only if the corresponding local trivialization is smooth.

Remark 32.9. In the smooth context, it is also appropriate to modify the definition of a **vector bundle isomorphism** $\Psi : E \to F$ and require Ψ to be a diffeomorphism rather than just a homeomorphism. This detail makes no actual difference to the question of whether two bundles are isomorphic, i.e. one can show that if two smooth bundles admit a continuous bundle isomorphism, then they also admit a smooth one. This is not trivial to prove, but it follows from general results on approximation of continuous maps by smooth ones as in [**Hir94**], together with the important fact from algebraic topology that the isomorphism classes of vector bundles of a given rank over a given space M have a natural bijective correspondence with the homotopy types of maps from M into some *classifying space* B. One can then appeal to the fact that two smooth maps $M \to B$ are continuously homotopic if and only if they are smoothly homotopic.

There are obvious generalizations of the notions above to allow only finite differentiability: if M is a manifold of class C^k, one can define a C^k*-structure* on a vector bundle $E \to M$ by requiring all transition functions to be of class C^k, and the notion of a C^k-section of $E \to M$ then makes sense because products of C^k-functions are also C^k-functions. However, if $E \to M$ only has a C^k-structure and $k < \infty$, then the notion of a C^ℓ-section cannot be defined for $\ell > k$, as it will typically depend on the choice of local trivialization.

Example 32.10. For any smooth n-dimensional manifold M, the tangent bundle $TM \to M$ is naturally a smooth real vector bundle of rank n, whose smooth sections are the smooth vector fields

$$\Gamma(TM) = \mathfrak{X}(M).$$

For any smooth chart $x = (x^1, \ldots, x^n) : \mathcal{U} \to \mathbb{R}^n$, the coordinate vector fields $\partial_1, \ldots, \partial_n \in \mathfrak{X}(\mathcal{U}) = \Gamma(TM|_{\mathcal{U}})$ define a smooth frame for TM over $\mathcal{U}$ and thus a local trivialization. The smoothness of the resulting transition functions follows easily from the smoothness of the transition maps in the atlas of M. More generally, if M is a manifold of class C^k for some $k \geqslant 1$, then $TM \to M$ becomes a bundle of class C^{k-1}; one derivative is lost because the transition functions for TM depend on first derivatives of the transition maps for M.

An analogous notion in complex geometry is worth mentioning in this context. Recall that on an open subset $\mathcal{U} \subset \mathbb{C}^n$, a function $f : \mathcal{U} \to \mathbb{C}^m$ is called **holomorphic** (*holomorph*) if its complex partial derivatives

$$(32.3) \qquad \frac{\partial f}{\partial z^j}(z^1, \ldots, z^n) = \lim_{z \to 0} \frac{f(z^1, \ldots, z^j + z, \ldots, z^n) - f(z^1, \ldots, z^n)}{z} \in \mathbb{C}^m$$

exist for all $j = 1, \ldots, n$ at all points $(z^1, \ldots, z^n) \in \mathcal{U}$, where it should be emphasized that the parameter z appearing in the limit is complex. As in the standard story of one complex variable, one can show that holomorphic functions of several complex variables are always smooth, where "smooth" in this case means the same thing that it means in real analysis, i.e. we identify $\mathbb{C}^n$ with $\mathbb{R}^{2n}$ in order to view f as a function of $2n$ real variables. Assembling the complex partial derivatives at a point $p = (z^1, \dots, z^n) \in \mathcal{U}$ into a Jacobian matrix produces a differential

$$Df(p) : \mathbb{C}^n \to \mathbb{C}^m,$$

which is a complex-linear map, and it is not too hard to show in fact that a smooth function $f : \mathcal{U} \to \mathbb{C}^m$ is holomorphic if and only if its differential at every point is complex linear. (Note that for a real-linear map $A : \mathbb{C}^n \to \mathbb{C}^m$, one requires $A(\lambda v) = \lambda A v$ for all $\lambda \in \mathbb{R}$, but not necessarily for $\lambda \in \mathbb{C} \backslash \mathbb{R}$.)

With the notion of holomorphic functions understood, one defines an **n-dimensional complex manifold** to be a smooth manifold of real dimension $2n$, equipped with a maximal atlas of charts having the form

$$M \overset{\text{open}}{\supset} \mathcal{U} \xrightarrow{(z^1,\dots,z^n)} \mathbb{C}^n = \mathbb{R}^{2n}$$

such that all the transition maps are holomorphic. An atlas with this property is called a **holomorphic atlas**, or equivalently a **complex structure** (*komplexe Struktur*) on M. When M is a complex manifold, one can define the notion of a **holomorphic function** $f : M \to V$ with values in any complex vector space V; it means simply that f looks holomorphic when expressed as a function of n complex variables in any of the charts in its holomorphic atlas. This notion makes sense due to the fact that compositions of holomorphic functions are always holomorphic.

The most obvious example of a complex manifold is $\mathbb{C}^n$, and the next most obvious is an arbitrary open subset of $\mathbb{C}^n$. The most popular *compact* example is the complex projective space, defined in the following exercise.

Exercise 32.11. The **complex projective n-space** (*komplexer projektiver Raum*) is defined as the set of all complex lines through the origin in $\mathbb{C}^{n+1}$: more precisely,

$$\mathbb{CP}^n = (\mathbb{C}^{n+1} \backslash \{0\}) / \sim,$$

where two nontrivial vectors $v, w \in \mathbb{C}^{n+1}$ are considered equivalent if and only if $v = \lambda w$ for some $\lambda \in \mathbb{C}$. It is convenient to denote points in $\mathbb{CP}^n$ via so-called **homogeneous coordinates**, in which the symbol

$$[z_0 : \ldots : z_n] \in \mathbb{CP}^n$$

means the equivalence class containing the vector $(z_0, \dots, z_n) \in \mathbb{C}^{n+1} \backslash \{0\}$. For $j = 0, \dots, n$, define the open subset $\mathcal{U}_j := \left\{ [z_0 : \ldots : z_n] \in \mathbb{CP}^n \;\middle|\; z_j \neq 0 \right\}$ and a map $\varphi_j : \mathbb{C}^n \to \mathbb{CP}^n$ by

$$\varphi_j(w_1, \dots, w_n) := [w_1 : \ldots : w_j : 1 : w_{j+1} : \ldots : w_n].$$

(a) Show that for each $j = 0, \dots, n$, φ_j is an injective map onto $\mathcal{U}_j$, thus its inverse defines a chart.
(b) Show that the charts $\varphi_j^{-1} : \mathcal{U}_j \to \mathbb{C}^n$ for $j = 0, \dots, n$ define a holomorphic atlas on $\mathbb{CP}^n$.
(c) Show that $\mathbb{CP}^1$ is diffeomorphic to S^2.

Definition 32.12. Suppose M is a complex manifold and $\pi : E \to M$ is a complex vector bundle. A **holomorphic structure** on $\pi : E \to M$ is a maximal bundle atlas with the property that all transition functions are holomorphic. With this extra data, $\pi : E \to M$ is called a **holomorphic vector bundle**, and the trivializations in its bundle atlas are called **holomorphic trivializations**.

Exercise 32.13. Show that for any complex manifold M, the tangent bundle $TM \to M$ has a natural holomorphic structure.

For a holomorphic bundle $E \to M$, one calls $s \in \Gamma(E)$ a **holomorphic section** if its local representatives $s_\alpha : \mathcal{U}_\alpha \to \mathbb{C}^m$ in holomorphic trivializations $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{C}^m$ are all holomorphic functions. This notion makes sense due to the fact that whenever $g_{\beta\alpha}$ and s_α are both holomorphic functions on $\mathcal{U}_\alpha \cap \mathcal{U}_\beta \subset M$, the *product* $s_\beta = g_{\beta\alpha} s_\alpha$ is also holomorphic. Note that on an arbitrary smooth complex vector bundle over a complex manifold—without a holomorphic structure—the notion of holomorphic sections is not well defined, because the question of whether s_α is holomorphic will depend on the choice of trivialization Φ_α.

Exercise 32.14. Show that on any holomorphic vector bundle $\pi : E \to M$, the total space E inherits a natural complex structure such that the projection π and the holomorphic local trivializations $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{C}^m$ are all holomorphic maps. Moreover, a section $s : M \to E$ is holomorphic if and only if it defines a holomorphic map between complex manifolds.

Remark 32.15. One serious qualitative difference between smooth and holomorphic vector bundles is the following. On a smooth vector bundle $E \to M$, the space $\Gamma(E)$ of smooth sections is always very large, and in fact any smooth section defined in coordinates near a given point p can be extended globally to a smooth section $M \to E$ just by multiplying it by a smooth bump function supported near p and extending it to the rest of M as 0. For holomorphic sections this trick does not work, because compactly-supported bump functions are *never* holomorphic, and this is one symptom of the fact that in general, one cannot expect nontrivial globally-defined holomorphic sections to exist. But holomorphic sections do always exist *locally*, i.e. every point $p \in M$ has an open neighborhood $\mathcal{U} \subset M$ such that the restriction $E|_{\mathcal{U}} \to \mathcal{U}$ is a holomorphic vector bundle with an abundance of holomorphic sections. For this reason, complex geometry makes extensive use of *sheaf theory* in order to extract interesting global information from nontrivial data of a purely local nature.

Holomorphic vector bundles play a large role in complex geometry, just as smooth vector bundles do in the geometry of smooth manifolds. We will focus for most of this course on smooth manifolds without complex structures, but we may come back to this subject near the end of the semester.

I want to mention one more definition that is in the same spirit as the last two. One way of characterizing holomorphicity for functions on a domain $\mathcal{U} \subset \mathbb{C}^n$ with complex coordinates $z^j = x^j + iy^j$ for $j = 1, \dots, n$ is via the differential operators

$$\frac{\partial}{\partial z^j} := \frac{1}{2}\left(\frac{\partial}{\partial x^j} - i\frac{\partial}{\partial y^j}\right), \qquad \frac{\partial}{\partial \bar{z}^j} := \frac{1}{2}\left(\frac{\partial}{\partial x^j} + i\frac{\partial}{\partial y^j}\right), \tag{32.4}$$

which are obtained via a formal application of the chain rule if one pretends that $f : \mathcal{U} \to \mathbb{C}^n$ is a function of $2n$ "independent" variables $z^1, \dots, z^n, \bar{z}^1, \dots, \bar{z}^n$ with $\frac{\partial z^j}{\partial x^j} = 1$, $\frac{\partial \bar{z}^j}{\partial y^j} = -i$ and so forth. I put the word "independent" in quotation marks here because it is a fiction: the variables $\bar{z}^j = x^j - iy^j$ are not independent of $z^j = x^j + iy^j$, but pretending they are leads to a few useful conventions such as the definition (32.4), which makes many equations and computations in complex geometry more concise. One can check in particular that $f : \mathcal{U} \to \mathbb{C}^m$ is holomorphic if and only if

$$\frac{\partial f}{\partial \bar{z}^1} = \dots = \frac{\partial f}{\partial \bar{z}^n} \equiv 0,$$

and if this holds, then the remaining partial derivatives $\frac{\partial f}{\partial z^1}, \dots \frac{\partial f}{\partial z^n}$ are precisely the limits that were written in (32.3). From this perspective, a holomorphic structure is a smooth bundle atlas for which all the transition functions are annihilated by a certain set of differential operators. One could now take this further: what happens if we demand for the transition functions to be annihilated by *all* first-order partial derivative operators?

Definition 32.16. A **flat structure** on a vector bundle $\pi : E \to M$ is a maximal bundle atlas with the property that all transition functions are locally constant. With this extra data, $\pi : E \to M$ is called a **flat vector bundle**, and the trivializations in its bundle atlas are called **flat trivializations**.

On a flat vector bundle $\pi : E \to M$, one can define the notion of a **flat section**, meaning a section $s : M \to E$ whose local representatives $s_\alpha : \mathcal{U}_\alpha \to \mathbb{F}^m$ in arbitrary flat trivializations $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^m$ are locally constant. If $\pi : E \to M$ is not equipped with a flat structure, then there is no special significance to the condition of $s_\alpha : \mathcal{U}_\alpha \to \mathbb{F}^m$ being constant, because the same section will generally have another local representative $s_\beta : \mathcal{U}_\beta \to \mathbb{F}^m$ that is not constant. This issue—the fact that vector bundles in general do not come with any natural definition of "constant sections"—is the fundamental motivation for the notion of connections.

32.3. Connections. While flat vector bundles come with a natural notion of "locally constant" sections, flat structures do not actually arise very often in nature: one can for instance use the first Chern class to show that *most* complex vector bundles do not admit any flat structure at all. Connections provide a less stringent notion that makes "locally constant" a well-defined notion without restricting the class of smooth bundles that we consider.

The definition of a connection comes in several equivalent variants: we surveyed all of them in Lectures 19 and 20 last semester, and will do so again in the more general setting of fiber bundles later this semester. For now, the following variant will be the most useful:

Definition 32.17. A **connection** (*Zusammenhang*) on the vector bundle $\pi : E \to M$ is a real-bilinear operator

$$\mathfrak{X}(M) \times \Gamma(E) \to \Gamma(E) : (X, s) \mapsto \nabla_X s$$

satisfying the following two properties for all $X \in \mathfrak{X}(M)$ and $s \in \Gamma(E)$:

(1) (C^∞-linearity) $\nabla_{fX} s = f \nabla_X s$ for all $f \in C^\infty(M, \mathbb{R})$;
(2) (Leibniz rule) $\nabla_X (fs) = df(X)s + f \nabla_X s$ for all $f \in C^\infty(M, \mathbb{F})$.

Observe that if $\mathbb{F} = \mathbb{C}$, then $(X, s) \mapsto \nabla_X s$ is real linear in X but complex linear in s, as one sees by taking constant complex-valued functions in the Leibniz rule. By C^∞-linearity, the value of $\nabla_X s \in \Gamma(E)$ at a single point $p \in M$ depends linearly on the value of X at p but not otherwise on the vector field $X \in \mathfrak{X}(M)$, thus it is sensible to write

$$\nabla_{X(p)} s := (\nabla_X s)(p) \in E_p$$

and interpret this as the directional derivative of s at $p \in M$ in the direction $X(p) \in T_p M$. Directional derivatives defined in this way via a connection are called **covariant derivatives** (*kovariante Ableitungen*).

Connections on a bundle $E \to M$ are not unique, but for any two connections ∇, ∇' on the same bundle, the difference $A(X, s) := \nabla'_X s - \nabla_X s$ is C^∞-linear in both X and s, thus for each $p \in M$ there is a bilinear map $A_p : T_p M \times E_p \to E_p$ satisfying

$$\nabla'_X s = \nabla_X s + A_p(X, s(p))$$

for all $s \in \Gamma(E)$ and $X \in T_p M$. The set of connections is thus an affine space over the space of all bilinear bundle maps $TM \oplus E \to E$ (see the next lecture for the notions of direct sum and bundle maps); in particular, it is a convex set. One can use this fact to piece local constructions of connections together via a partition of unity and thus show that every vector bundle admits a connection, a fact that we will prove in much more general terms when we discuss principal bundles. One sees a special case of the relation between two different connections whenever one writes down a connection in local coordinates. Recall that any local trivialization $\Phi : E|_{\mathcal{U}} \to \mathcal{U} \times \mathbb{F}^m$ defines a bijection between $\Gamma(E|_{\mathcal{U}})$ and $C^\infty(\mathcal{U}, \mathbb{F}^m)$, thus it also determines a so-called **trivial connection**

∇^0 on $E|_{\mathcal{U}}$ whose action on sections is just the usual directional derivative of $\mathbb{F}^m$-valued functions. Now, if $\mathcal{U} \subset M$ is an open subset admitting both a chart $x = (x^1, \ldots, x^n) : \mathcal{U} \to \mathbb{R}^n$ and a frame $e_1, \ldots, e_m \in \Gamma(E|_{\mathcal{U}})$ for E, then any connection ∇ can be described over $\mathcal{U}$ via its **Christoffel symbols**

$$\Gamma^a_{ib} : \mathcal{U} \to \mathbb{F}, \qquad a, b \in \{1, \ldots, m\}, \quad i \in \{1, \ldots, n\},$$

defined as the components of the covariant derivative of e_b in the ith coordinate direction,

$$\Gamma^a_{ib} := (\nabla_i e_b)^a, \qquad \text{i.e.} \qquad \nabla_i e_b = \Gamma^a_{ib} e_a,$$

where we abbreviate

$$\nabla_i := \nabla_{\partial_i}.$$

Applying the Leibniz rule to an arbitrary section $s = s^a e_a$ over $\mathcal{U}$ now gives the formula

$$(\nabla_i s)^a = \partial_i s^a + \Gamma^a_{ib} s^b, \tag{32.5}$$

in which the first term on the right hand side is a coordinate representation for the trivial connection ∇^0, and the Christoffel symbols describe the bilinear bundle map $\nabla - \nabla^0$.

33. Affine connections and geodesics

We have a few more things to recall about connections on general vector bundles before specializing to the case of a tangent bundle, so that we can talk about geodesics and venture into Riemannian geometry.

33.1. The pullback connection. For any continuous map $f : N \to M$, a vector bundle $E \to M$ determines a vector bundle over N called the **pullback** or **induced bundle**

$$f^*E \to N,$$

which has fibers $(f^*E)_p := E_{f(p)}$ for $p \in N$. Sections of $f^*E \to N$ are maps s that send points $p \in N$ to vectors $s(p) \in E_{f(p)}$, and these are also often called **sections of** E **along** f. For example, if $s \in \Gamma(E)$ is any section of E, then the composition $s \circ f$ is a section of f^*E, though in general sections of f^*E may take more general forms than this. Any frame $e_1, \ldots, e_m$ for $E \to M$ over an open subset $\mathcal{U} \subset M$ gives rise to a frame for f^*E over $f^{-1}(\mathcal{U}) \subset N$ consisting of the sections $e_1 \circ f, \ldots, e_m \circ f$, and in this way any bundle atlas $\{(\mathcal{U}_\alpha, \Phi_\alpha)\}_{\alpha \in I}$ for $E \to M$ with transition functions $g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{GL}(m, \mathbb{F})$ gives rise to a bundle atlas for $f^*E \to N$ whose transition functions are

$$g_{\beta\alpha} \circ f : f^{-1}(\mathcal{U}_\alpha) \cap f^{-1}(\mathcal{U}_\beta) \to \mathrm{GL}(m, \mathbb{F}).$$

If f is smooth, it follows that any smooth structure on $E \to M$ induces a natural smooth structure on $f^*E \to N$.

Continuing under the assumption that $f : N \to M$ is a smooth map, any connection ∇ on $E \to M$ now induces a natural connection on $f^*E \to N$, called the **pullback connection**, via the condition that for any $s \in \Gamma(E)$, $p \in N$ and $X \in T_pM$,

$$\nabla_X (s \circ f) := \nabla_{f_* X} s \in E_{f(p)} = (f^*E)_p. \tag{33.1}$$

Exercise 33.1. If you haven't seen pullback connections before, convince yourself that a connection on f^*E satisfying (33.1) does indeed exist and is unique.
Hint: Uniqueness is easy since any section of f^*E can be written locally in terms of a frame of the form $e_1 \circ f, \ldots, e_m \circ f$ for some frame $e_1, \ldots, e_m$ on E. You need to check that the resulting local definition of the pullback connection does not depend on the choice of frame.

Given a frame $e_1, \dots, e_m$ for E near $f(p)$ and charts $x = (x^1, \dots, x^k)$ for N near p and $y = (y^1, \dots, y^n)$ for M near $f(p)$, let us write

$$y \circ f = (f^1, \dots, f^n)$$

on a neighborhood of p. The relation (33.1) then determines the covariant derivatives of $e_j \circ f$ for each $j = 1, \dots, m$, and the Christoffel symbols $\widehat{\Gamma}^a_{ib}$ for the pullback connection in the chosen coordinates are thus related to the Christoffel symbols Γ^a_{jb} of the connection on $E \to M$ by

$$\widehat{\Gamma}^a_{ib}(p) = \left(\nabla_{\frac{\partial}{\partial x^i}(p)}(e_b \circ f)\right)^a = \left(\nabla_{\frac{\partial f}{\partial x^i}(p)} e_b\right)^a = \left(\frac{\partial f^j}{\partial x^i}(p) \nabla_{\frac{\partial}{\partial y^j}(f(p))} e_b\right)^a = \Gamma^a_{jb}(f(p)) \frac{\partial f^j}{\partial x^i}(p).$$

The local formula for the covariant derivative of a section $s \in \Gamma(f^*E)$ along f using the pullback connection is therefore

$$(\nabla_i s)^a = \partial_i s^a + (\Gamma^a_{jb} \circ f)\, (\partial_i f^j) s^b. \tag{33.2}$$

Given a connection ∇ on $\pi : E \to M$ and a smooth map $f : N \to M$, a section $s \in \Gamma(f^*E)$ along f is called **parallel** if $\nabla s \equiv 0$. While parallel sections cannot generally be expected to exist, even locally, there is an important exception: if N is an interval $I \subset \mathbb{R}$ and $I \to M : t \mapsto \gamma(t)$ is a smooth path, then the equation $\nabla_t s := \nabla_{\frac{\partial}{\partial t}} s \equiv 0$ for a section $s \in \Gamma(\gamma^* E)$ along the path becomes a first-order linear ODE with smooth coefficients, and thus has unique solutions determined linearly by an initial value. If $0 \in I$, then we shall write

$$P^t_\gamma : E_{\gamma(0)} \to E_{\gamma(t)}, \qquad t \in I$$

for the unique smooth family of vector space isomorphisms such that $t \mapsto P^t_\gamma(v)$ is a parallel section along γ for each $v \in E_{\gamma(0)}$, and we call these the **parallel transport** (*Parallelverschiebung*) isomorphisms along γ. For an arbitrary section $s(t) \in E_{\gamma(t)}$, the Leibniz rule now gives another formula for the covariant derivative of s with respect to t at $t = 0$: it is the ordinary derivative after parallel transporting $s(t)$ along γ so that its values lie in a single fiber,

$$\nabla_t s(0) = \left.\frac{d}{dt}(P^t_\gamma)^{-1}(s(t))\right|_{t=0} \in E_{\gamma(0)}. \tag{33.3}$$

33.2. Algebraic constructions. To conclude our quick survey on connections, recall that various algebraic operations one can perform on vector spaces give rise to similarly natural operations on vector bundles, and for each of there are also natural notions of *induced* connections. In case you have doubts about any of the objects described below having natural smooth vector bundle structures, I recommend thinking about how you might use a choice of local frames on the given bundles E and/or F to derive the most natural choice of frame on the constructed bundle. A similar trick using parallel transport and the formula (33.3) leads to the most natural choice of a connection on each. (For more detailed discussions of these bundles, see §17.4 and §21.3 from the first semester.)

33.2.1. *Duals.* Any bundle $E \to M$ has a **dual bundle** (*Dualbündel*)

$$E^* \to M$$

whose fiber over a point $p \in M$ is the dual space $E^*_p = \mathrm{Hom}(E_p, \mathbb{F})$ of the corresponding fiber of E. Given a connection on E, the induced connection on E^* is uniquely determined by the Leibniz rule

$$\mathcal{L}_X(\lambda(\eta)) = (\nabla_X \lambda)(\eta) + \lambda(\nabla_X \eta) \qquad \text{for all } X \in \mathfrak{X}(M),\ \eta \in \Gamma(E) \text{ and } \lambda \in \Gamma(E^*),$$

where the pairing $\lambda(\eta)$ is understood as the smooth scalar-valued function $M \to \mathbb{F} : p \mapsto \lambda(p)\,(\eta(p))$.

EXERCISE 33.2. Suppose $(x^1, \dots, x^n)$ and $e_1, \dots, e_m$ are a chart for M and a frame for E repsectively, both defined over an open subset $\mathcal{U} \subset M$, and let $e_*^1, \dots, e_*^m$ denote the **dual frame** for E^* over $\mathcal{U}$, defined by the condition

$$e_*^a(e_b) \equiv \delta_b^a.$$

Writing sections $\lambda \in \Gamma(E^*)$ on $\mathcal{U}$ in the form $\lambda = \lambda_a e_*^a$ in terms of component functions $\lambda_a : \mathcal{U} \to \mathbb{F}$, show that the Christoffel symbols Γ_{ib}^a of a chosen connection on E give rise to the local formulas

$$(\nabla_i e_*^b)_a = -\Gamma_{ia}^b, \qquad \text{and thus} \qquad (\nabla_i \lambda)_a = \partial_i \lambda_a - \Gamma_{ia}^b \lambda_b$$

for the induced connection on E^*.

33.2.2. *Direct sums.* For any two bundles $E, F \to M$, there is a **direct sum** (*direkte Summe*) bundle

$$E \oplus F \to M$$

whose fiber over a point $p \in M$ is the Cartesian product (or equivalently direct sum) of vector spaces $E_p \times F_p$. Given connections on E and F, the induced connection on $E \oplus F$ is described very simply via the formula

$$\nabla_X(\eta, \xi) = (\nabla_X \eta, \nabla_X \xi) \qquad \text{for all } X \in \mathfrak{X}(M),\ \eta \in \Gamma(E) \text{ and } \xi \in \Gamma(F),$$

where we use the obvious identification of $\Gamma(E \oplus F)$ with $\Gamma(E) \times \Gamma(F)$.

33.2.3. *Tensor products.* For any two bundles $E, F \to M$, there is a **tensor product** (*Tensorprodukt*) bundle

$$E \otimes F \to M$$

whose fiber over a point $p \in M$ is the tensor product of vector spaces $E_p \otimes F_p$. Given connections on E and F, the induced connection on $E \otimes F$ is uniquely determined by the Leibniz rule

$$\nabla_X(\eta \otimes \xi) = \nabla_X \eta \otimes \xi + \eta \otimes \nabla_X \xi \qquad \text{for all } X \in \mathfrak{X}(M),\ \eta \in \Gamma(E) \text{ and } \xi \in \Gamma(F),$$

where we denote $(\eta \otimes \xi)(p) := \eta(p) \otimes \xi(p) \in E_p \otimes F_p$. This definition determines the covariant derivatives of arbitrary sections of $E \otimes F$ since locally all of them are linear combinations of products of the form $\eta \otimes \xi$. By induction, any finite collection of vector bundles over M has a tensor product that inherits a natural connection from any choice of connections on its factors. An important special case is

$$E_\ell^k := E^{\otimes k} \otimes (E^*)^{\otimes \ell} := \underbrace{E \otimes \ldots \otimes E}_{k} \otimes \underbrace{E^* \otimes \ldots \otimes E^*}_{\ell}$$

for integers $k, \ell \geqslant 0$, whose fiber over a point $p \in M$ has a natural identification with the space of multilinear maps

$$\underbrace{E_p^* \times \ldots \times E_p^*}_{k} \times \underbrace{E_p \times \ldots \times E_p}_{\ell} \to \mathbb{F},$$

and we adopt the convention that $(E_p)_\ell^k := \mathbb{F}$ for $k = \ell = 0$, so E_0^0 is the trivial line bundle over M. When E is the tangent bundle TM, we prefer to write

$$T_\ell^k M := (TM)_\ell^k,$$

so that sections in $\Gamma(T_\ell^k M)$ are tensor fields of type (k, ℓ). Any connection on E induces a connection first on E^* via §33.2.1 and then on E_ℓ^k via the tensor product construction, so in particular, a connection on $TM \to M$ gives rise to a natural covariant derivative operator on tensor fields

$$\nabla : \Gamma(T_\ell^k M) \to \Gamma(T_{\ell+1}^k M),$$

where $\nabla S \in \Gamma(T_{\ell+1}^k M)$ is defined for $S \in \Gamma(T_\ell^k M)$ by

$$\nabla S(\lambda^1, \dots, \lambda^k, X_0, \dots, X_\ell) := (\nabla_{X_0} S)(\lambda^1, \dots, \lambda^k, X_1, \dots, X_\ell).$$

EXERCISE 33.3. Assume a connection on $E \to M$ has been chosen and $E^k_\ell \to M$ is equipped with the induced connection.

(a) Show that the connections on E and E^* determine the connection on E^k_ℓ via the Leibniz rule

$$\begin{aligned}\mathcal{L}_X\big(S(\lambda^1,\dots,\lambda^k,\eta_1,\dots,\eta_\ell)\big) &= (\nabla_X S)(\lambda^1,\dots,\lambda^k,\eta_1,\dots,\eta_\ell)\\ &\quad + S(\nabla_X\lambda^1,\dots,\lambda^k,\eta_1,\dots,\eta_\ell) + \ldots + S(\lambda^1,\dots,\nabla_X\lambda^k,\eta_1,\dots,\eta_\ell)\\ &\quad + S(\lambda^1,\dots,\lambda^k,\nabla_X\eta_1,\dots,\eta_\ell) + \ldots + S(\lambda^1,\dots,\lambda^k,\eta_1,\dots,\nabla_X\eta_\ell)\end{aligned}$$

for $S \in \Gamma(E^k_\ell)$, $\lambda^1,\dots,\lambda^k \in \Gamma(E^*)$, $\eta_1,\dots,\eta_\ell \in \Gamma(E)$ and $X \in \mathfrak{X}(M)$.
Hint: It suffices (why?) to consider sections of the form $S = \xi_1 \otimes \ldots \otimes \xi_k \otimes \alpha^1 \otimes \ldots \otimes \alpha^\ell \in \Gamma(E^k_\ell)$ for $\xi_1,\dots,\xi_k \in \Gamma(E)$ and $\alpha^1,\dots,\alpha^\ell \in \Gamma(E^*)$.

(b) In the case $k = 1$, there is a canonical identification of sections of E^1_ℓ with multilinear bundle maps $E^{\oplus \ell} \to E$; show that from this perspective, the connection on E^1_ℓ is determined by the connection on E via the Leibniz rule

$$\nabla_X(S(\eta_1,\dots,\eta_\ell)) = (\nabla_X S)(\eta_1,\dots,\eta_\ell) + S(\nabla_X\eta_1,\dots,\eta_\ell) + \ldots + S(\eta_1,\dots,\nabla_X\eta_\ell)$$

for $S \in \Gamma(E^1_\ell)$, $\eta_1,\dots,\eta_\ell \in \Gamma(E)$ and $X \in \mathfrak{X}(M)$.

EXERCISE 33.4. Fix a chart $(x^1,\dots,x^n)$ and frame $e_1,\dots,e_m$ over a subset $\mathcal{U} \subset M$ as in Exercise 33.2, let Γ^a_{ib} denote the correpsonding Christoffel symbols for a chosen connection on E, $e^1_*,\dots,e^m_*$ denote the dual frame for E^* determined by $e_1,\dots,e_m$, and write sections $S \in \Gamma(E^k_\ell)$ over $\mathcal{U}$ in terms of their **components** ${S^{a_1\ldots a_k}}_{b_1\ldots b_\ell} : \mathcal{U} \to \mathbb{F}$ as

$$S = {S^{a_1\ldots a_k}}_{b_1\ldots b_\ell}\, e_{a_1} \otimes \ldots \otimes e_{a_k} \otimes e^{b_1}_* \otimes \ldots \otimes e^{b_\ell}_*.$$

Show that the induced connection on E^k_ℓ is then given locally by the formula

$$\begin{aligned}{(\nabla_i S)^{a_1\ldots a_k}}_{b_1\ldots b_\ell} = \partial_i {S^{a_1\ldots a_k}}_{b_1\ldots b_\ell} + \Gamma^{a_1}_{ic}{S^{c a_2\ldots a_k}}_{b_1\ldots b_\ell} + \ldots + \Gamma^{a_k}_{ic}{S^{a_1\ldots a_{k-1}c}}_{b_1\ldots b_\ell}&\\ - \Gamma^c_{ib_1}{S^{a_1\ldots a_k}}_{cb_2\ldots b_\ell} - \ldots - \Gamma^c_{ib_\ell}{S^{a_1\ldots a_k}}_{b_1\ldots b_{\ell-1}c}&.\end{aligned}$$

33.2.4. *Bundle maps.* For any two bundles $E, F \to M$, the **bundle of linear maps**

$$\operatorname{Hom}(E,F) \to M$$

has fiber $\operatorname{Hom}(E_p,F_p) := \{A : E_p \to F_p \mid A \text{ linear}\}$ for a point $p \in M$, and is canonically isomorphic to $E^* \otimes F$. The special case $E = F$ comes up often and has its own notation: it gives us the **endomorphism bundle**

$$\operatorname{End}(E) := \operatorname{Hom}(E,E) \to M.$$

The connection that $\operatorname{Hom}(E,F) \cong E^* \otimes F$ inherits from any choice of connections on E and F via §33.2.1 and §33.2.3 can also be described via the Leibniz rule

$$\nabla_X(A\eta) = (\nabla_X A)\eta + A(\nabla_X\eta) \qquad \text{for all } \eta \in \Gamma(E),\ A \in \Gamma(\operatorname{Hom}(E,F)) \text{ and } X \in \mathfrak{X}(M).$$

Sections of $\operatorname{Hom}(E,F)$ are known as **smooth linear bundle maps** $E \to F$; we will call them *bundle maps* for short when the rest is understood from context.

The notation $\operatorname{Hom}(\cdot,\cdot)$ allows us also to regard connections ∇ on $E \to M$ as linear operators

$$\nabla : \Gamma(E) \to \Gamma(\operatorname{Hom}(TM,E)), \qquad (\nabla\eta)(X) := \nabla_X\eta$$

that satisfy the Leibniz rule $\nabla(f\eta) = df(\cdot)\eta + f\,\nabla\eta$ for all $f \in C^\infty(M,\mathbb{F})$ and $\eta \in \Gamma(E)$. From this perspective, the C^∞-linearity condition in Definition 32.17 is redundant, as it is implied by the assumption that ∇s is a section of $\operatorname{Hom}(TM,E)$.

33.2.5. *Subbundles and quotients.* A subset $F \subset E$ of a vector bundle $\pi : E \to B$ is called a **subbundle** (*Unterbündel*) if F admits the structure of a vector bundle over B such that the inclusion $F \hookrightarrow E$ is a linear bundle map. This means in particular that the fibers of F are linear subspaces of the fibers of E. There is then also a **quotient bundle** (*Quotientenbündel*)

$$E/F \to B,$$

whose fiber at $p \in B$ is the quotient vector space E_p/F_p, and the quotient projections $E_p \to E_p/F_p$ define a smooth fiberwise-surjective linear bundle map

$$E \to E/F : v \mapsto [v].$$

Any connection ∇ on E then descends to a connection on E/F by defining

$$\nabla_X[\eta] := [\nabla_X \eta] \qquad \text{for all } \eta \in \Gamma(E).$$

It is not true in general that a connection on E naturally determines a connection on every subbundle $F \subset E$, as covariant derivatives of sections of F using a connection on E take values in E, but not necessarily in F. There are some important exceptions, however, such as the **exterior** (*äußere*) tensor bundles

$$\Lambda^k E \subset E_0^k = E^{\otimes k}, \qquad \Lambda^k E^* \subset E_k^0 = (E^*)^{\otimes k},$$

whose fibers $\Lambda^k E_p$ and $\Lambda^k E_p^*$ over a point $p \in M$ can be regarded as the space of antisymmetric k-fold multilinear forms on E_p^* or E_p respectively. The most familiar example is $\Lambda^k T^*M$, whose sections are the smooth differential k-forms

$$\Gamma(\Lambda^k T^*M) = \Omega^k(M).$$

The next exercise shows that any choice of connection on E also determines natural connections on $\Lambda^k E$ and $\Lambda^k E^*$ for each $k \geqslant 0$.

EXERCISE 33.5. Given a connection on $E \to M$ and an integer $k \geqslant 2$, show that if $\omega \in \Gamma(E_k^0)$ is antisymmetric, then so is $\nabla_X \omega \in \Gamma(E_k^0)$ for every $X \in \mathfrak{X}(M)$.

33.3. Affine connections. A connection on the tangent bundle $TM \to M$ of a smooth manifold M is often called an **affine connection** (*affiner Zusammenhang*), or simply a **connection on** M. In a local chart $(x^1, \ldots, x^n)$ on some subset $\mathcal{U} \subset M$, it is natural to use $\partial_1, \ldots, \partial_n$ as a frame for TM over $\mathcal{U}$ and write the covariant derivative in terms of the corresponding Christoffel symbols

$$\Gamma_{jk}^i = (\nabla_j(\partial_k))^i, \qquad i, j, k \in \{1, \ldots, n\}.$$

We must now recall two important facts about connections in this special case.

The first is that any affine connection ∇ on M determines a distinguished class of smooth paths $\gamma : I \to M$; here $I \subset \mathbb{R}$ is an interval. We call γ a **geodesic** (*Geodäte* or *Geodätische Linie*) if its velocity vector $\dot{\gamma}(t) := \frac{d}{dt}\gamma(t) \in T_{\gamma(t)}M$ is parallel along γ, meaning γ satisfies the **geodesic equation**

$$\nabla_t \dot{\gamma}(t) = 0.$$

If γ takes values in the domain of a chart $x = (x^1, \ldots, x^n)$, we can write $x \circ \gamma = (\gamma^1, \ldots, \gamma^n)$ and apply the formula (33.2) to write the geodesic equation in the form

$$\ddot{\gamma}^i(t) + \Gamma_{jk}^i(\gamma(t))\dot{\gamma}^j(t)\dot{\gamma}^k(t) = 0,$$

or more succintly,

$$\ddot{\gamma}^i + \Gamma_{jk}^i \dot{\gamma}^j \dot{\gamma}^k = 0. \tag{33.4}$$

As a second-order ODE, this equation has a unique solution with any given initial position $\gamma(0) \in M$ and velocity $\dot\gamma(0) \in T_{\gamma(0)}M$, and we use this fact to define the so-called **exponential map** (*Exponentialabbildung*)

$$TM \supset \mathcal{O} \xrightarrow{\exp} M,$$

namely by setting

$$\exp(X) = \exp_p(X) := \gamma(1), \qquad \text{for } p \in M,\ X \in T_pM,$$

where γ is the unique geodesic satisfying $\gamma(0) = p$ and $\dot\gamma(0) = X$. The domain

$$\mathcal{O} \subset TM$$

of exp is an open subset consisting of all $X \in TM$ for which the maximal solution γ to $\nabla_t\dot\gamma = 0$ with $\dot\gamma(0) = X$ has 1 in its domain. It is straightforward to check that for every constant $c \in \mathbb{R}$ and geodesic $\gamma(t)$, the path $t \mapsto \gamma(ct)$ is also a geodesic, and as a consequence, the path

$$\gamma(t) := \exp_p(tX)$$

is in fact the unique geodesic with $\gamma(0) = p$ and $\dot\gamma(0) = X \in T_pM$. The exponential map is often used in this manner to write down parametrizations of geodesics.

Exercise 33.6. Given a manifold M with a chart $M \overset{\text{open}}{\supset} \mathcal{U} \xrightarrow{(x^1,\dots,x^n)} \mathbb{R}^n$ and affine connection ∇, suppose $\gamma : I \to \mathcal{U}$ is a nonconstant geodesic with image in $\mathcal{U}$, write $\gamma^i := x^i \circ \gamma$ for $i = 1,\dots,n$ and define $\rho : I \to [0,\infty)$ by

$$\rho(t) := [\gamma^1(t)]^2 + \ldots + [\gamma^n(t)]^2.$$

Prove: there exists an $\epsilon > 0$ such that every $t \in I$ with $\rho(t) < \epsilon$ satisfies $\rho''(t) > 0$. What can you conclude about the paths of geodesics in small coordinate balls about a point?
Hint: Using the geodesic equation, derive a formula for $\rho''(t)$ involving no second derivatives of the γ^i. Then prove and make use of the estimate $\left|\sum_{i,j}\dot\gamma^i\dot\gamma^j\right| \leqslant n^2\sum_k(\dot\gamma^k)^2$.

The second fact to recall about affine connections is that since ∇ defines a bilinear map $\mathfrak{X}(M) \times \Gamma(TM) \to \Gamma(TM) : (X,Y) \mapsto \nabla_XY$ and $\Gamma(TM) = \mathfrak{X}(M)$, there is a symmetry condition that can be imposed. It is most easily stated in terms of the **torsion tensor**

$$T \in \Gamma(T^1_2M), \qquad T(X,Y) := \nabla_XY - \nabla_YX - [X,Y] \in \mathfrak{X}(M) \quad \text{for} \quad X,Y \in \mathfrak{X}(M),$$

where one checks easily that the right hand side is C^∞-linear with respect to both X and Y and thus defines a smooth bilinear bundle map $T : TM \oplus TM \to TM$, i.e. a tensor field of type $(1,2)$. In local coordinates, one derives from (32.5) a formula for the components of the torsion tensor in terms of the Christoffel symbols, namely

$$T^i_{\ jk} = \Gamma^i_{jk} - \Gamma^i_{kj}. \tag{33.5}$$

The connection ∇ is called **symmetric** (*symmetrisch*) if its torsion tensor vanishes identically. The most useful consequence of this condition arises when one considers maps of the form

$$\mathbb{R}^2 \overset{\text{open}}{\supset} \mathcal{V} \to M : (s,t) \mapsto f(s,t),$$

in which case the partial derivatives ∂_sf and ∂_tf define **vector fields on** M **along** f, i.e. sections of the pullback bundle f^*TM. If f takes values in the domain of a chart $x = (x^1,\dots,x^n)$ and we write $x \circ f = (f^1,\dots,f^n)$, then applying (33.2) in this situation gives

$$(\nabla_s\partial_tf)^i = \partial_s\partial_tf^i + (\Gamma^i_{jk} \circ f)\,\partial_sf^j\,\partial_tf^k.$$

Combined with (33.5) and the definition of the torsion tensor, this implies

$$\nabla_s\partial_tf - \nabla_t\partial_sf = T(\partial_sf,\partial_tf), \tag{33.6}$$

and it follows that the relation

$$\nabla_s \partial_t f = \nabla_t \partial_s f \tag{33.7}$$

holds for all smooth maps $\mathbb{R}^2 \supset \mathcal{V} \xrightarrow{f} M$ if and only if ∇ is symmetric.

It is straightforward to show (see Exercise 33.7) that any affine connection can be "symmetrized" to produce one that is symmetric, thus symmetric connections exist on every manifold. However, one obtains a much more important variant of this statement when bundle metrics are incorporated into the picture; see Theorem 33.10 below.

Exercise 33.7. Show that for any affine connection ∇ on a manifold M, there exists a symmetric affine connection ∇' that has the same geodesics as ∇.
Hint: Write $\nabla' = \nabla + A$ for a bundle map A.

33.4. Riemannian metrics. A **Riemannian metric** (*Riemannsche Metrik*) on a manifold M is a positive bundle metric $\langle\ ,\ \rangle$ on the tangent bundle $TM \to M$. Such a bundle metric defines a tensor field $g \in \Gamma(T_2^0 M)$ such that

$$g_p(X, Y) = \langle X, Y\rangle \qquad \text{for } X, Y \in T_pM,\ p \in M,$$

and the pair (M, g) is then called a **Riemannian manifold** (*Riemannsche Mannigfaltigkeit*). If we drop the condition that $\langle\ ,\ \rangle$ is positive and allow it to be an indefinite bundle metric (cf. Remark 32.4), then g is called a **pseudo-Riemannian** metric and (M, g) a pseudo-Riemannian manifold. We will generally refer to both simply as "metrics" when the distinction does not matter and there is no danger of confusion.

In local coordinates $(x^1, \dots, x^n)$, we write the components of a metric $g \in \Gamma(T_2^0 M)$ as $g_{ij} := \langle \partial_i, \partial_j\rangle$, and the symmetry $g(X, Y) = g(Y, X)$ allows us to write

$$g = g_{ij}\, dx^i \otimes dx^j = \sum_{i \leqslant j} g_{ij}\, dx^i\, dx^j,$$

where in the second expression we are refraining from the Einstein summation convention and using the *symmetrization*

$$dx^i\, dx^j := \frac{1}{2}\left(dx^i \otimes dx^j + dx^j \otimes dx^i\right).$$

One example that comes up frequently is the **Euclidean metric** on $\mathbb{R}^n$, which in the standard global coordinates $(x^1, \dots, x^n)$ takes the form

$$g_E := (dx^1)^2 + \ldots + (dx^n)^2.$$

The nondegeneracy of the pairing $\langle\ ,\ \rangle$ on each tangent space implies that it defines bundle isomorphisms

$$\flat : TM \to T^*M : X \mapsto X_\flat := \langle X, \cdot\rangle, \qquad \sharp := \flat^{-1} : T^*M \to TM : \lambda \mapsto \lambda^\sharp,$$

often called **musical isomorphisms** due to the notation. This gives rise to a natural definition of a bundle metric on T^*M by

$$\langle \alpha, \beta\rangle := \langle \alpha^\sharp, \beta^\sharp\rangle,$$

and in local coordinates, the resulting pairing of coordinate differentials is often denoted by

$$g^{ij} := \langle dx^i, dx^j\rangle.$$

The notational convention for musical isomorphisms in terms of components in local coordinates is to use the same symbol for the corresponding vectors and 1-forms but raise or lower the index accordingly, thus

$$X = X^i \partial_i \quad \Leftrightarrow \quad X_\flat = X_i\, dx^i, \qquad \lambda = \lambda_i\, dx^i \quad \Leftrightarrow \quad \lambda^\sharp = \lambda^i \partial_i.$$

The local coordinate formula for the isomorphisms themselves then takes the form

$$X_i = g_{ij} X^j, \qquad \lambda^i = g^{ij} \lambda_j,$$

which shows in particular that the matrices with entries g_{ij} and g^{ij} are inverse to each other:

$$g^{ij} g_{jk} = \delta^i_k.$$

DEFINITION 33.8. Suppose $g = \langle\ ,\ \rangle \in \Gamma(E^0_2)$ is a smooth bundle metric on a real vector bundle $E \to M$. A connection ∇ on $E \to M$ is called a **metric connection** if it satisfies any of the following equivalent conditions:

(1) $\mathcal{L}_X \langle \eta, \xi \rangle = \langle \nabla_X \eta, \xi \rangle + \langle \eta, \nabla_X \xi \rangle$ for all $\eta, \xi \in \Gamma(E)$ and $X \in \mathfrak{X}(M)$;
(2) $\nabla g \equiv 0$ for the induced connection on $E^0_2 \to M$;
(3) The parallel transport maps $P^t_\gamma : E_{\gamma(0)} \to E_{\gamma(t)}$ along any path $\gamma(t) \in M$ satisfy $\langle P^t_\gamma(v), P^t_\gamma(w) \rangle = \langle v, w \rangle$ for all $v, w \in E_{\gamma(0)}$.

We also say in this case that ∇ is **compatible** with the bundle metric g.

EXERCISE 33.9. Show that the three conditions in Definition 33.8 really are equivalent.

The next result is sometimes called the *fundamental theorem of Riemannian geometry*:

THEOREM 33.10. *On any pseudo-Riemannian manifold (M, g), there exists a unique symmetric connection that is compatible with g.*

The connection provided by Theorem 33.10 is known as the **Levi-Cività connection** on (M, g). A proof of the theorem appeared in Lecture 22 last semester, and we will not repeat it here. That proof also produced a local coordinate formula for the Christoffel symbols of the Levi-Cività connection in terms of the components of the metric:

$$\Gamma^\ell_{ij} = \frac{1}{2} g^{k\ell} \left(\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij} \right). \tag{33.8}$$

We will henceforth *always* use the Levi-Cività connection when discussing covariant derivatives of vector or tensor fields on pseudo-Riemannian manifolds. In particular, the connection appearing in the geodesic equation will be assumed to be the Levi-Cività connection whenever there is a pseudo-Riemannian metric in the picture.

34. Geodesics in a Riemannian manifold

For most of this lecture, (M, g) is a Riemannian manifold and ∇ is its Levi-Cività connection. We will occasionally be able to generalize slightly and allow $g = \langle\ ,\ \rangle$ to be an indefinite pseudo-Riemannian metric, but for the most important results, the assumption that g is positive definite will be crucial. The main reason is that we want to use g to define distances between points, and examine the special role that geodesics play in measuring such distances. Note that, by a standard argument involving partitions of unity, every vector bundle admits a positive bundle metric, and consequently, every smooth manifold admits a Riemannian metric.

Every connected Riemannian manifold (M, g) can in turn be viewed in a natural way as a metric space: the idea is to define the **length** (*Länge*) of any smooth curve $\gamma : [a, b] \to M$ by

$$\ell(\gamma) := \int_a^b |\dot\gamma(t)|\, dt, \qquad \text{where} \qquad |X| := \sqrt{\langle X, X \rangle} \text{ for } X \in TM, \tag{34.1}$$

and then define the distance between points $p, q \in M$ as

$$\operatorname{dist}(p, q) := \inf_\gamma \ell(\gamma), \tag{34.2}$$

where the infimum is taken over all smooth paths γ from p to q. This trick does not work for indefinite pseudo-Riemannian metrics: if $\langle X, X\rangle$ is not always positive for $X \neq 0$, then g does not define a natural notion of distance. The length of a path γ can still be an interesting concept if restricted to so-called *time-like* or *space-like* paths—for which the sign of $\langle\dot\gamma,\dot\gamma\rangle$ is prescribed—but with the exception of a few remarks, we will not consider that in the present lecture.

34.1. Length-minimizing paths. A smooth path $\gamma : [a, b] \to M$ in a Riemannian manifold (M, g) with end points $\gamma(a) = p$ and $\gamma(b) = q$ is called **length-minimizing** (*minimierend*) if its length as defined in (34.1) satisfies

$$\ell(\gamma) \leqslant \ell(\hat\gamma)$$

for all other smooth paths $\hat\gamma : [a', b'] \to M$ with the same end points $\hat\gamma(a') = p$ and $\hat\gamma(b') = q$. If M is connected, one can use the distance function (34.2) to rephrase this condition as

$$\ell(\gamma) = \operatorname{dist}(p, q).$$

Two remarks are in order. First, one should not generally expect a length-minimizing path between two distinct points $p, q \in M$ to exist: the function $\gamma \mapsto \ell(\gamma)$ on the set of paths connecting them is certainly bounded from below, but it need not attain a minimum. Second, as parametrized maps, length-minimizing paths are never truly unique if they exist, since the length functional (34.1) is parametrization-invariant, i.e. one has $\ell(\gamma) = \ell(\gamma \circ \varphi)$ for any diffeomorphism $\varphi : [a', b'] \to [a, b]$. It may happen however that all paths *other than* the reparametrizations of γ have strictly larger length than γ, in which case we can say that there is a unique length-minimizing path *up to parametrization.*

If $\gamma : [a, b] \to M$ is an immersion, then it admits a distinguished class of reparametrizations, namely those which have **constant speed**

$$|\dot\gamma| \equiv \text{const}.$$

If the constant speed is 1, then one says also that γ is **parametrized by arc length** (*nach Bogenlänge parametrisiert*), because the length of γ along any segment $[t_0, t_1] \subset [a, b]$ is precisely $t_1 - t_0$.

Theorem 34.1. *In a Riemannian manifold (M, g), every length-minimizing path with constant speed is a geodesic.*

Exercise 34.2. Prove Theorem 34.1 via the following standard variational argument. Suppose γ is length-minimizing with constant speed, and $\{\gamma_s : [a, b] \to M\}_{s\in(-\epsilon,\epsilon)}$ is a smooth family of paths with fixed end points $\gamma_s(a) = p$ and $\gamma_s(b) = q$ for all s, such that $\gamma_0 = \gamma$. The assumption $\ell(\gamma) \leqslant \ell(\gamma_s)$ for all s implies $\frac{d}{ds}\ell(\gamma_s)\big|_{s=0} = 0$.

(a) Consider the **energy** (*Energie*) functional

$$E(\gamma) := \frac{1}{2}\int_a^b |\dot\gamma(t)|^2\, dt$$

for smooth paths $\gamma : [a, b] \to M$ from p to q. Show that γ satisfies the geodesic equation if and only if $\frac{d}{ds}E(\gamma_s)\big|_{s=0} = 0$ for all possible smooth families $\{\gamma_s\}$ as described above.

(b) Prove that if γ has constant speed and $\frac{d}{ds}\ell(\gamma_s)\big|_{s=0} = 0$, then $\frac{d}{ds}E(\gamma_s)\big|_{s=0}$ also vanishes.

Hint: If you get stuck, see §22.4 from last semester's course.

The argument in Exercise 34.2 also shows conversely that a compact geodesic segment is always a critical point of the length functional on paths connecting its two end points, though this does not immediately imply that it is a length-minimizing path. In general, geodesics need not even be local minima of the length functional, e.g. one can easily think up examples of *long* geodesic paths on the unit sphere $S^2 \subset \mathbb{R}^3$ that are longer than a family of nearby paths. But as we will recall in §34.4 below, compact geodesic segments are always guaranteed to be length-minimizing if they are short enough.

REMARK 34.3. One minor irritation about Theorem 34.1 is that at first glance, it only applies to *immersed* paths, since only these can be reparametrized to have constant speed. We will see however in Exercise 34.16 that this is not a loss of generality.

REMARK 34.4. Very little in this section admits straightforward adaptations to the case of an indefinite pseudo-Riemannian metric, though the result of Exercise 34.2(a) does hold in general for the (not necessarily positive) energy functional $E(\gamma) := \frac{1}{2}\int_a^b \langle\dot\gamma(t), \dot\gamma(t)\rangle\, dt$. If one restricts to so-called *time-like* or *space-like* paths for which $\langle\dot\gamma, \dot\gamma\rangle$ is required to be either positive or negative, then the length functional $\ell(\gamma)$ can also be defined (possibly after inserting a sign under the square root), and Exercise 34.2(b) then also holds and applies equally well to length-*maximizing* paths. The latter are important in general relativity, where g always has Lorentzian signature $(1, n-1)$ or $(n-1, 1)$; cf. Remark 34.17.

34.2. Injectivity radius. For any affine connection ∇ on M, consider the exponential map

$$\exp_p : \mathcal{O}_p \to M$$

restricted to a fixed point $p \in M$, where $\mathcal{O}_p \subset T_pM$ is the (necessarily open) largest neighborhood of $0 \in T_pM$ on which $\exp_p$ is defined. As an open subset of the vector space T_pM, all tangent spaces of $\mathcal{O}_p$ are canonically identified with T_pM, and since $\gamma(t) := \exp_p(tX)$ for each $X \in T_pM$ defines the unique geodesic with $\dot\gamma(0) = X$, the derivative of $\exp_p$ at $0 \in T_pM$ is then canonically identified with the identity map,

$$T_0(\exp_p) : T_0\mathcal{O}_p = T_pM \xrightarrow{\mathbb{1}} T_pM.$$

This implies via the inverse function theorem that $\exp_p$ maps any sufficiently small neighborhood of 0 in T_pM diffeomorphically onto a neighborhood of p in M.

If (M, g) is a Riemannian manifold and ∇ the Levi-Cività connection, then the observation above implies that the following number is always positive:

DEFINITION 34.5. On a Riemannian manifold (M, g), the **injectivity radius** (*Injektivitätsradius*)

$$\text{inj}(p) \in (0, \infty]$$

at a point $p \in M$ is defined as the supremum of all numbers $r > 0$ such that $\exp_p$ is well defined on $B_r(0) := \{X \in T_pM \mid |X| < r\}$ and maps it diffeomorphically onto a neighborhood of p in M. When $0 < r < \text{inj}(p)$, the set

$$\exp_p(B_r(0)) \subset M,$$

is then called the **geodesic ball of radius** r **about** p. The **injectivity radius of** (M, g) is defined as

$$\text{inj}(M, g) := \inf_{p \in M} \text{inj}(p).$$

PROPOSITION 34.6. *On any Riemannian manifold (M, g), the function $\text{inj} : M \to (0, \infty]$ is lower semicontinuous.*

PROOF. For each $p \in M$, let $B_r^p \subset T_pM$ denote the open ball of radius r about the origin, using the norm on T_pM determined by the Riemannian metric, and let $\bar B_r^p$ denote its closure. One needs to show that for any given $p \in M$ and $\epsilon > 0$, there exists a neighborhood $\mathcal{U} \subset M$ of p such that $\text{inj}(q) > \text{inj}(p) - \epsilon$ for all $q \in \mathcal{U}$. Such a neighborhood is provided by the following claim: if $\exp_p$ is a well-defined embedding on $\bar B_r^p$ for some $r > 0$, then $\exp_q$ is a well-defined embedding on $\bar B_r^q$ for all q sufficiently close to p. Since it is slightly inconvenient to consider a family of maps defined on different domains, let us instead consider a family of maps $\Phi_Y : \bar{B}^p_r \to M$ dependent on a parameter $Y \in T_pM$ lying in a small neighborhood of 0, defined by

$$\Phi_Y := \exp_{\exp_p(Y)} \circ \Psi_Y\Big|_{\bar{B}^p_r} : \bar{B}^p_r \to M,$$

where $\Psi_Y : T_pM \to T_{\exp_p(Y)}M$ denotes the parallel transport isomorphism along the path $[0,1] \to M : t \mapsto \exp_p(tY)$. This isomorphism is orthogonal since ∇ is compatible with g, thus it maps $\bar{B}^p_r$ diffeomorphically to $\bar{B}^q_r$ for $q := \exp_p(Y)$. Clearly the claim is true if the maps Φ_Y are all embeddings for $Y \in T_pM$ sufficiently close to 0, and the latter holds because for smooth maps on a compact domain, the condition of being an embedding is C^1-open; see Exercises 34.7 and 34.8 below. $\square$

EXERCISE 34.7. Suppose $K \subset \mathbb{R}^k$ is a compact set, $f_j : \mathcal{U}_j \to \mathbb{R}^n$ is a sequence of C^1-smooth maps defined on open neighborhoods $\mathcal{U}_j \subset \mathbb{R}^k$ of K, and $f : \mathcal{U} \to \mathbb{R}^n$ is another such map such that f_j and their first partial derivatives all converge uniformly on K to f and its respective first derivatives. Prove that if $f|_K : K \to \mathbb{R}^n$ is an injective immersion, then the same is true of $f_j|_K : K \to \mathbb{R}^n$ for all j sufficiently large.
Hint: Show that if $f_j(p_j) = f_j(q_j)$ for some sequences of distinct points $p_j, q_j \in K$, then $Df(p) : \mathbb{R}^k \to \mathbb{R}^n$ must have nontrivial kernel for some $p \in K$. Use the definition of the derivative.

EXERCISE 34.8. For smooth manifolds M and N, let us say that a sequence of continuously differentiable maps $f_j \in C^1(M,N)$ is C^1_{loc}**-convergent** to a map $f \in C^1(M,N)$ if for all charts $(\mathcal{U}, x)$ on M and $(\mathcal{V}, y)$ on N, the maps $y \circ f_j \circ x^{-1}$ and their first derivatives converge uniformly to $y \circ f_j \circ x^{-1}$ and its respective first derivatives on all compact subsets of their domains. Deduce from Exercise 34.7 that under this condition, if f is an embedding and M is compact, then f_j is also an embedding for all j sufficiently large.

COROLLARY 34.9. *For any compact Riemannian manifold (M,g), $\mathrm{inj}(M,g) > 0$.* $\square$

EXERCISE 34.10. Find an example of a noncompact surface $\Sigma \subset \mathbb{R}^3$ such that, if Σ is endowed with the Riemannian metric g determined by the Euclidean inner product, then all geodesics on Σ exist for all time, but $\mathrm{inj}(\Sigma, g) = 0$. Prove it with a picture.

REMARK 34.11. We will not use this, but it can be shown in fact that $\mathrm{inj} : M \to (0, \infty]$ is also upper semicontinuous, and thus continuous. In the case where (M,g) is geodesically complete, a proof of continuity may be found in [**Lee18**, Prop. 10.37]. Without any completeness assumption, [**Bou**, §10.8] carries out another proof (attributed to Stephen McKeown and John Lee).

34.3. Normal coordinates. The notion of Riemann normal coordinates was introduced last semester in §23.1; here is a quick review. Suppose p is a point in a pseudo-Riemannian manifold (M,g), and $X_1, \dots, X_n \in T_pM$ is a choice of orthonormal basis, meaning

$$\langle X_i, X_j \rangle = \pm\delta_{ij},$$

where the signs $\pm$ may vary depending on the signature of g, and in particular they are all positive if g is positive. The map

$$(t^1, \dots, t^n) \mapsto \exp_p(t^i X_i) \tag{34.3}$$

then defines a diffeomorphism from some neighborhood of 0 in $\mathbb{R}^n$ to a neighborhood of p in M, and the coordinates $(x^1, \dots, x^n)$ defined as the inverse of this map are called **Riemann normal coordinates** about p. These coordinates identify the point p with the origin in $\mathbb{R}^n$. If the metric g is positive, then the embedding (34.3) is well-defined on any open ball of radius $r < \mathrm{inj}(p)$, and the domain of the normal coordinate system is then the geodesic ball of radius r about p. More generally, so-called **geodesic normal coordinates** about a point $p \in M$ can be defined in a similar manner on any manifold equipped with an affine connection, after choosing a (not necessarily orthonormal) basis of T_pM.

EXERCISE 34.12. Assume $(x^1, \dots, x^n)$ are coordinates near $p \in M$ defined as the inverse of the map (34.3) for any choice of affine connection ∇ on M and basis $X_1, \dots, X_n$ of T_pM.

(a) Show that for any vector field $Y \in \mathfrak{X}(M)$ whose components in the chart $(x^1, \dots, x^n)$ are constant near p, $\nabla_{Y(p)} Y = 0$.

(b) Deduce via the multilinearity of $\nabla_{\partial_i + \partial_j}(\partial_i + \partial_j)$ that if the connection ∇ is symmetric, then $\nabla_i \partial_j = 0$ at p for all $i, j \in \{1, \dots, n\}$. In particular, the Christoffel symbols Γ^k_{ij} in these coordinates all vanish at p.

It follows from Exercise 34.12 that in Riemann normal coordinates, the components $g_{ij} := \langle \partial_i, \partial_j \rangle$ of the metric satisfy $g_{ij}(p) = \pm \delta_{ij}$ and $\partial_k g_{ij}(p) = \langle \nabla_k \partial_i, \partial_j \rangle + \langle \partial_i, \nabla_k \partial_j \rangle|_p = 0$, thus if we write g_{ij} as a function of the coordinates $x^1, \dots, x^n$ in $\mathcal{U}$, we have

$$g_{ij}(x^1, \dots, x^n) = \pm \delta_{ij} + O(|x|^2). \tag{34.4}$$

At this level, one sees no local distinction between any two given Riemannian metrics—they both look locally the same up to first order. We will see next week that the second-order term in the Taylor expansion of g_{ij} is determined by the curvature of (M, g), and thus cannot generally be eliminated by clever choices of coordinates.

34.4. The Gauss lemma. Here is the promised result stating that geodesics on a Riemannian manifold form unique shortest paths between sufficiently nearby points.

THEOREM 34.13. *For any point $p \in M$ in a Riemannian manifold (M, g) and $r < \mathrm{inj}(p)$, let $B_r(p)$ denote the geodesic ball of radius r about p. Then for every $q \in B_r(p)$, M contains a unique (up to parametrization) length-minimizing path from p to q, and it is a geodesic contained in $B_r(p)$.*

COROLLARY 34.14. *In any Riemann normal coordinate system $(\mathcal{U}, (x^1, \dots, x^n))$ about a point p in a Riemannian manifold (M, g), one has*

$$\mathrm{dist}(p, q) = \sqrt{[x^1(q)]^2 + \ldots + [x^n(q)]^2}$$

for every $q \in \mathcal{U}$. □

It is obvious from the definition of geodesic balls that for $q = \exp_p(X) \in B_r(p)$, the only geodesic segment $[0, 1] \to M$ from p to q contained in $B_r(p)$ is $\exp_p(tX)$. There may in general be more geodesic segments connecting p to q in M, but they cannot be contained fully in the neighborhood $B_r(p)$. That this particular geodesic is the *shortest* path from p to q follows via the Pythagorean theorem from a result known as the *Gauss lemma*, whose statement and proof we now recall:

PROPOSITION 34.15 (Gauss lemma). *Assume $p \in M$ is a point in a Riemannian manifold (M, g) and $\gamma(t) = \exp_p(tX)$ is a nonconstant geodesic based at p, defined for t in some interval $I \subset \mathbb{R}$. Then for every $t \in I$ and every vector $Y \in T_{tX}(T_pM)$ that is tangent to the sphere of radius $|tX|$ in T_pM, we have*

$$\langle \dot{\gamma}(t), T_{tX}(\exp_p)(Y) \rangle = 0.$$

In particular, for every $r \in (0, \mathrm{inj}(p))$, the image under $\exp_p$ of the sphere of radius r is a submanifold of M orthogonal to every geodesic emerging from p.

PROOF. We can assume without loss of generality $|X| = 1$. After possibly shrinking the interval $I \subset \mathbb{R}$ a bit, we can then consider a smooth map of the form

$$f : I \times (-\epsilon, \epsilon) \to M : (s, t) \mapsto \exp_p(sX(t))$$

for some small $\epsilon > 0$ and a smooth path of unit vectors $X(t) \in T_pM$ such that $X(0) = X$. The lemma follows from the claim that for any map of this form,

$$\langle \partial_s f, \partial_t f \rangle \equiv 0.$$

When $s = 0$ this is immediate, because $f(0,t) = p$ for all t and thus $\partial_t f(0,t) = 0$. Using the properties of the Levi-Cività connection and the fact that $s \mapsto f(s,t) = \exp_p(sX(t))$ is a geodesic for each fixed t, we also have

$$\partial_s \langle \partial_s f, \partial_t f \rangle = \langle \nabla_s \partial_s f, \partial_t f \rangle + \langle \partial_s f, \nabla_s \partial_t f \rangle = \langle \partial_s f, \nabla_t \partial_s f \rangle. \tag{34.5}$$

Next observe that for each fixed t, the geodesic equation implies that $\langle \partial_s f, \partial_s f \rangle$ is a constant independent of s (i.e. geodesics have constant speed), from which it follows that $\langle \partial_s f(s,t), \partial_s f(s,t) \rangle = \langle \partial_s f(0,t), \partial_s f(0,t) \rangle = \langle X(t), X(t) \rangle = 1$. This proves

$$0 = \partial_t \langle \partial_s f, \partial_s f \rangle = 2 \langle \nabla_t \partial_s f, \partial_s f \rangle,$$

so that (34.5) now vanishes, thus establishing that $\langle \partial_s f(s,t), \partial_t f(s,t) \rangle = \langle \partial_s f(0,t), \partial_t f(0,t) \rangle = 0$ for all (s,t). □

For a reminder of why the Gauss lemma implies that the distinguished geodesics in Theorem 34.13 or shorter than all other paths between the same points, see the proof of Theorem 23.5 in Lecture 23 from last semester.

Exercise 34.16. Deduce from Theorem 34.13 that in a Riemannian manifold (M,g), every length-minimizing path between two distinct points has the same image as one that is embedded. *Hint: If $\gamma : [a,b] \to M$ is a path with $\dot{\gamma} = 0$ on some compact subinterval $[t_0, t_1] \subset [a,b]$, what can you say about the shortest path from $\gamma(t_0 - \epsilon)$ to $\gamma(t_1 + \epsilon)$ for $\epsilon > 0$ small?*

Remark 34.17. The Gauss lemma admits a relatively straightforward generalization to indefinite pseudo-Riemannian manifolds (M,g), and if g has Lorentz signature $(1, n-1)$ or $(n-1, 1)$, then the proof of Theorem 34.13 can be adapted to show that sufficiently short time-like geodesics are always length-*maximizing* (not minimizing). For details, see Proposition 23.6 and Remark 23.8 from last semester's course.

34.5. Geodesic convexity. A subset $K \subset M$ in a Riemannian manifold is called **geodesically convex** if for every pair of points $p, q \in K$, there exists a unique (up to parametrization) length-minimizing path in M from p to q, and that path is contained in K. Here is a basic fact about the local geometry of Riemannian manifolds:

Theorem 34.18. *For every point p in a Riemannian manifold (M,g) and every $r \in (0, \mathrm{inj}(p))$ sufficiently small, the geodesic ball $B_r(p) \subset M$ of radius r about p is geodesically convex.*

Proof. Since the injectivity radius is a lower-semicontinuous function on M, we can find $r > 0$ small enough to ensure that $\mathrm{inj}(q) > 2r$ for every $q \in B_r(p)$. Let us also assume r is small enough so that in Riemann normal coordinates on $B_r(p)$, any geodesic passing through $B_r(p)$ satisfies the conclusions of Exercise 33.6. For any two points $x, q \in B_r(p)$, the fact (due to Corollary 34.14) that both have distance less than r from p implies $\mathrm{dist}(x,q) < 2r$, thus each is contained in a geodesic ball of radius $2r$ about the other, implying the existence of a unique length-minimizing geodesic from x to q. That this geodesic is contained in $B_r(p)$ then follows from Exercise 33.6. □

Theorem 34.18 is frequently applied in differential topology to show that all smooth manifolds admit open coverings with a particularly nice property: we say that an open covering $\{\mathcal{U}_\alpha\}_{\alpha \in I}$ of M is a **good cover** if for every finite subset $J \subset I$ such that the intersection $\bigcap_{\alpha \in J} \mathcal{U}_\alpha$ is nonempty, this intersection is also smoothly contractible.[81] Good covers are especially useful for studying the Čech cohomology of manifolds.

EXERCISE 34.19. Let us say that a subset K of a Riemannian n-manifold (M, g) is *small* if for every pair of points $p, q \in K$, p is contained in a geodesic ball about q of some radius $r < \text{inj}(q)$. Prove:

(a) The intersection of any two small subsets is also small.
(b) The intersection of any two geodesically convex subsets is also geodesically convex.
(c) Any small geodesically convex open subset is diffeomorphic to an n-dimensional star-shaped domain, i.e. a set of the form $\{r\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} \in S^{n-1},\ -f(\mathbf{x}) < r < f(\mathbf{x})\}$ for some (not necessarily continuous) function $f : S^{n-1} \to (0, \infty]$.
(d) Every n-dimensional star shaped domain is smoothly contractible.
(e) Every smooth n-manifold admits a good cover.

REMARK 34.20. For topological manifolds without a smooth structure, one can reasonably define the term "good cover" to mean that all the nonempty finite intersections are either contractible or homeomorphic to open balls. It does not seem to be known, however, whether all topological manifolds admit coverings with these properties.[82]

34.6. The Hopf-Rinow theorem. A pseudo-Riemannian manifold (M, g) is called **geodesically complete** if all of its geodesics can be defined for all time, which is equivalent to saying that the domain of the exponential map is all of TM. We proved in Lecture 23 last semester that this is true on any *compact* (not pseudo-) Riemannian manifold: this follows essentially from the fact that geodesics can be derived from the flow of a vector field on the tangent bundle TM, called the *geodesic flow*, and this flow preserves the submanifolds $\{X \in TM \mid \langle X, X\rangle = r^2\}$, which are compact if the metric is positive and M itself is compact. On the other hand, you already know examples of Riemannian manifolds that are noncompact but nonetheless geodesically complete, e.g. Euclidean space. The Hopf-Rinow theorem gives a useful characterization that also applies in the noncompact case, in addition to providing an existence (though not uniqueness) result for geodesics connecting two points that are allowed to be arbitrarily far from each other.

THEOREM 34.21 (Hopf-Rinow). *For a connected Riemannian manifold (M, g), the following conditions are equivalent:*

i (M, g) is geodesically complete;
ii (M, g) is a complete metric space with the distance function defined by (34.2);
iii For some point $p \in M$, $\exp_p$ is well defined on all of T_pM.

Moreover, if any of these conditions hold, then for every pair of points $p, q \in M$, there exists a length-minimizing geodesic segment $\gamma : [0, 1] \to M$ from $\gamma(0) = p$ to $\gamma(1) = q$.

PROOF. The implication (ii) $\Rightarrow$ (i) holds because if (M, g) is not geodesically complete, then it contains a geodesic $\gamma : (a, b) \to M$ that cannot be continued to some finite time $b \in \mathbb{R}$, and for any sequence $t_j \in (a, b)$ with $t_j \to b$, $\gamma(t_j)$ is then a Cauchy sequence in M that does not converge.

The implication (i) $\Rightarrow$ (iii) is immediate from the definition of geodesic completeness.

Before proving (iii) $\Rightarrow$ (ii), we prove the following version of the existence statement: if M is connected and $\exp_p$ is well defined on all of T_pM, then for every $q \in M$, there exists a geodesic $\gamma : [0, 1] \to M : t \mapsto \exp_p(tX)$ that ends at $\gamma(1) = q$ and satisfies $\ell(\gamma) = \text{dist}(p, q)$. To find a candidate for γ, the idea is to apply Theorem 34.13 and start with a short geodesic from p to some

[81] Various minor modifications of this definition can also be found in the literature: one could instead require that the finite intersections are topologically (but maybe not smoothly) contractible, or that they are homeomorphic or diffeomorphic to open balls.

[82] See for instance https://mathoverflow.net/questions/165850/good-covers-of-manifolds.

nearby point x, but use the distance between x and q as a means of "aiming" the geodesic so that its continuation will eventually hit q. Concretely, pick a positive number $\epsilon < \mathrm{inj}(p)/2$ such that the geodesic 2ϵ-ball $B_{2\epsilon}(p) \subset M$ about p does not contain q. The boundary of the metric ball of radius ϵ about p is then the $(n-1)$-dimensional sphere

$$S_\epsilon(p) := \left\{ \exp_p(X) \;\middle|\; X \in T_pM \text{ with } |X| = \epsilon \right\} \subset M,$$

and since this is compact, the function $S_\epsilon(p) \to (0,\infty) : x \mapsto \mathrm{dist}(x,q)$ attains a minimum at some point $x \in S_\epsilon(p)$. (Note that x might not be unique—this is why the geodesic we end up with might also be non-unique.) Any path from p to q must have length at least ϵ as it travels from p to $S_\epsilon(p)$, and at least $\mathrm{dist}(x,q)$ as it travels from $S_\epsilon(p)$ to q, implying $\mathrm{dist}(p,q) \geqslant \epsilon + \mathrm{dist}(x,q)$. But since $\mathrm{dist}(p,x) = \epsilon$ by construction, the triangle inequality turns this into an equality:

$$\mathrm{dist}(p,q) = \epsilon + \mathrm{dist}(x,q).$$

We claim now that following the shortest geodesic from p through x far enough will eventually reach q in the shortest time possible: in other words, the unique geodesic of the form

$$\gamma(t) := \exp_p(tX), \qquad |X| = 1$$

satisfying $\gamma(\epsilon) = x$ also satisfies $\gamma(\mathrm{dist}(p,q)) = q$. To see this, let

$$T := \sup \left\{ t \in [0, \mathrm{dist}(p,q)] \;\middle|\; \mathrm{dist}(p,q) = t + \mathrm{dist}(\gamma(t),q) \right\} \geqslant \epsilon.$$

Note that whenever t satisfies the relation defining this set, the triangle inequality also implies $\mathrm{dist}(p,q) \leqslant \mathrm{dist}(p,\gamma(t)) + \mathrm{dist}(\gamma(t),q) \leqslant t + \mathrm{dist}(\gamma(t),q)$, thus both inequalities must be equalities and it follows that $\mathrm{dist}(p,\gamma(t)) = t$. Moreover, since the distance function is continuous, the set in question is closed, implying that T itself satisfies

$$\mathrm{dist}(p,q) = T + \mathrm{dist}(\gamma(T),q) \qquad \text{and} \qquad \mathrm{dist}(p,\gamma(T)) = T. \tag{34.6}$$

We claim that $T = \mathrm{dist}(p,q)$, in which case (34.6) will imply $\gamma(\mathrm{dist}(p,q)) = q$. Arguing by contradiction, suppose $T < \mathrm{dist}(p,q)$, and repeat the previous step with p replaced by the point $p_1 := \gamma(T) \neq q$, i.e. for $\epsilon_1 > 0$ sufficiently small, choose

$$x_1 \in S_{\epsilon_1}(p_1)$$

to minimize the distance from $S_{\epsilon_1}(p_1)$ to q. It follows that

$$\mathrm{dist}(p_1,x_1) = \epsilon_1 \qquad \text{and} \qquad \mathrm{dist}(p_1,q) = \epsilon_1 + \mathrm{dist}(x_1,q),$$

thus

$$\mathrm{dist}(p,q) = T + \mathrm{dist}(p_1,q) = T + \epsilon_1 + \mathrm{dist}(x_1,q). \tag{34.7}$$

Here, we can interpret $T + \epsilon_1$ as the length of the piecewise smooth path that follows the geodesic γ from p to p_1 and then follows a possibly different geodesic from p_1 to x_1. It is easy to see however that if these two geodesics are actually *different*, i.e. if their intersection at p_1 is not tangential, then the path from p to x_1 can be made strictly shorter: indeed, pick a point p_2 along γ close to p_1 and deduce from Theorem 34.13 that the shortest path from p_2 to some point on the geodesic between p_1 and x must be a *smooth* geodesic, not just piecewise smooth. This would imply $\mathrm{dist}(p,x_1) < T + \epsilon_1$ and thus contradict the triangle inequality when combined with (34.7):

$$\mathrm{dist}(p,q) \leqslant \mathrm{dist}(p,x_1) + \mathrm{dist}(x_1,q) < T + \epsilon_1 + \mathrm{dist}(x_1,q) = \mathrm{dist}(p,q).$$

It follows that the geodesic from p_1 to x_1 is actually just a continuation of our original geodesic γ, so $x_1 = \gamma(T + \epsilon_1)$, and (34.7) thus becomes

$$\mathrm{dist}(p,q) = (T + \epsilon_1) + \mathrm{dist}(\gamma(T + \epsilon_1), q),$$

contradicting the definition of T. This establishes the existence of a length-minimizing geodesic from p to q.

The proof that (iii) $\Rightarrow$ (ii) now goes as follows: given a Cauchy sequence $q_j \in M$, we can choose $X_j \in T_pM$ such that $[0,1] \to M : t \mapsto \exp_p(tX_j)$ is a length-minimizing geodesic from p to q_j for each j. Since Cauchy sequences are bounded, X_j is now also a bounded sequence and thus has a subsequence convergent to some $X \in T_pM$, implying that the corresponding subsequence of q_j converges to $q := \exp_p(X)$. The Cauchy condition now implies that the entire sequence also converges to q, so (M, g) is a complete metric space. □

REMARK 34.22. For an indefinite pseudo-Riemannian manifold (M, g), the statement of the Hopf-Rinow theorem does not immediately make sense since g does not naturally define a metric space structure on M. One might still wonder however whether there is a similar existence result for geodesics: if (M, g) is geodesically complete and connected, does it follow that any two points in M can be connected by a geodesic? Remarkably, the answer is no: an explicit counterexample is furnished by the *de Sitter spacetime*, a geodesically complete n-dimensional Lorentzian hyperboloid $H^n_L \cong \mathbb{R} \times S^{n-1}$ that appeared in Exercise 31.22 at the end of last semester's notes, and is connected for $n \geqslant 3$. For a more detailed discussion, see [**Bär**, §1.2]. Less naive adaptations of the Hopf-Rinow theorem to the Lorentzian setting are nonetheless possible under some technical assumptions, e.g. showing that whenever two points are connected by a smooth time-like path, they are also connected by a time-like geodesic (see [**Sei67**, **Sán01**]).

EXERCISE 34.23. Show that a Riemannian manifold (M, g) with $\text{inj}(M, g) > 0$ must be geodesically complete.

35. The Riemann curvature tensor

In this lecture we examine various ways of measuring curvature on a Riemannian or pseudo-Riemannian manifold.

35.1. Definition of the Riemann tensor. Given any vector bundle $E \to M$ with a connection ∇, one defines the **Riemann curvature tensor** (*Riemannscher Krümmungstensor*) as the unique smooth multilinear bundle map

$$R : TM \oplus TM \oplus E \to E : (X, Y, v) \mapsto R(X, Y)v$$

satisfying the relation

$$R(X, Y)s = \nabla_X \nabla_Y s - \nabla_Y \nabla_X s - \nabla_{[X,Y]} s$$

for all $X, Y \in \mathfrak{X}(M)$ and $s \in \Gamma(E)$. It is straightforward to check that this last expression is C^∞-linear with respect to each of its three arguments. In local coordinates, writing $R^a{}_{jkb} := (R(\partial_j, \partial_k)e_b)^a$, one derives from (32.5) a formula for the Riemann tensor in terms of the Christoffel symbols,

$$R^a{}_{jkb} = \partial_j \Gamma^a_{kb} - \partial_k \Gamma^a_{jb} + \Gamma^a_{jc}\Gamma^c_{kb} - \Gamma^a_{kc}\Gamma^c_{jb}.$$

The Riemann tensor often arises in calculations involving the following scenario. Given a vector bundle $E \to M$ and a smooth map

$$\mathbb{R}^2 \overset{\text{open}}{\supset} \mathcal{V} \to M : (s, t) \mapsto f(s, t),$$

we can consider the covariant derivatives with respect to s and t of a section $\eta \in \Gamma(f^*E)$ along f. If one temporarily adds the assumption that f is an embedding, then after restricting $\mathcal{V}$ to a sufficiently small neighborhood of one point $(s_0, t_0) \in \mathcal{V}$, there exists an extension of f to a diffeomorphism $\hat{f} : \hat{\mathcal{V}} \to M$ onto an open neighborhood of $f(s_0, t_0)$ in M, where $\hat{\mathcal{V}}$ is an open neighborhood of $\mathcal{V} \times \{0\}$ in $\mathbb{R}^n$ and $\hat{f}|_{\mathcal{V}\times\{0\}} = f$; in this situation the inverse of $\hat{f}$ can be viewed as a chart for which the vector fields $\partial_s f$ and $\partial_t f$ along f are just the restrictions to the submanifold $f(\mathcal{V}) \subset M$ of the coordinate vector fields ∂_1 and ∂_2 respectively. One can now also extend $\eta \in \Gamma(f^*E)$ to a section $\widehat{\eta} \in \Gamma(E)$ that matches η along $f(\mathcal{V})$, so $R(\partial_s f, \partial_t f)\eta$ is just $R(\partial_1, \partial_2)\widehat{\eta}$ restricted to the submanifold $f(\mathcal{V}) \subset M$. At any point on $f(\mathcal{V})$, the definition of the Riemann tensor then gives

$$R(\partial_s f, \partial_t f)\eta = R(\partial_1, \partial_2)\widehat{\eta} = \nabla_1\nabla_2\widehat{\eta} - \nabla_2\nabla_1\widehat{\eta} = \nabla_s\nabla_t\eta - \nabla_t\nabla_s\eta,$$

where the Lie bracket term does not appear since coordinate vector fields commute with each other. The result is the useful relation

$$\nabla_s\nabla_t\eta - \nabla_t\nabla_s\eta = R(\partial_s f, \partial_t f)\eta, \tag{35.1}$$

which is reminiscent of the relation (33.6) that holds for the covariant derivatives of the vector fields $\partial_s f, \partial_t f \in \Gamma(f^*TM)$. Just like that relation, we claim that (35.1) also holds for any smooth map $f : \mathcal{V} \to M$ and section $\eta \in \Gamma(f^*E)$ along f, without needing to assume that f is an embedding. There are at least two possible ways to see this: if $\dim M \geqslant 2$, then one can always perturb f near any given point $(s_0, t_0) \in \mathcal{V}$ to make it an embedding near that point, in which case our proof above for (35.1) works, and if the relation is valid under such perturbations, then continuity implies that it must also have been valid beforehand. (The perturbation trick is impossible when $\dim M = 1$, but for this case, see Exercise 35.1 below.) Alternatively, one can verify (35.1) directly by writing down both sides in local coordinates, as we did for (33.6).

Exercise 35.1. Show that for any connection ∇ on a vector bundle $E \to M$ whose base M is 1-dimensional, the Riemann tensor vanishes and Equation (35.1) is also valid, meaning in this case that $\nabla_s\nabla_t\eta = \nabla_t\nabla_s\eta$ for all sections $\eta \in \Gamma(f^*E)$ along a map $\mathcal{V} \to M : (s,t) \mapsto f(s,t)$. *Hint: Near any given point in $\mathcal{V}$, you can write η in the form $\eta(s,t) = \eta^i(s,t)e_i(f(s,t))$ for a set of scalar-valued functions $\eta^i(s,t)$ and parallel sections e_i of E. (This is not possible for arbitrary connections on arbitrary vector bundles, but is always possible when $\dim M = 1$.)*

Theorems 35.2 and 35.3 below are the most important results about the Riemann tensor, though they will not yet play a serious role until later in this course. A connection ∇ on $E \to M$ is called **flat** (*flach*) if for every $p \in M$ and $v \in E_p$, there exists a neighborhood $\mathcal{U} \subset M$ of p and a parallel section $s \in \Gamma(E|_{\mathcal{U}})$ such that $s(p) = v$. In this case one can always form local frames out of parallel sections, so flatness is equivalent to saying that ∇ looks like the *trivial* connection in some local trivialization defined near any given point. Similarly, a pseudo-Riemannian metric g on M is called **locally flat** (*lokal flach*) if every point $p \in M$ has a neighborhood $\mathcal{U} \subset M$ admitting a chart $(x^1, \ldots, x^n) : \mathcal{U} \to \mathbb{R}^n$ such that $g = \pm(dx^1)^2 \pm \ldots \pm (dx^n)^2$ on $\mathcal{U}$, where as usual the signs in front of the individual terms may vary depending on the signature of g, but are positive if g is positive.

Theorem 35.2. *A connection on a vector bundle is flat if and only if its Riemann tensor vanishes identically.* □

Theorem 35.3. *A pseudo-Riemannian metric g on a manifold M is locally flat if and only if its Levi-Cività connection is flat.* □

A proof of Theorem 35.2 appeared last semester in §26.3, where it followed mainly from two ingredients: (1) a calculation identifying the Riemann tensor with a different object defined in terms of Lie brackets of horizontal lifts, called the *curvature* 2*-form*, and (2) the Frobenius integrability theorem, which implies in the situation at hand that the horizontal subbundle $HE \subset TE$ defined by the connection is integrable if and only if the curvature 2-form vanishes. We will give a more general version of this proof when we discuss fiber bundles later in the course. Given this result, Theorem 35.3 is a relatively easy corollary, and we will also prove a more general version of it involving Riemannian manifolds of constant sectional curvature later in this semester.

EXERCISE 35.4. Show that any flat connection on a vector bundle $E \to M$ determines a natural flat structure (see Definition 32.16) on $E \to M$, for which the flat sections on open subsets of M are precisely the parallel sections. Conversely, show that any flat vector bundle over a smooth manifold carries a natural flat connection with the same property.

35.2. Symmetries of the Riemann tensor. On the tangent bundle of an n-manifold M, the Riemann tensor is determined in local coordinates by n^4 components, which is a large number even in the simplest nontrivial case, i.e. when $n = 2$. In order to make the information carried by these components seem more manageable, it will be useful to be aware of certain nontrivial relations that they satisfy. One of them is immediate from the definition: for every connection on every vector bundle $E \to M$ we have

$$R(X,Y)v + R(Y,X)v = 0 \qquad \text{for all } X,Y \in T_pM,\ v \in E_p,\ p \in M, \tag{35.2}$$

which translates into local coordinates as the relation

$$R^a{}_{ijb} + R^a{}_{jib} = 0. \tag{35.3}$$

Further relations hold if we impose extra conditions on the connection, conditions that will all be satisfied in the case of the Levi-Cività connection on a pseudo-Riemannian manifold.

PROPOSITION 35.5. *If the connection ∇ is compatible with a bundle metric $\langle\ ,\ \rangle$ on $E \to M$, then for every $X, Y \in T_pM$ at a point $p \in M$, the linear map $R(X,Y) : E_p \to E_p$ is antisymmetric with respect to $\langle\ ,\ \rangle$, i.e. we have*

$$\langle R(X,Y)v, w\rangle + \langle v, R(X,Y)w\rangle = 0 \qquad \text{for all } X,Y \in T_pM,\ v,w \in E_p,\ p \in M. \tag{35.4}$$

PROOF. For any two vector fields $X, Y \in \mathfrak{X}(M)$, the differential operator $D := \mathcal{L}_X\mathcal{L}_Y - \mathcal{L}_Y\mathcal{L}_X - \mathcal{L}_{[X,Y]}$ on $C^\infty(M,\mathbb{F})$ is zero by the definition of the Lie bracket. Given two sections $v, w \in \Gamma(E)$, apply this operator to the function $\langle v, w\rangle : M \to \mathbb{F}$, use the compatibility of ∇ with the metric, plug in the definition of the Riemann tensor and cancel all terms that can be cancelled: the result is (35.4). ☐

For an affine connection ∇ on M, the Riemann tensor is a multilinear map $TM \oplus TM \oplus TM \to TM$ and can thus be regarded as a type $(1,3)$ tensor field,

$$R \in \Gamma(T_3^1 M).$$

PROPOSITION 35.6 (First Bianchi identity). *For any symmetric affine connection ∇ on M, the Riemann tensor satisfies*

$$R(X,Y)Z + R(Y,Z)X + R(Z,X)Y = 0 \qquad \textit{for all } X,Y,Z \in \mathfrak{X}(M).$$

PROOF. We calculate, using the definitions of R and the symmetry relation $\nabla_X Y - \nabla_Y X = [X,Y]$:

$$\begin{aligned}
R(X,Y)Z + R(Y,Z)X + R(Z,X)Y &= \nabla_X\nabla_Y Z - \nabla_Y\nabla_X Z + \nabla_Y\nabla_Z X - \nabla_Z\nabla_Y X \\
&\quad + \nabla_Z\nabla_X Y - \nabla_X\nabla_Z Y - \nabla_{[X,Y]}Z - \nabla_{[Y,Z]}X - \nabla_{[Z,X]}Y \\
&= \nabla_X\left(\nabla_Y Z - \nabla_Z Y\right) + \nabla_Y\left(\nabla_Z X - \nabla_X Z\right) + \nabla_Z\left(\nabla_X Y - \nabla_Y X\right) \\
&\quad - \nabla_{[X,Y]}Z - \nabla_{[Y,Z]}X - \nabla_{[Z,X]}Y \\
&= \nabla_X[Y,Z] - \nabla_{[Y,Z]}X + \nabla_Y[Z,X] - \nabla_{[Z,X]}Y + \nabla_Z[X,Y] - \nabla_{[X,Y]}Z \\
&= [X,[Y,Z]] + [Y,[Z,X]] + [Z,[X,Y]].
\end{aligned}$$

This last term vanishes by the Jacobi identity for the Lie bracket. ☐

Assume now that (M, g) is a pseudo-Riemannian manifold and ∇ is the Levi-Cività connection. Writing $\langle\ ,\ \rangle := g$ for the bundle metric on TM, it is often convenient in this setting to replace $R \in \Gamma(T_3^1 M)$ with its **fully covariant** version

$$\mathrm{Riem} \in \Gamma(T_4^0 M), \qquad \mathrm{Riem}(V, X, Y, Z) := \langle V, R(X, Y)Z\rangle,$$

which carries all the same information as R since nondegeneracy of the bundle metric means that $TM \to T^*M : V \mapsto V_\flat := \langle V, \cdot\rangle$ is a bundle isomorphism. In local coordinates $(x^1, \ldots, x^n)$, it is conventional to write the components of Riem with the same symbol as the Riemann tensor but four lower indices, thus

$$R_{ijk\ell} := \mathrm{Riem}(\partial_i, \partial_j, \partial_k, \partial_\ell) = \langle \partial_i, R(\partial_j, \partial_k)\partial_\ell\rangle = \langle \partial_i, R^m{}_{jk\ell}\partial_m\rangle = g_{im} R^m{}_{jk\ell}.$$

The next result assembles all the symmetry properties of R proved above into a statement about Riem.

THEOREM 35.7. *On any pseudo-Riemannian manifold (M, g), the covariant Riemann tensor* $\mathrm{Riem} \in \Gamma(T_4^0 M)$ *satisfies the following relations for all $p \in M$ and $V, X, Y, Z \in T_pM$:*

i $\mathrm{Riem}(V, X, Y, Z) + \mathrm{Riem}(V, Y, X, Z) = 0$
ii $\mathrm{Riem}(V, X, Y, Z) + \mathrm{Riem}(Z, X, Y, V) = 0$
iii $\mathrm{Riem}(V, X, Y, Z) + \mathrm{Riem}(V, Y, Z, X) + \mathrm{Riem}(V, Z, X, Y) = 0$ *(first Bianchi identity)*
iv $\mathrm{Riem}(V, X, Y, Z) = \mathrm{Riem}(Y, Z, V, X)$ *(interchange symmetry)*

PROOF. Properties (i), (ii) and (iii) follow from (35.2) and Propositions 35.5 and 35.6 respectively. We claim that these three properties imply property (iv). The main idea of the proof is, well—cleverness:

$$\begin{aligned}
\mathrm{Riem}(V, X, Y, Z) &\overset{\text{(ii)}}{=} -\mathrm{Riem}(Z, X, Y, V) \overset{\text{(iii)}}{=} \mathrm{Riem}(Z, Y, V, X) + \mathrm{Riem}(Z, V, X, Y) \\
&\overset{\text{(ii)}}{=} -\mathrm{Riem}(X, Y, V, Z) - \mathrm{Riem}(Y, V, X, Z) \\
&\overset{\text{(iii)}}{=} \mathrm{Riem}(X, V, Z, Y) + \mathrm{Riem}(X, Z, Y, V) + \mathrm{Riem}(Y, X, Z, V) + \mathrm{Riem}(Y, Z, V, X) \\
&\overset{\text{(i)+(ii)}}{=} 2\mathrm{Riem}(Y, Z, V, X) - \mathrm{Riem}(V, Z, Y, X) - \mathrm{Riem}(V, X, Z, Y) \\
&\overset{\text{(iii)}}{=} 2\mathrm{Riem}(Y, Z, V, X) + \mathrm{Riem}(V, Y, X, Z) \overset{\text{(i)}}{=} 2\mathrm{Riem}(Y, Z, V, X) - \mathrm{Riem}(V, X, Y, Z).
\end{aligned}$$

□

In local coordinates, relations (i), (ii) and (iv) in Theorem 35.7 become

$$R_{ijk\ell} = -R_{\ell jki} = -R_{ikj\ell} = R_{k\ell ij}$$

and the Bianchi identity (iii) becomes

$$R_{ijk\ell} + R_{ik\ell j} + R_{i\ell jk} = 0.$$

EXERCISE 35.8. Show that in the case $\dim M = 2$, the relations (iii) and (iv) in Theorem 35.7 are redundant, i.e. they follow from (i) and (ii).
Hint: Use (i) and (ii) to show that in local coordinates, all components $R_{ijk\ell}$ are determined by R_{1122}.

35.3. Gaussian curvature. Exercise 35.8 shows that in local coordinates, the Riemann tensor on a pseudo-Riemannian 2-manifold is determined by one of its components, the real-valued function R_{1122}. We shall now derive a global coordinate-invariant version of this statement, under the simplifying assumption that the metric g is positive. (This assumption is not essential; for a discussion of the case with signature $(1, 1)$, see §31.4 at the end of last semester's notes.)

Assuming the bundle metric $g = \langle\ ,\ \rangle$ is positive, let

$$\mathrm{End}^{\mathrm{anti}}(TM) \subset \mathrm{End}(TM)$$

denote the subbundle of $\mathrm{End}(TM) = \mathrm{Hom}(TM, TM) = T_1^1 M$ whose fiber over $p \in M$ is the space of linear maps $T_pM \to T_pM$ that are antisymmetric with respect to the inner product $\langle\ ,\ \rangle$. By Theorem 35.7(i) and (ii), R can be viewed as a section of the vector bundle

$$\Lambda^2 T^*M \otimes \mathrm{End}^{\mathrm{anti}}(TM),$$

whose fibers are canonically isomorphic to the spaces of antisymmetric bilinear maps $T_pM \times T_pM \to \mathrm{End}^{\mathrm{anti}}(T_pM)$. In the case $\dim M = 2$, both $\Lambda^2 T^*M$ and $\mathrm{End}^{\mathrm{anti}}(TM)$ are line bundles, and so therefore is $\Lambda^2 T^*M \otimes \mathrm{End}^{\mathrm{anti}}(TM)$. In fact, if we assume additionally that M is oriented, then both even come with canonical frames determined by the orientation and the metric. For $\Lambda^2 T^*M$, this is the **Riemannian volume form**

$$d\mathrm{vol} \in \Omega^2(M),$$

uniquely determined by the condition that $d\mathrm{vol}(X, Y) = 1$ for every positively-oriented orthonormal basis (X, Y) of each tangent space T_pM. For $\mathrm{End}^{\mathrm{anti}}(TM)$, we define the unique tensor field $J \in \Gamma(T_1^1 M) = \mathrm{End}(TM)$ such that for each $p \in M$, $J_p : T_pM \to T_pM$ is the 90-degree counterclockwise rotation, where "counterclockwise" means that (X, JX) is a positively-oriented basis of T_pM whenever $X \in T_pM$ is nonzero. This operator satisfies

$$J^2 = -\mathbb{1} \qquad \text{and} \qquad \langle JX, JY\rangle = \langle X, Y\rangle$$

for all $X, Y \in T_pM$ and $p \in M$, and is therefore antisymmetric:

$$\langle JX, Y\rangle = \langle J^2X, JY\rangle = -\langle X, JY\rangle.$$

There is a simple relationship between $d\mathrm{vol}$ and J: since the bilinear form $(X, Y) \mapsto \langle JX, Y\rangle$ is antisymmetric, it is a scalar multiple of $d\mathrm{vol}$ at every point, and plugging in a basis $(X, Y) := (X, JX)$ with $|X| = 1$ reveals that the constant of proportionality is 1, so

$$d\mathrm{vol}(X, Y) = \langle JX, Y\rangle \qquad \text{for all } X, Y \in T_pM,\ p \in M. \tag{35.5}$$

More importantly, $d\mathrm{vol} \otimes J$ defines a global frame for $\Lambda^2 T^*M \otimes \mathrm{End}^{\mathrm{anti}}(TM)$, and even better, this frame does not depend on the choice of orientation for M, as reversing the orientation puts a sign in front of both $d\mathrm{vol}$ and J, so these two signs cancel out. For this reason, the following definition does not require M to be orientable, but requires only a choice of *local* orientation on a neighborhood of any given point, so that $d\mathrm{vol}$ and J can be defined on such a neighborhood—what matters in the definition is their product, which does not depend on the choice.

DEFINITION 35.9. For a Riemannian 2-manifold (M, g), the **Gaussian curvature** (*Gaußkrümmung*) is the unique function $K_G : M \to \mathbb{R}$ such that for any choice of orientation on a neighborhood of each point $p \in M$,

$$R(X, Y)Z = -K_G(p)\, d\mathrm{vol}(X, Y) JZ \qquad \text{for all } X, Y, Z \in T_pM.$$

For surfaces embedded in Euclidean $\mathbb{R}^3$, the Gaussian curvature has an elegant and geometrically intuitive interpretation that we do not have space to review here, but I highly recommend reading §27.3 in the notes from the first semester if you have not seen it before.

For computational purposes, recall that on any real oriented vector space V with an inner product and a positively-oriented orthonormal basis $e_1, \ldots, e_n \in V$, its dual basis $e_*^1, \ldots, e_*^n \in V^*$ can be used to write down the canonical volume form

$$\mu := e_*^1 \wedge \ldots \wedge e_*^n \in \Lambda^n V^*,$$

which satisfies $\mu(v_1, \ldots, v_n) = 1$ whenever $v_1, \ldots, v_n \in V$ is a positively-oriented orthonormal basis. Now for any $v_1, \ldots, v_n \in V$, not necessarily linearly independent or orthonormal, the result of Exercise 11.12 gives

$$\mu(v_1, \ldots, v_n) = \sqrt{\det \begin{pmatrix} \langle v_1, v_1 \rangle & \cdots & \langle v_1, v_n \rangle \\ \vdots & \ddots & \vdots \\ \langle v_n, v_1 \rangle & \cdots & \langle v_n, v_n \rangle \end{pmatrix}},$$

where the determinant vanishes if $v_1, \ldots, v_n$ are linearly dependent and is otherwise positive since the symmetric bilinear form $\langle \ , \ \rangle$ is positive definite. We can use this and (35.5) to derive a formula for K_G in terms of the Riemann tensor: for any $X, Y \in T_pM$, we have

$$\begin{aligned} \operatorname{Riem}(X, X, Y, Y) &= \langle X, R(X, Y)Y \rangle = -K_G(p)\langle X, \operatorname{dvol}(X, Y)JY \rangle = K_G(p) \operatorname{dvol}(X, Y)\langle JX, Y \rangle \\ &= K_G(p) \left|\operatorname{dvol}(X, Y)\right|^2 = K_G(p) \cdot \det \begin{pmatrix} \langle X, X \rangle & \langle X, Y \rangle \\ \langle Y, X \rangle & \langle Y, Y \rangle \end{pmatrix} \\ &= K_G(p) \cdot \left(\langle X, X \rangle \langle Y, Y \rangle - \langle X, Y \rangle^2 \right). \end{aligned}$$

Whenever X and Y are linearly independent, the determinant in this last expression is positive, so we conclude

$$K_G(p) = \frac{\operatorname{Riem}(X, X, Y, Y)}{\langle X, X \rangle \langle Y, Y \rangle - \langle X, Y \rangle^2}, \tag{35.6}$$

or in local coordinates,

$$K_G = \frac{R_{1122}}{g_{11}g_{22} - g_{12}^2}.$$

REMARK 35.10. The formula (35.6) can also be taken as a definition of the Gaussian curvature $K_G : M \to \mathbb{R}$ for pseudo-Riemannian 2-manifolds (M, g) of arbitrary signature, though in the indefinite case, a slightly different argument is required for showing that $K_G(p)$ does not depend on the choice of basis $X, Y \in T_pM$ (see §31.4).

35.4. Sectional curvature. The next definition provides a way of measuring the same geometric information as Gaussian curvature when the dimension is greater than 2. We continue under the assumption that g is positive, though as explained in §31.4, this condition can be lifted with a bit of care.

DEFINITION 35.11. Assume (M, g) is a Riemannian manifold with $\dim M \geqslant 2$, $p \in M$ is a point and $P \subset T_pM$ is a 2-dimensional linear subspace. The **sectional curvature** (*Schnittkrümung*) of (M, g) **along** P, denoted by

$$K_S(P) \in \mathbb{R},$$

is defined as the Gaussian curvature at p of the embedded surface $\Sigma_P \subset M$ defined by

$$\Sigma_P := \exp_p(\mathcal{O} \cap P) \subset M,$$

where $\mathcal{O} \subset T_pM$ is a neighborhood of 0 on which the map $\mathcal{O} \xrightarrow{\exp_p} M$ is an embedding, and the Riemannian metric on Σ_P is the restriction of g.

PROPOSITION 35.12. *The sectional curvature of (M, g) along $P \subset T_pM$ satisfies*

$$K_S(P) = \frac{\operatorname{Riem}(X, X, Y, Y)}{\langle X, X \rangle \langle Y, Y \rangle - \langle X, Y \rangle^2}$$

for every linearly-independent pair of tangent vectors $X, Y \in P$.

Proof. This follows from (35.6) and the fact that the Riemann tensor $\widehat{R} \in \Gamma(T^1_3\Sigma_P)$ of Σ_P at p is simply the restriction to $T_p\Sigma_P = P \subset T_pM$ of the Riemann tensor $R \in \Gamma(T^1_3M)$ of (M, g). The latter follows from a general result about Riemannian submanifolds, the *Gauss equation* (see Prop. 28.5 from the first semester),

$$\widehat{\mathrm{Riem}}(V, X, Y, Z) = \mathrm{Riem}(V, X, Y, Z) + \langle \mathrm{II}(V, X), \mathrm{II}(Y, Z)\rangle - \langle \mathrm{II}(V, Y), \mathrm{II}(X, Z)\rangle,$$

where $\widehat{\mathrm{Riem}} \in \Gamma(T^0_4\Sigma_P)$ denotes the fully covariant version of $\widehat{R}$ and $\mathrm{II} : T\Sigma_P \oplus T\Sigma_P \to (T\Sigma_P)^\perp$ is the *second fundamental form*, defined so that $\mathrm{II}(X, Y)$ is the part of $\nabla_X Y \in \Gamma(TM|_{\Sigma_P})$ orthogonal to $T\Sigma_P$ for any $X, Y \in \mathfrak{X}(\Sigma_P)$. At the point $p \in \Sigma_P$, the second fundamental form vanishes for the following reason: for any $X \in T_p\Sigma_P$, the path $\gamma(t) = \exp_p(tX)$ for t close to 0 is a geodesic of (M, g) that is also contained in Σ_P, implying that X can be extended to a vector field on Σ_P near p that matches $\dot\gamma$ along γ and therefore satisfies $\nabla_X X = 0$ at p. This implies

$$\mathrm{II}(X, X) = 0$$

for all $X \in T_p\Sigma_P$. Expanding $\mathrm{II}(X + Y, X + Y)$ for any $X, Y \in T_p\Sigma_P$ and using the fact that II is symmetric then proves $\mathrm{II}(X, Y) = 0$. □

One can think of sectional curvature as a real-valued function

$$K_S : \mathrm{Gr}_2(TM) \to \mathbb{R},$$

where the **Grassmannian** of 2-planes in TM is defined as the set

$$\mathrm{Gr}_2(TM) := \bigcup_{p\in M} \mathrm{Gr}_2(T_pM), \qquad \mathrm{Gr}_2(T_pM) := \{\text{2-dimensional subspaces } P \subset T_pM\}.$$

We will see later that $\mathrm{Gr}_2(TM)$ inherits a natural smooth manifold structure from that of M; in fact it is an important example of a *smooth fiber bundle*, whose fiber over each point p is the compact smooth manifold $\mathrm{Gr}_2(T_pM)$. We say that a Riemannian manifold (M, g) has **positive/negative** sectional curvature if the values of the function $K_S : \mathrm{Gr}_2(TM) \to \mathbb{R}$ are everywhere positive/negative, and the notions of **nonpositive/nonnegative** or **vanishing** sectional curvature are defined similarly. The ability to state such definitions and prove theorems about them is one of the principal advantages of the notion of sectional curvature in comparison with the Riemann tensor. The following result shows however that, secretly, the function $K_S : \mathrm{Gr}_2(TM) \to \mathbb{R}$ and tensor $R \in \Gamma(T^1_3M)$ are completely equivalent objects.

Theorem 35.13. *On any Riemannian manifold (M, g), the Riemann tensor $R_p : T_pM \times T_pM \times T_pM \to T_pM$ at a point $p \in M$ is determined by the values of the sectional curvature function $K_S(P)$ on the set of all 2-planes $P \in \mathrm{Gr}_2(T_pM)$ at p. In particular, the tensor field $R \in \Gamma(T^1_3M)$ vanishes if and only if the function $K_S : \mathrm{Gr}_2(TM) \to \mathbb{R}$ vanishes.*

Proof. By Proposition 35.12, the main thing we need to show here is that if the values $\mathrm{Riem}(X, X, Y, Y) \in \mathbb{R}$ are known for every pair $X, Y \in T_pM$, then these determine $\mathrm{Riem}(V, X, Y, Z)$ for all tuples $V, X, Y, Z \in T_pM$, and in particular the latter will always vanish if $\mathrm{Riem}(X, X, Y, Y)$ always vanishes. The proof is a purely algebraic argument based on the symmetries listed in Theorem 35.7.

Suppose $\mathrm{Riem}, \widehat{\mathrm{Riem}} \in \Gamma(T^0_4M)$ are two tensor fields that both satisfy the relations in Theorem 35.7 and satisfy $\mathrm{Riem}(X, X, Y, Y) = \widehat{\mathrm{Riem}}(X, X, Y, Y)$ for all pairs X, Y. Then $D := \widehat{\mathrm{Riem}} - \mathrm{Riem} \in \Gamma(T^0_4M)$ also satisfies the relations in Theorem 35.7, and additionally

$$D(X, X, Y, Y) = 0 \qquad \text{for all } X, Y \in T_pM,\ p \in M. \tag{35.7}$$

We claim that these assumptions imply $D \equiv 0$. As a first step, exploiting the symmetries in Theorem 35.7 gives

$$D(V, X, Y, Y) \overset{\text{(iv)}}{=} D(Y, Y, V, X) \overset{\text{(i)+(ii)}}{=} D(X, V, Y, Y)$$

for any three vectors $V, X, Y \in T_pM$ at a point $p \in M$, and using (35.7) to eliminate terms that are trivial, multilinearity then implies

$$0 = D(V + X, V + X, Y, Y) = D(V, X, Y, Y) + D(X, V, Y, Y) = 2D(V, X, Y, Y).$$

Choosing a fourth vector $Z \in T_pM$, the latter implies

$$0 = D(V, X, Y + Z, Y + Z) = D(V, X, Y, Z) + D(V, X, Z, Y),$$

so $D(V, X, Y, Z)$ is antisymmetric under the interchange of Y and Z. It is also antisymmetric under the interchange of X and Y by Theorem 35.7(i), so applying the Bianchi identity (iii), we conclude

$$0 = D(V, X, Y, Z) + D(V, Y, Z, X) + D(V, Z, X, Y) = 3D(V, X, Y, Z).$$

□

REMARK 35.14. The formula in Proposition 35.12 can be taken as a general definition of sectional curvature for pseudo-Riemannian manifolds of arbitrary signature, but in addition to the need for an additional argument to prove independence of the basis $X, Y \in P$, there is a further caveat that does not arise in the Riemannian setting: if $\langle \, , \, \rangle$ is nondegenerate but not positive, then the denominator $\langle X, X\rangle\langle Y, Y\rangle - \langle X, Y\rangle^2$ can vanish. This reflects the fact that g does not have a nondegenerate restriction to arbitrary 2-dimensional submanifolds $\Sigma_P \subset M$, i.e. they are not all *pseudo-Riemannian submanifolds*, and without nondegeneracy, the Gaussian curvature of Σ_P cannot be defined. For this reason, K_S is only defined on the open and dense subset of $\mathrm{Gr}_2(TM)$ consisting of planes $P \subset T_pM$ on which the bundle metric is nondegenerate. For more details, see §31.4.

36. Jacobi, Ricci and Cartan-Hadamard

36.1. Ricci and scalar curvature. In addition to the Gaussian and sectional curvatures discussed in the previous lecture, there are other ways of reducing the Riemann tensor $R \in \Gamma(T^1_3M)$ of a Riemannian manifold (M, g) to a more manageable set of information. The object we define next turns out to play a large role in many deep theorems about smooth manifolds, including Perelman's solution to the Poincarè conjecture. In the setting of pseudo-Riemannian manifolds with Lorentz signature $(1, n-1)$ or $(n-1, 1)$, it is also one of the key ingredients in Einstein's equation, which describes the evolution of the metric representing gravitation on the spacetime manifold of General Relativity.

DEFINITION 36.1. The **Ricci curvature** (*Ricci-Krümmung*) of a pseudo-Riemannian manifold (M, g) is the type $(0, 2)$ tensor field $\mathrm{Ric} \in \Gamma(T^0_2M)$ defined by

$$\mathrm{Ric}(Y, Z) := \mathrm{tr}\,(T_pM \to T_pM : X \mapsto R(X, Y)Z)\,, \qquad Y, Z \in T_pM,\ p \in M,$$

where for any finite-dimensional vector space V, one defines the **trace** (*Spur*) of a linear map $A : V \to V$ as the trace of its matrix representative with respect to any basis.

EXERCISE 36.2. Show that the trace of a linear map $A : V \to V$ as described in Definition 36.1 does not depend on the choice of basis for V.
Hint: Use the relation $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$.

The trace in Definition 36.1 is a special case of a general operation on tensors known as **contraction**. For a tensor field $S \in \Gamma(T^k_\ell M)$ with $k, \ell \geqslant 1$, there are in general $k\ell$ ways of contracting S to produce a tensor field of type $(k-1, \ell-1)$. In the case of the Riemann tensor, there are two other contractions we could have considered besides the one in Definition 36.1; let's temporarily call them $\mathrm{Ric}_1, \mathrm{Ric}_2 \in \Gamma(T^0_2 M)$ and define them by

$$\mathrm{Ric}_1(X, Z) := \mathrm{tr}\,(Y \mapsto R(X,Y)Z)\,, \qquad \mathrm{Ric}_2(X,Y) := \mathrm{tr}\,(Z \mapsto R(X,Y)Z)\,.$$

A moment's thought about the algebraic properties of the Riemann tensor reveals why we did not define Ric_2 before: the compatibility of ∇ with the metric implies that $Z \mapsto R(X,Y)Z$ is an antisymmetric map $T_pM \to T_pM$ for any $X, Y \in T_pM$. It follows that in any orthonormal basis, the matrix representing this map has zeroes along the diagonal, and Ric_2 is thus trivial. The variant Ric_1 is not as uninteresting, but actually it is just $-\mathrm{Ric}$, since

$$\mathrm{Ric}_1(Y, Z) = \mathrm{tr}\,(X \mapsto R(Y,X)Z) = \mathrm{tr}\,(X \mapsto -R(X,Y)Z) = -\mathrm{Ric}(Y,Z).$$

We conclude that up to a sign, Ric is the *only* potentially interesting contraction of the Riemann tensor.

Exercise 36.3. Show that for any orthonormal basis $e_1, \dots, e_n \in T_pM$ at a point $p \in M$ and any $Y, Z \in T_pM$,

$$\mathrm{Ric}(Y,Z) = \sum_{j=1}^n \langle e_j, e_j \rangle \cdot \mathrm{Riem}(e_j, e_j, Y, Z).$$

Remark 36.4. In books on Riemannian geometry, one sometimes sees

$$\mathrm{Ric}(Y,Z) = \sum_{j=1}^n \mathrm{Riem}(e_j, e_j, Y, Z)$$

given as a definition of the Ricci tensor. This is just Exercise 36.3 in the case where $\langle e_j, e_j \rangle = 1$ for all $j = 1, \dots, n$, which is true if and only if the metric is positive. Since we are allowing indefinite pseudo-Riemannian metrics in our discussion, the term "orthonormal" means for us that each $\langle e_j, e_j \rangle$ is ± 1, with k positive terms and ℓ negative terms if the metric has signature (k, ℓ).

Here is another important consequence of the symmetries of the Riemann tensor:

Proposition 36.5. *The Ricci tensor is symmetric:* $\mathrm{Ric}(X,Y) = \mathrm{Ric}(Y,X)$ *for all* X, Y.

Proof. We use relations (i), (ii) and (iv) from Theorem 35.7 and compute the Ricci tensor via Exercise 36.3 with a choice of orthonormal basis $e_1, \dots, e_n \in T_pM$:

$$\begin{aligned}\mathrm{Ric}(Y,Z) &= \sum_{j=1}^n \langle e_j, e_j \rangle \cdot \mathrm{Riem}(e_j, e_j, Y, Z) \overset{\text{(i)+(ii)}}{=} \sum_{j=1}^n \langle e_j, e_j \rangle \cdot \mathrm{Riem}(Z, Y, e_j, e_j) \\ &\overset{\text{(iv)}}{=} \sum_{j=1}^n \langle e_j, e_j \rangle \cdot \mathrm{Riem}(e_j, e_j, Z, Y) = \mathrm{Ric}(Z, Y).\end{aligned}$$

□

Remark 36.6. Like the Riemann tensor, the Ricci curvature can be defined for any affine connection ∇ on a manifold M, but if ∇ is something other than the Levi-Cività connection for a metric, then Ric need not be symmetric since the Riemann tensor might not satisfy the relations (ii)–(iv) in Theorem 35.7.

Since Ric is symmetric, we can view it as a quadratic form on each tangent space, and its general values $\mathrm{Ric}(X,Y)$ are determined by the values $\mathrm{Ric}(X,X)$ for arbitrary $X \in T_pM$; indeed, $\mathrm{Ric}(X,Y)$ can be deduced from $\mathrm{Ric}(X+Y,X+Y)$, $\mathrm{Ric}(X,X)$ and $\mathrm{Ric}(Y,Y)$ via bilinearity. If the metric is positive, then Exercise 36.3 provides an interpretation of $\mathrm{Ric}(X,X)$ as a kind of "average" value (more accurately a sum) of sectional curvatures along planes tangent to X. More precisely, if we normalize X so that $|X| = 1$ and choose the orthonormal basis $e_1, \ldots, e_n$ so that $e_1 = X$, then since $\mathrm{Riem}(e_1,e_1,e_1,e_1) = 0$, Exercise 36.3 and Prop. 35.12 give
(36.1)

$$\mathrm{Ric}(X,X) = \sum_{j=2}^{n} \mathrm{Riem}(e_j,e_j,e_1,e_1) = \sum_{j=2}^{n} K_S(P_j), \qquad \text{where} \qquad P_j := \mathbb{R}e_1 \oplus \mathbb{R}e_j \subset T_pM.$$

It is therefore possible when $\dim M \geqslant 3$ for the Ricci curvature to vanish even if the sectional curvatures do not. Metrics with $\mathrm{Ric} \equiv 0$ are called **Ricci flat**, and they enjoy a special status both in Riemannian and in pseudo-Riemannian geometry; in a Lorentzian manifold, in particular, the Einstein equations give Ricci flat metrics an interpretation as possible configurations for the gravitational field on spacetime in the absence of matter. Another geometric interpretation of Ricci curvature in terms of volume will appear in §36.3 below.

It is possible to reduce the Riemann tensor even further via another contraction. This makes more explicit use of the metric: we can transform the fully covariant Ricci tensor $\mathrm{Ric} \in \Gamma(T_2^0 M)$ into a mixed tensor $\mathrm{Ric}^\# \in \Gamma(T_1^1 M)$, defined as the unique linear bundle map $\mathrm{Ric}^\# : TM \to TM$ satisfying

$$\mathrm{Ric}(X,Y) = \langle X, \mathrm{Ric}^\#(Y) \rangle \qquad \text{for all } X, Y \in T_pM,\ p \in M.$$

This works for the same reason that the musical isomorphism $\# : T^*M \to TM$ is well defined: the nondegeneracy of the bundle metric implies that there is a bundle isomorphism $\flat : T_1^1 M \to T_2^0 M$ defined by $A_\flat(X,Y) := \langle X, AY \rangle$, and the inverse $\# : T_2^0 M \to T_1^1 M$ of this isomorphism sends Ric to $\mathrm{Ric}^\sharp$. We can now contract $\mathrm{Ric}^\#$ to a tensor field of type $(0,0)$, i.e. a real-valued function, defning the **scalar curvature** (*Skalarkrümmung*) $\mathrm{Scal} : M \to \mathbb{R}$ by

$$\mathrm{Scal}(p) := \mathrm{tr}_g(\mathrm{Ric}_p) := \mathrm{tr}\left(T_pM \to T_pM : X \mapsto \mathrm{Ric}^\#(X)\right).$$

EXERCISE 36.7. Show that for any orthonormal basis $e_1, \ldots, e_n \in T_pM$ at a point $p \in M$,

$$\mathrm{Scal}(p) = \sum_{i=1}^{n} \sum_{j=1}^{n} \langle e_i, e_i \rangle \cdot \langle e_j, e_j \rangle \cdot \mathrm{Riem}(e_i, e_i, e_j, e_j).$$

Together with Prop. 35.12, Exercise 36.7 interprets the scalar curvature of a Riemannian manifold as a sum of the sectional curvatures along any complete orthogonal set of 2-planes in T_pM. Its vanishing at p thus means that the sectional curvatures along all possible 2-planes at p balance each other out.

If M is 2-dimensional, the sum in Exercise 36.7 contains only two nontrivial terms, both of which are the Gaussian curvature, thus

$$\mathrm{Scal} = 2K_G \qquad \text{when} \qquad \dim = 2.$$

We see in this way that the scalar curvature completely determines the Riemann tensor when M is 2-dimensional.

EXERCISE 36.8. Show that on any pseudo-Riemannian 2-manifold (M, g), $\mathrm{Ric} = K_G \cdot g$.

In local coordinates, it is conventional to write the components of the Ricci tensor with the same letter as for the Riemann tensor, but only two indices:

$$R_{ij} := \mathrm{Ric}(\partial_i, \partial_j) = \mathrm{tr}\left(R(\cdot, \partial_i)\partial_j\right) = R^k{}_{kij},$$

and the symmetry of the Ricci tensor then means

$$R_{ij} = R_{ji}.$$

To compute the scalar curvature, one can write the components of $\mathrm{Ric}^\sharp$ as $R^i{}_j := \mathrm{Ric}^\sharp(dx^i, \partial_j) = dx^i(\mathrm{Ric}^\sharp \partial_j)$, so the relation $\mathrm{Ric}(X,Y) = R_{ij} X^i Y^j = \langle X, \mathrm{Ric}^\#(Y)\rangle = g_{ij} X^i R^j{}_k Y^k = g_{ik} R^k{}_j X^i Y^j$ implies $R_{ij} = g_{ik} R^k{}_j$. Using the fact that the matrix with entries $g^{ij} := \langle dx^i, dx^j \rangle := \langle (dx^i)^\sharp, (dx^j)^\sharp \rangle$ is inverse to the matrix with entries g_{ij}, we deduce

$$R^i{}_j = g^{ik} R_{kj},$$

and thus

$$\mathrm{Scal} = R^i{}_i = g^{ik} R_{ki} = g^{ik} R^j{}_{jki}.$$

36.2. Jacobi vector fields. The geodesic equation is nonlinear, as one can see clearly from the quadratic term in its local coordinate expression, $\ddot{\gamma}^i + \Gamma^i_{jk} \dot{\gamma}^j \dot{\gamma}^k = 0$.[83] But nonlinear equations can be linearized: in general this means that one imagines a smooth 1-parameter family of solutions $\{\gamma_s\}_{s \in (-\epsilon,\epsilon)}$ and derives a linear differential equation that must be satisfied by $\eta := \partial_s \gamma_s|_{s=0} \in \Gamma(\gamma^* TM)$. If the maps $\gamma_s : [a,b] \to M$ are all geodesics in particular, then $\nabla_t \dot{\gamma}_s = 0$ for all s, so using (35.1) and (33.6), we have

$$\begin{aligned} 0 = \nabla_s \nabla_t \partial_t \gamma_s|_{s=0} &= \nabla_t \nabla_s \partial_t \gamma_s|_{s=0} + R(\partial_s \gamma_s, \partial_t \gamma_s) \partial_t \gamma_s|_{s=0} \\ &= \nabla_t \nabla_t \partial_s \gamma_s|_{s=0} + \nabla_t \left(T(\partial_s \gamma_s, \partial_t \gamma_s) \right)|_{s=0} + R(\eta, \dot{\gamma})\dot{\gamma} \\ &= \nabla_t^2 \eta + \nabla_t \left(T(\eta, \dot{\gamma}) \right) + R(\eta, \dot{\gamma})\dot{\gamma}. \end{aligned}$$

The linear differential equation

$$\nabla_t^2 \eta + \nabla_t \left(T(\eta, \dot{\gamma}) \right) + R(\eta, \dot{\gamma})\dot{\gamma} = 0 \tag{36.2}$$

for vector fields $\eta \in \Gamma(\gamma^* TM)$ along a geodesic γ in a manifold with an affine connection ∇ is called the **Jacobi equation**, and its solutions are called **Jacobi vector fields**. (Note that they are not actually "vector fields" on M, but are instead vector fields along γ, i.e. sections of the pullback bundle $\gamma^* TM$.) In the situation we consider most frequently, where ∇ is the Levi-Cività connection on a Riemannian manifold, the torsion term vanishes and (36.2) thus simplifies to

$$\nabla_t^2 \eta + R(\eta, \dot{\gamma})\dot{\gamma} = 0. \tag{36.3}$$

EXERCISE 36.9. If γ takes values in the domain of a chart $x = (x^1, \dots, x^n)$ and we write $x \circ \gamma = (\gamma^1, \dots, \gamma^n)$ and $\eta = \eta^i \partial_i \in \Gamma(\gamma^* TM)$, show that the Jacobi equation for η takes the form $\ddot{\eta}^i(t) + F^i{}_j(t) \dot{\eta}^j(t) + G^i{}_j(t) \eta^j(t) = 0$ for suitable smooth functions $F^i{}_j$ and $G^i{}_j$. (Don't worry too much about what these functions are—it's a bit of a mess, and we will never actually need to know.)

As solutions to a second-order linear ODE, Jacobi vector fields along a geodesic $\gamma : (a,b) \to M$ with $t_0 \in (a,b)$ are uniquely determined by an initial value $\eta(t_0) \in T_{\gamma(t_0)} M$ and an initial velocity, which in this situation can be taken to mean the first covariant derivative $\nabla_t \eta(t_0) \in T_{\gamma(t_0)} M$. With this understood, Jacobi vector fields provide a way of writing down the derivative of the map

$$\exp_p : \mathcal{O}_p \to M$$

for $p \in M$, where $\mathcal{O} \subset TM$ denotes the domain of exp and $\mathcal{O}_p := \mathcal{O} \cap T_p M$. Suppose $\gamma_s : [0,1] \to M$ is a smooth 1-parameter family of geodesics starting at $\gamma_s(0) = p$ for every s, write $\gamma := \gamma_0$,

[83] Strictly speaking, *every* differential equation for maps taking values in a manifold M must be considered nonlinear—in the sense that its solution set will not form a vector space—unless M itself happens to be a vector space. But the quadratic term means that the geodesic equation is nonlinear even in the latter case, unless the connection is trivial so that the Christoffel symbols vanish.

$\eta := \partial_s \gamma_s|_{s=0}$, $X_s := \dot\gamma_s(0) \in T_pM$, $X := X_0$ and $Y := \partial_s X_s|_{s=0} \in T_pM$. Since the domain $\mathcal{O}_p$ is an open subset of the vector space T_pM, we have a canonical identification of its tangent spaces with T_pM, and can thus write the derivative of $\exp_p$ at a point $X \in \mathcal{O}_p$ as a linear map

$$T_X(\exp_p) : T_pM \to T_{\exp_p(X)}M.$$

By definition, $\exp_p(\dot\gamma_s(0)) = \gamma_s(1)$, and since $s \mapsto \gamma_s(0)$ is a constant path, $\nabla_t\eta(0) = \nabla_t\partial_s\gamma_s(0)|_{s=0} = \nabla_s\partial_t\gamma_s(0)|_{s=0} + T(\dot\gamma(0), \partial_s\gamma_s(0)|_{s=0}) = \partial_s X_s|_{s=0} + T(X, 0) = Y$. Differentiating the expression $\exp_p(\dot\gamma_s(0)) = \gamma_s(1)$ with respect to s thus proves:

PROPOSITION 36.10. *For $p \in M$, $X \in \mathcal{O}_p$ and $Y \in T_pM$, we have*

$$T_X(\exp_p)Y = \eta(1),$$

where $\eta \in \Gamma(\gamma^*TM)$ is the unique Jacobi vector field along $\gamma(t) := \exp_p(tX)$ satisfying the initial conditions $\eta(0) = 0$ and $\nabla_t\eta(0) = Y$. □

We'll discuss two applications of the Jacobi equation in the next two subsections.

36.3. Geometry in normal coordinates. Recall from §34.3 that for any point p in a pseudo-Riemannian manifold (M, g), one can choose a Riemann normal coordinate system that identifies p with the origin in $\mathbb{R}^n$ so that the metric looks like a standard flat metric up to first order:

$$g_{ij}(x^1, \ldots, x^n) = \pm\delta_{ij} + O(|x|^2). \tag{36.4}$$

According to Theorem 35.3, we can achieve $g_{ij} \equiv \pm\delta_{ij}$ to all orders if and only if the Riemann tensor of (M, g) vanishes. We will now show that if R does not vanish, then the failure of the equation $g_{ij} \equiv \pm\delta_{ij}$ near the point p is visible already in the *quadratic* term on the right hand side of (36.4). This requires computing the Hessian of g_{ij} at p, which is determined by the individual second derivatives of the form

$$\mathcal{L}_Y^2 g_{ij}(p) = \mathcal{L}_Y\mathcal{L}_Y\langle\partial_i, \partial_j\rangle|_p = \left\langle \nabla_Y^2\partial_i\big|_p, \partial_j \right\rangle + \left\langle \partial_i, \nabla_Y^2\partial_j\big|_p \right\rangle, \tag{36.5}$$

where $Y \in \mathfrak{X}(M)$ is any vector field that has constant components in normal coordinates near p, and we have eliminated terms involving first covariant derivatives since these vanish at p. To compute the second covariant derivatives in this expression, the key trick is to notice that in normal coordinates, every radial path of constant velocity emerging from the origin is a geodesic, and as a consequence, one can easily find some Jacobi vector fields. Suppose in particular that $X(s) \in T_pM$ is a smooth family of tangent vectors at p with $X(0) = Y$ and $\partial_s X(0) = \partial_j$; then $\eta(t) := \partial_s \exp_p(tX(s))\big|_{s=0}$ is a Jacobi vector field along $\gamma(t) := \exp_p(tY)$, and since the paths $t \mapsto \exp_p(tX(s))$ appear in normal coordinates as $t \mapsto t(X^1(s), \ldots, X^n(s))$, we have $\eta(t) = t\partial_j$. The Jacobi equation for η at the point $\gamma(t) = \exp_p(tY)$ thus takes the form

$$\begin{aligned} 0 &= \nabla_t^2\eta + R(\eta, \dot\gamma)\dot\gamma = \nabla_t\nabla_t(t\partial_j) + R(t\partial_j, Y)Y = \nabla_t(\partial_j + t\nabla_Y\partial_j) + tR(\partial_j, Y)Y \\ &= \nabla_Y\partial_j + \left(\nabla_Y\partial_j + t\nabla_Y^2\partial_j\right) + tR(\partial_j, Y)Y = 2\nabla_Y\partial_j + t\left(\nabla_Y^2\partial_j + R(\partial_j, Y)Y\right). \end{aligned}$$

If we now take the inner product of this relation with ∂_i, symmetrize with respect to i and j, and divide by t, we find

$$\begin{aligned} 0 &= 2\frac{\langle\partial_i, \nabla_Y\partial_j\rangle + \langle\nabla_Y\partial_i, \partial_j\rangle}{t} + \langle\partial_i, \nabla_Y^2\partial_j\rangle + \langle\nabla_Y^2\partial_i, \partial_j\rangle + \mathrm{Riem}(\partial_i, \partial_j, Y, Y) + \mathrm{Riem}(\partial_j, \partial_i, Y, Y) \\ &= 2\frac{\mathcal{L}_Y g_{ij}}{t} + \langle\partial_i, \nabla_Y^2\partial_j\rangle + \langle\nabla_Y^2\partial_i, \partial_j\rangle + \mathrm{Riem}(\partial_i, \partial_j, Y, Y) + \mathrm{Riem}(\partial_j, \partial_i, Y, Y) \\ &= 2\frac{\mathcal{L}_Y g_{ij}}{t} + \langle\partial_i, \nabla_Y^2\partial_j\rangle + \langle\nabla_Y^2\partial_i, \partial_j\rangle + 2\mathrm{Riem}(\partial_i, \partial_j, Y, Y), \end{aligned}$$

where in the last line we've used the symmetries of the Riemann tensor. This relation is valid specifically at the point $\exp_p(tY)$, so taking the limit as $t \to 0$ and plugging in (36.5), we deduce

$$0 = 2\mathcal{L}_Y^2 g_{ij} + \mathcal{L}_Y^2 g_{ij} + 2\mathrm{Riem}(\partial_i, \partial_j, Y, Y) = 3\mathcal{L}_Y^2 g_{ij} + 2\mathrm{Riem}(\partial_i, \partial_j, Y, Y),$$

or in other words,

$$\mathcal{L}_Y^2 g_{ij} = -\frac{2}{3}\mathrm{Riem}(\partial_i, \partial_j, Y, Y) \qquad \text{at } p. \tag{36.6}$$

The bilinear form $\partial_k \partial_\ell g_{ij}\, dx^k \otimes dx^\ell$ is now uniquely determined by the condition that it is symmetric and must match this expression whenever two of the same vector (Y, Y) are fed into it, so we conclude

$$\partial_k \partial_\ell g_{ij} = -\frac{1}{3}\left(R_{ijk\ell} + R_{ij\ell k}\right) \qquad \text{at } p, \tag{36.7}$$

and the Taylor expansion of g_{ij} at p therefore becomes:

PROPOSITION 36.11. *In Riemann normal coordinates about a point $p \in M$ in a pseudo-Riemannian manifold (M, g), the components of the metric take the form*

$$g_{ij}(x^1, \dots, x^n) = \pm\delta_{ij} - \frac{1}{3} R_{ijk\ell}(p) x^k x^\ell + O(|x|^3).$$

PROOF. By (36.7), the quadratic term in the Taylor expansion is $-\frac{1}{6}\left(R_{ijk\ell}(p) + R_{ij\ell k}(p)\right) x^k x^\ell$, which simplifies after observing that the roles of the summed indices k and ℓ can be interchanged in the second (implied) summation without changing the sum. $\square$

Let us now assume for simplicity that g is positive and write down a similar local approximation formula for the Riemannian volume form

$$d\mathrm{vol} = \sqrt{\det \mathbf{g}}\, dx^1 \wedge \dots \wedge dx^n,$$

where $\mathbf{g}$ is the matrix with entries g_{ij}.

EXERCISE 36.12. Suppose $\mathbf{A}(t) \in \mathbb{F}^{m \times m}$ is a smooth 1-parameter family of m-by-m matrices over $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ such that $\mathbf{A}(0) = \mathbb{1}$ and $\partial_t \mathbf{A}(0) = \dots = \partial_t^{k-1}\mathbf{A}(0) = 0$ for some $k \in \mathbb{N}$. Show that

$$\partial_t \det(\mathbf{A}(t))|_{t=0} = \dots = \partial_t^{k-1} \det(\mathbf{A}(t))\big|_{t=0} = 0 \quad \text{and} \quad \partial_t^k \det(\mathbf{A}(t))\big|_{t=0} = \mathrm{tr}\left(\partial_t^k \mathbf{A}(0)\right).$$

Hint: Write $\det \mathbf{A}(t)$ in terms of the columns of $\mathbf{A}(t)$ and look at its Taylor expansion to order k.

Choosing again a vector field $Y \in \mathfrak{X}(M)$ with constant components in normal coordinates near p, Exercise 36.12 and (36.6) enable us to compute

$$\mathcal{L}_Y^2 \sqrt{\det \mathbf{g}}\Big|_p = \frac{1}{2\sqrt{\det(\mathbf{g})|_p}} \mathrm{tr}\left(\mathcal{L}_Y^2 (g_{ij})\big|_p\right) = -\frac{1}{3}\sum_{i=1}^n \mathrm{Riem}(\partial_i, \partial_i, Y, Y) = -\frac{1}{3}\mathrm{Ric}(Y, Y) \qquad \text{at } p.$$

In light of the symmetry of the Ricci tensor, this implies

$$\partial_k \partial_\ell \sqrt{\det \mathbf{g}}\Big|_p = -\frac{1}{3} R_{k\ell},$$

so that the Taylor expansion for the volume form near p becomes:

PROPOSITION 36.13. *In Riemann normal coordinates about a point $p \in M$ in a Riemannian manifold (M, g), the Riemannian volume form takes the form*

$$d\mathrm{vol}_{(x^1, \dots, x^n)} = \left(1 - \frac{1}{6} R_{k\ell} x^k x^\ell + O(|x|^3)\right) dx^1 \wedge \dots \wedge dx^n.$$

$\square$

This gives a nice geometric interpretation of the Ricci curvature at a point p in the Riemannian setting: it measures the degree to which volume gets distorted in small neighborhoods of p.

36.4. Nonpositive sectional curvature. At the end of last semester, we introduced sectional curvature in the context of geodesics between two fixed points, and computed the second variation of the energy functional in order to show that such geodesics always occur in isolation if the sectional curvature is nonpositive. We can now use the Jacobi equation to reprove and improve upon that result.

Definition 36.14. In a Riemannian manifold (M, g), suppose $\gamma : [a, b] \to M$ is a nonconstant geodesic from $\gamma(a) = p$ to $\gamma(b) = q$. We say that the points p and q are **conjugate along** γ if there exists a nontrivial Jacobi vector field $\eta \in \Gamma(\gamma^*TM)$ with $\eta(a) = 0$ and $\eta(b) = 0$.

The standard example to think of is on $S^2 \subset \mathbb{R}^3$, where any two antipodal points are conjugate along any great circle that connects them, the reason being that there exists a whole 1-parameter family of such great circles, so differentiating them with respect to the parameter gives a nontrivial Jacobi field that vanishes at the end points. Intuitively, this is the kind of scenario in which we expect conjugate points to arise, though in general a nontrivial Jacobi vector field might exist even if there is no corresponding 1-parameter family of geodesics.[84] The main observation we'd like to make now is that in the example of the sphere, the curvature is positive, and if it weren't, then conjugate points could not exist:

Proposition 36.15. *Suppose (M, g) is a Riemannian manifold with everywhere nonpositive sectional curvature $K_S \leqslant 0$. Then no two points in M are conjugate along any geodesic.*

Proof. Suppose $\gamma : [a, b] \to M$ is a nonconstant geodesic and $\eta \in \Gamma(\gamma^*TM)$ is a Jacobi field satisfying $\eta(a) = 0$ and $\eta(b) = 0$. The Jacobi equation and the assumption $K_S \leqslant 0$ then imply

$$\langle \eta(t), \nabla_t^2 \eta(t) \rangle = -\langle \eta(t), R(\eta(t), \dot\gamma(t))\dot\gamma(t) \rangle = -\mathrm{Riem}(\eta(t), \eta(t), \dot\gamma(t), \dot\gamma(t)) \geqslant 0$$

for all t. But since $\partial_t \langle \eta, \nabla_t \eta \rangle = |\nabla_t \eta|^2 + \langle \eta, \nabla_t^2 \eta \rangle$ and η vanishes at the end points, we can integrate by parts, giving

$$0 \leqslant \int_a^b \langle \eta(t), \nabla_t^2 \eta(t) \rangle \, dt = -\int_a^b |\nabla_t \eta(t)|^2 \, dt \leqslant 0,$$

which implies that η must be parallel along γ. Since it vanishes at the end points, it follows that η is trivial. $\square$

Recall that a smooth map $f : M \to N$ is called a **local diffeomorphism** if every point $p \in M$ has a neighborhood $\mathcal{U} \subset M$ such that $f|_{\mathcal{U}}$ is a diffeomorphism onto an open set $f(\mathcal{U}) \subset N$. By the inverse function theorem, this is equivalent to the condition that $T_p f : T_p M \to T_{f(p)} N$ is invertible for every $p \in M$.

Corollary 36.16. *If (M, g) is a Riemannian manifold with nonpositive sectional curvature, then for every $p \in M$, the map $\exp_p : \mathcal{O}_p \to M$ is a local diffeomorphism.*

Proof. For any $X \in \mathcal{O}_p$ and $Y \in T_X(\mathcal{O}_p) = T_p M$, Proposition 36.10 gives $T_X(\exp_p) Y$ as the end value $\eta(1)$ for the unique Jacobi vector field η along $\gamma(t) := \exp_p(tX)$ such that $\eta(0) = 0$ and $\nabla_t \eta(0) = Y$. By Proposition 36.15, $\eta(1)$ cannot be 0 for any $Y \neq 0$, otherwise the points p and $\exp_p(X)$ would be conjugate along γ. This proves that $T_X(\exp_p) : T_p M \to T_{\exp_p(X)} M$ is an injective map, and it is therefore also surjective. $\square$

[84] The following is certainly true: if $\eta \in \Gamma(\gamma^*TM)$ is a Jacobi vector field along $\gamma : [a, b] \to M$ such that $\eta(a)$ and $\eta(b)$ both vanish, then $\eta = \partial_s \gamma_s|_{s=0}$ for some 1-parameter family of geodesics $\gamma_s : [a, b] \to M$ with $\gamma_0 = \gamma$, $\gamma_s(a) = \gamma(a)$ for all s and $\partial_s \gamma_s(b)|_{s=0} = 0$. However, the latter does not guarantee that $\gamma_s(b) = \gamma(b)$ for all s, and it may or may not be possible to find a family with this property.

To appreciate the implications of this result, it helps to be familiar with covering space theory, as treated in a standard first course on algebraic topology (see e.g. [**Wen18**, Lectures 14–17]). Such familiarity will be assumed for the remainder of this lecture. Here is the main result:

THEOREM 36.17 (Cartan-Hadamard). *Suppose (M, g) is a complete and connected Riemannian n-manifold with nonpositive sectional curvature. Then for any point $p \in M$, $\exp_p : T_pM \to M$ is a covering map. In particular, $\exp_p$ is the universal cover of M, so if M is simply connected, it follows that M is diffeomorphic to $\mathbb{R}^n$.*

Before getting into the proof, here is a typical application.

COROLLARY 36.18. *If (M, g) is a complete and connected Riemannian manifold with nonpositive sectional curvature, then for any two points $p, q \in M$, there exists a unique geodesic segment in every homotopy class of paths from p to q.*

PROOF. By standard results on covering spaces, every continuous path γ from p to q in M has a unique lift $\tilde{\gamma}$ to the covering space T_pM that begins at the point $\tilde{p} := 0 \in T_pM$, ending at some point $X \in T_pM$ with $\exp_p(X) = q$, and since T_pM is simply connected, this lifting construction gives a bijection between the set of homotopy classes (with fixed end points) of paths from p to q and the set of all possible lifted end points $\left\{X \in T_pM \;\middle|\; \exp_p(X) = q\right\}$. Each homotopy class thus corresponds to a unique X in this set, and $t \mapsto \exp_p(X)$ is then the unique geodesic from p to q in that homotopy class. □

Recall that for two Riemannian manifolds (M, g) and (N, h), a smooth map $\varphi : M \to N$ is called an **isometry** if it is a diffeomorphism and $\varphi^*h = g$. If $\varphi : M \to N$ is a local diffeomorphism, then for any Riemannian metric h on N, the pullback $g := \varphi^*h$ is a Riemannian metric on M such that every point $p \in M$ has a neighborhood $(\mathcal{U}, g)$ that φ maps isometrically to a neighborhood of $\varphi(p)$ in (N, h). We say in this case that φ is a **local isometry** from (M, g) to (N, h).[85]

PROOF OF THEOREM 36.17. Using the popular notation in topology for the universal cover, let us fix $p \in M$ and write

$$\widetilde{M} := T_pM, \qquad \pi := \exp_p : \widetilde{M} \to M.$$

Since π is (according to Corollary 36.16) a local diffeomorphism, the pullback

$$\tilde{g} := \pi^*g \in \Gamma(T_2^0\widetilde{M})$$

is a Riemannian metric on $\widetilde{M}$, and π is then a local isometry from $(\widetilde{M}, \tilde{g})$ to (M, g). It follows that whenever $\tilde{\gamma}$ is a path in $\widetilde{M}$ and $\gamma := \pi \circ \tilde{\gamma}$, $\tilde{\gamma}$ is a geodesic in $(\widetilde{M}, \tilde{g})$ if and only if γ is a geodesic in (M, g). In particular, the paths $t \mapsto tX$ for every $X \in T_pM$ are therefore geodesics in $(\widetilde{M}, \tilde{g})$, and since these are defined for all $t \in \mathbb{R}$, it follows via the Hopf-Rinow theorem that $(\widetilde{M}, \tilde{g})$ is also complete. The theorem now follows from Lemma 36.19 below. □

We used:

LEMMA 36.19. *For any local isometry $f : (M, g) \to (N, h)$ between Riemannian manifolds, if (M, g) is complete, then f is a covering map.*

PROOF. We need to show that every $p \in N$ has a neighborhood $\mathcal{U}_p \subset N$ that is evenly covered, meaning that $f^{-1}(\mathcal{U}_p)$ is a union of disjoint open subsets $\mathcal{V}_\alpha \subset M$ that are each mapped by f homeomorphically (and in this case diffeomorphically) onto $\mathcal{U}_p$. Choose $\mathcal{U}_p$ to be the geodesic ball $\exp_p(B_\epsilon^p) \subset N$ for some $\epsilon < \operatorname{inj}(p)$, where $B_\epsilon^p \subset T_pN$ denotes the open ball of radius ϵ around the

[85] For reasons of time, the proof of Theorem 36.17 was omitted from the actual lecture, excepting some brief remarks about the important role played by Corollary 36.16.

origin. Since f is a local diffeomorphism, the set $f^{-1}(p) \subset M$ is discrete, and for each $q \in f^{-1}(p)$, we can define $\mathcal{V}_q := \exp_q(B^q_\epsilon) \subset M$, noting that $\exp_q$ is guaranteed to be defined on B^q_ϵ for every $q \in M$ since (M,g) is complete. In fact, we claim that $\mathcal{V}_q$ is also a geodesic ball, and is mapped by f diffeomorphically onto $\mathcal{U}_p$. This follows mainly from the local isometry condition, which implies that $T_q f$ maps B^q_ϵ bijectively onto B^p_ϵ, and also that $\exp_p \circ T_q f = f \circ \exp_q$, hence f maps geodesics in (M,g) to geodesics in (N,h), and the map $\mathcal{U}_p \to \mathcal{V}_q$ defined as the composition

$$\begin{tikzcd} & T_pN \arrow[r, "(T_qf)^{-1}"] & T_qM & \\ & \cup & \cup & \\ \mathcal{U}_p \arrow[r, "(\exp_p)^{-1}"] \arrow[rrr, bend right] & B^p_\epsilon \arrow[r, "(T_qf)^{-1}"] & B^q_\epsilon \arrow[r, "\exp_q"] & \mathcal{V}_q \end{tikzcd}$$

is a smooth inverse for $\mathcal{V}_q \xrightarrow{f} \mathcal{U}_p$. We claim next that $\mathcal{V}_q \cap \mathcal{V}_{q'} = \emptyset$ for any two distinct points $q, q' \in f^{-1}(p)$. To see this, observe that any point $x \in \mathcal{V}_q \cap \mathcal{V}_{q'}$ would be the end point of both a geodesic segment $\tilde{\alpha}$ through $\mathcal{V}_q$ starting at q and a geodesic segment $\tilde{\beta}$ through $\mathcal{V}_{q'}$ starting at q'. Then $\alpha := f \circ \tilde{\alpha}$ and $\beta := f \circ \tilde{\beta}$ are both geodesic segments through $\mathcal{U}_p$ from p to $f(x)$, and must therefore be identical (up to parametrization). It follows that $\tilde{\alpha}$ and $\tilde{\beta}$ are tangent to each other when they intersect at x, and must therefore also be identical up to parametrization; this is only possible if $q = q'$.

Finally, to see that $f^{-1}(\mathcal{U}_p) = \bigcup_{q \in f^{-1}(p)} \mathcal{V}_q$, note that whenever $x \in f^{-1}(\mathcal{U}_p)$, there is a unique geodesic segment through $\mathcal{U}_p$ from $f(x)$ to p, which we can denote by $[0,1] \to \mathcal{U}_p : t \mapsto \gamma(t) := \exp_{f(x)} X$ for some $X \in T_{f(x)}N$, and note that $\dot{\gamma}(1) \in B^p_\epsilon \subset T_pN$. Writing $\widetilde{X} := (T_x f)^{-1}(X) \subset T_xM$, the geodesic segment $[0,1] \to M : t \mapsto \tilde{\gamma}(t) := \exp_x(t\widetilde{X})$ then satisfies $f \circ \tilde{\gamma} = \gamma$, thus it ends at a point $q \in f^{-1}(p)$ with velocity $\dot{\gamma}(1) \in B^q_\epsilon \subset T_qM$, from which it follows that γ is contained in $\mathcal{V}_q$. $\square$

Exercise 36.20. Show that Lemma 36.19 becomes false in general if the assumption that (M,g) is complete is dropped.
Hint: Take a well-behaved covering space and remove one point.

Exercise 36.21. On a pseudo-Riemannian manifold (M,g) with a geodesic segment $\gamma : [a,b] \to M$ from $\gamma(a) = p$ to $\gamma(b) = q$, prove that the following conditions are equivalent:

(i) p and q are not conjugate along γ;
(ii) For all $X \in T_pM$ and $Y \in T_qM$, there exists a unique Jacobi vector field $\eta \in \Gamma(\gamma^*TM)$ satisfying $\eta(a) = X$ and $\eta(b) = Y$.

37. Lie groups and their Lie algebras

37.1. Main definitions and examples. We have encountered several examples so far of groups that are also smooth manifolds in a natural way. The most popular are the matrix groups $\mathrm{GL}(n,\mathbb{F})$ for $\mathbb{F} \in \{\mathbb{R},\mathbb{C}\}$ and their well-known subgroups $\mathrm{O}(n), \mathrm{SO}(n), \mathrm{SL}(n,\mathbb{R}) \subset \mathrm{GL}(n,\mathbb{R})$, $\mathrm{U}(n), \mathrm{SU}(n), \mathrm{SL}(n,\mathbb{C}) \subset \mathrm{GL}(n,\mathbb{C})$ and so forth. These arise whenever one considers a manifold or vector bundle with some extra geometric structure such as a bundle metric or volume form; the so-called *structure group* of the bundle is then a matrix group consisting of all linear transformations that preserve the relevant structure on a single fiber of the bundle. Moving forward, we will increasingly also have to consider examples like the isometry group $\mathrm{Isom}(M,g)$ of a Riemannian manifold (M,g), which is not a group of matrices in any natural way, but we will see that it is nonetheless a smooth manifold. For all these reasons and more, we are somewhat overdue for a more systematic examination of the intersection between differential geometry and group theory.

DEFINITION 37.1. A **topological group** (*topologische Gruppe*) G is a group with a topology such that the maps

$$G \times G \to G : (a, b) \mapsto ab \qquad \text{and} \qquad G \to G : a \mapsto a^{-1}$$

are both continuous.

Similarly, a **Lie group** (*Liesche Gruppe* or *Lie-Gruppe*) is a group that is also a smooth manifold (without boundary) such that the two maps above are both smooth.

For a Lie group G, a **Lie subgroup** is a subgroup $H \subset G$ that is also a smooth submanifold.

If G and H are both Lie groups, a **Lie group homomorphism** $G \to H$ is a group homomorphism that is also a smooth map. We call it a **Lie group isomorphism** if it is a group isomorphism and a diffeomorphism.

REMARK 37.2. Since smooth maps restrict smoothly to smooth submanifolds, a Lie subgroup $H \subset G$ of a Lie group G is (according to the definition above) also a Lie group in a natural way, and the inclusion map $H \hookrightarrow G$ is then a Lie group homomorphism, as well as an embedding.

REMARK 37.3. The reader should be warned that our definition of the term "Lie subgroup" is stricter than what is found in many other sources: these allow a Lie subgroup to be a subgroup with its own Lie group structure such that the inclusion is an injective immersion but not necessarily an embedding (in which case the subgroup would not be a submanifold by our definition). Since the most interesting examples satisfy the stricter definition, I will stick with it until someone convinces me to change it.

EXAMPLE 37.4. For $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ and $n \in \mathbb{N}$, the general linear group $\mathrm{GL}(n, \mathbb{F})$ is an open set in the vector space $\mathbb{F}^{n \times n}$ of all n-by-n matrices over $\mathbb{F}$, and basic theorems in linear algebra imply that the maps $\mathrm{GL}(n, \mathbb{F}) \times \mathrm{GL}(n, \mathbb{F}) \to \mathrm{GL}(n, \mathbb{F}) : (\mathbf{A}, \mathbf{B}) \mapsto \mathbf{AB}$ and $\mathrm{GL}(n, \mathbb{F}) \to \mathrm{GL}(n, \mathbb{F}) : \mathbf{A} \mapsto \mathbf{A}^{-1}$ are both smooth, hence $\mathrm{GL}(n, \mathbb{F})$ is a Lie group with the smooth structure it inherits as an open subset of a finite-dimensional vector space. One can then use the implicit function theorem to show that the classical matrix groups such as $\mathrm{GL}_+(n, \mathbb{R})$, $\mathrm{O}(n)$, $\mathrm{O}(k, \ell)$, $\mathrm{SO}(n)$, $\mathrm{U}(n)$, $\mathrm{SU}(n)$, $\mathrm{SL}(n, \mathbb{R})$ and $\mathrm{SL}(n, \mathbb{C})$ are each Lie subgroups of $\mathrm{GL}(n, \mathbb{R})$ or $\mathrm{GL}(n, \mathbb{C})$. These were all worked out as exercises last semester—see especially §4.6 and §18.1.

EXAMPLE 37.5. Every finite-dimensional vector space V over $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ is naturally an abelian Lie group, with vector addition as the group operation. After identifying V with $\mathbb{F}^n$ via a choice of basis, the map

$$\mathbb{F}^n \to \mathrm{GL}(n+1, \mathbb{F}) : \mathbf{v} \mapsto \begin{pmatrix} \mathbb{1}_{n \times n} & \mathbf{v} \\ 0_{1 \times n} & 1 \end{pmatrix}$$

defines an injective Lie homomorphism, identifying V with a Lie subgroup of $\mathrm{GL}(n+1, \mathbb{F})$.

EXAMPLE 37.6. Every group with at most countably many elements can be regarded as a 0-dimensional Lie group by assigning it the discrete topology; we refer to Lie groups of this form as **discrete groups**. Whether this is actually a reasonable thing to do depends on the situation: it is universally appropriate for finite groups, though for instance the countable subgroup $\mathbb{Q}\backslash\{0\} = \mathrm{GL}(1, \mathbb{Q}) \subset \mathrm{GL}(1, \mathbb{R})$ inherits a natural topology from $\mathbb{R}^{1 \times 1} = \mathbb{R}$ that is quite different from the discrete topology; $\mathbb{Q}\backslash\{0\}$ with its natural topology is not a manifold of any dimension, and thus not a Lie group.

EXAMPLE 37.7. The quotient of any finite-dimensional vector space V by any discrete subgroup $\Gamma \subset V$ inherits from V a natural smooth structure that makes it an abelian Lie group. This fact will follow as a special case of a general theorem about quotients of Lie groups by subgroups, but it is easy to see it directly in the most popular examples, e.g. the n-torus $\mathbb{T}^n = S^1 \times \ldots \times S^1$ becomes a Lie group when identified with the quotient of $\mathbb{R}^n$ by the lattice $\mathbb{Z}^n$. In particular, the circle S^1 can be identified either with the quotient group $\mathbb{R}/\mathbb{Z}$ or with the unit circle in $\mathbb{C}^* := \mathbb{C}\backslash\{0\}$ with its multiplicative group structure; there is a Lie group isomorphism between these two models. Viewing S^1 as the unit circle in $\mathbb{C}$ also identifies it with the matrix group $\mathrm{U}(1) \subset \mathrm{GL}(1, \mathbb{C}) = \mathbb{C}^*$.

We will not have very much to say in this course about topological groups that are not Lie groups, but one or two popular examples are important to be aware of. If X is any topological space, it is natural to consider the group of all homeomorphisms $X \to X$,

$$\mathrm{Homeo}(X) := \left\{\varphi : X \to X \;\middle|\; \varphi \text{ is a homeomorphism}\right\},$$

where the group multiplication law is defined via composition of maps. A natural topology to define on $\mathrm{Homeo}(X)$ is the so-called *compact-open* topology, which is generated by all sets of the form

$$\mathcal{U}_{K,V} := \left\{\varphi \in \mathrm{Homeo}(X) \;\middle|\; \varphi(K) \subset V\right\}$$

where $K \subset X$ is an arbitrary compact set and $V \subset X$ an arbitrary open set. For our purposes, it is easiest to describe what this means under the additional assumption that X is a metric space: then a sequence $\varphi_k \in \mathrm{Homeo}(X)$ converges to $\varphi \in \mathrm{Homeo}(X)$ in the compact-open topology if and only if it converges uniformly on all compact subsets of X. For this reason, we also sometimes call it the C^0_{loc}**-topology**. Under additional mild technical assumptions on X such as the second countability axiom, one can define a metric on the space of all continuous maps $X \to X$ for which the induced notion of convergence is C^0_{loc}-convergence, so in this setting, the compact-open topology is metrizable. Unfortunately, it is not always true that $\mathrm{Homeo}(X)$ with the compact-open topology is a topological group, but it is true under fairly mild assumptions about the topology of X. Concretely, the map $\mathrm{Homeo}(X) \times \mathrm{Homeo}(X) \to \mathrm{Homeo}(X) : (\varphi, \psi) \mapsto \varphi \circ \psi$ is continuous whenever X is locally compact and Hausdorff, and if X is additionally either compact or locally connected, then $\mathrm{Homeo}(X) \to \mathrm{Homeo}(X) : \varphi \mapsto \varphi^{-1}$ is also continuous, making $\mathrm{Homeo}(X)$ a topological group. In particular, this is always true if X is a finite-dimensional topological manifold. (For a more detailed discussion of these facts, see [**Wen18**, Exercise 7.27].)

For a smooth manifold M, one similarly considers the group

$$\mathrm{Diff}(M) := \left\{\varphi : M \to M \;\middle|\; \varphi \text{ is a diffeomorphism}\right\},$$

which is a subgroup of $\mathrm{Homeo}(M)$, but it is natural to take advantage of the additional structure provided by differentiability and assign to $\mathrm{Diff}(M)$ a stronger topology:

Definition 37.8. Suppose M and N are smooth manifolds. A sequence of smooth maps $f_k : M \to N$ is called C^∞_{loc}**-convergent** to a smooth map $f : M \to N$ if for every pair of smooth charts $M \supset \mathcal{U} \xrightarrow{x} \mathbb{R}^m$ and $N \supset \mathcal{V} \xrightarrow{y} \mathbb{R}^n$, the partial derivatives of all nonnegative orders of the functions $y \circ f_k \circ x^{-1}$ converge uniformly on all compact subsets of their domains to the corresponding partial derivatives of $y \circ f \circ x^{-1}$.

One can use coverings of M and N by countably-many charts to construct a metric on the space of all smooth maps $C^\infty(M, N)$ for which the notion of convergence is C^∞_{loc}-convergence as described in Definition 37.8. Such a metric will inevitably depend on a multitude of arbitrary choices, but the induced topology on $C^\infty(M, N)$ does not; we call it the C^∞_{loc}**-topology**. As a subset of $C^\infty(M, M)$, the group $\mathrm{Diff}(M)$ inherits this topology and becomes a topological group. It cannot be a Lie group, however, at least according to our current definitions, as it is too large to be a finite-dimensional manifold.[86]

[86] There are various ways of defining the notion of an infinite-dimensional Lie group so that $\mathrm{Diff}(M)$ becomes an example, but this involves a multitude of thorny technical issues that call the usefulness of such a notion into

On the other hand, interesting examples of Lie groups do sometimes arise as natural subgroups of $\mathrm{Diff}(M)$ with the C^∞_{loc}-topology. One that we will examine later more closely is the isometry group of a pseudo-Riemannian manifold (M, g),

$$\mathrm{Isom}(M, g) := \left\{ \varphi \in \mathrm{Diff}(M) \;\middle|\; \varphi^* g = g \right\}.$$

This is our first example of a Lie group that is not generally a matrix group in any natural way, and actually proving that it's a Lie group will require a substantial effort. While most concrete examples we consider can at least be shown to admit injective Lie group homomorphisms into $\mathrm{GL}(n, \mathbb{R})$ for some $n \in \mathbb{N}$ sufficiently large, it can be proved that not all Lie groups have this property. (Contrast this with the standard theorem in differential topology that every smooth n-manifold admits a smooth embedding into $\mathbb{R}^N$ for N sufficiently large.)

EXAMPLE 37.9. A map $\varphi : \mathbb{F}^n \to \mathbb{F}^n$ is called *affine* if it has the form

$$\varphi(\mathbf{v}) = \mathbf{A}\mathbf{v} + \mathbf{b}$$

for some $\mathbf{A} \in \mathrm{GL}(n, \mathbb{F})$ and $\mathbf{b} \in \mathbb{F}^n$. These form the group of **affine transformations** (*affine Transformationen*) on $\mathbb{F}^n$, which we will denote by

$$\mathrm{Aff}(\mathbb{F}^n) \subset \mathrm{Diff}(\mathbb{F}^n),$$

and the obvious bijection $\mathrm{Aff}(\mathbb{F}^n) \leftrightarrow \mathrm{GL}(n, \mathbb{F}) \times \mathbb{F}^n$ endows it with a smooth structure that makes it into a Lie group. (Exercise: Convince yourself that the resulting topology on $\mathrm{Aff}(\mathbb{F}^n)$ matches the C^∞_{loc}-topology.) The groups $\mathrm{GL}(n, \mathbb{F})$ and $\mathbb{F}^n$ both live naturally inside $\mathrm{Aff}(\mathbb{F}^n)$ as the Lie subgroups

$$\left\{ \mathbf{v} \mapsto \mathbf{A}\mathbf{v} \;\middle|\; \mathbf{A} \in \mathrm{GL}(n, \mathbb{F}) \right\} \qquad \text{and} \qquad \left\{ \mathbf{v} \mapsto \mathbf{v} + \mathbf{b} \;\middle|\; \mathbf{b} \in \mathbb{F}^n \right\}$$

respectively, and $\mathrm{Aff}(\mathbb{F}^n)$ can be identified with a group-theoretic construction called the **semidirect product** of these two groups. The same trick can be used to produce many other Lie groups by replacing $\mathrm{GL}(n, \mathbb{F})$ with one of its subgroups: for instance, the semidirect product of $\mathrm{O}(n)$ with $\mathbb{R}^n$ is the **Euclidean group**

$$\{\varphi \in \mathrm{Diff}(\mathbb{R}^n) \mid \varphi(\mathbf{v}) = \mathbf{A}\mathbf{v} + \mathbf{b} \text{ for some } \mathbf{A} \in \mathrm{O}(n) \text{ and } \mathbf{b} \in \mathbb{R}^n\},$$

which we will later see is precisely the isometry group of $\mathbb{R}^n$ with its standard Euclidean metric. An important cousin of this group arises in Einstein's theory of Special Relativity: the **Poincaré group** is the semidirect product of the Lorentz group $\mathrm{O}(1, 3)$ with $\mathbb{R}^4$, and it turns out to be identical to the isometry group of $\mathbb{R}^4$ with the Minkowski metric. In light of the injective Lie group homomorphism

$$\mathrm{Aff}(\mathbb{F}^n) \to \mathrm{GL}(n+1, \mathbb{F}) : (\mathbf{v} \mapsto \mathbf{A}\mathbf{v} + \mathbf{b}) \mapsto \begin{pmatrix} \mathbf{A} & \mathbf{b} \\ 0_{1\times n} & 1 \end{pmatrix},$$

all Lie subgroups of $\mathrm{Aff}(\mathbb{F}^n)$ can be identified with Lie subgroups of $\mathrm{GL}(n+1, \mathbb{F})$.

NOTATION. Whenever G is an arbitrary group, as opposed to one of the concrete examples such as matrix groups or subgroups of $\mathrm{Diff}(M)$, we will typically denote the identity element by

$$e \in G,$$

hence the relation

$$ge = eg = g$$

is assumed to hold for every $g \in G$. We will also often use the notation

$$i : G \to G$$

for the inversion map $i(g) := g^{-1}$.

question, and place it in any case outside the realm of standard differential geometry. In other words, we won't discuss it any further here.

EXERCISE 37.10. Prove that on any Lie group G, $T_e i : T_e G \to T_e G$ is multiplication by -1.

37.2. Left and right translation. In a Lie group G, every element $g \in G$ determines two diffeomorphisms on G known as **left translation** (*Linkstranslation*) and **right translation** (*Rechtstranslation*):

$$L_g : G \to G : h \mapsto gh, \qquad R_g : G \to G : h \mapsto hg.$$

You should take a moment to convince yourself that both are smooth and have smooth inverses; the former follows from the smoothness of the group operations on G, and the latter from the fact that every element of G has an inverse.

DEFINITION 37.11. A tensor field $S \in \Gamma(T^k_\ell G)$ on a Lie group G is called **left-invariant** (*linksinvariant*) if $L_g^* S = S$ for every $g \in G$, and **right-invariant** (*rechtsinvariant*) if $R_g^* S = S$ for every $g \in G$. If S is both left- and right-invariant, it is called **bi-invariant**.

REMARK 37.12. Since we are dealing with diffeomorphisms and every group element has an inverse, Definition 37.11 could equally well have been stated in terms of pushforwards instead of pullbacks: for instance, the pushforward operator $(L_g)_*$ is the same thing as the pullback via $(L_g)^{-1} = L_{g^{-1}}$, so a tensor field S is left-invariant if and only if $(L_g)_* S = S$ for all $g \in G$.

Recall that for a finite-dimensional vector space V and integers $k, \ell \geqslant 0$, we denote $V^k_\ell := V^{\otimes k} \otimes (V^*)^{\otimes \ell}$, so a tensor field $S \in \Gamma(T^k_\ell M)$ of type (k, ℓ) on a manifold M assigns to each $p \in M$ an element $S_p \in (T_p M)^k_\ell$.

PROPOSITION 37.13. *For every $S_e \in (T_e G)^k_\ell$, there exists a unique left-invariant tensor field $S^L \in \Gamma(T^k_\ell G)$ and a unique right-invariant tensor field $S^R \in \Gamma(T^k_\ell G)$ whose values at $e \in G$ are both S_e.*

PROOF. The uniqueness of S^L is seen as follows: since $L_g(e) = g$ for each $g \in G$, left-invariance implies that for every $\lambda^1, \dots, \lambda^k \in T_g^* G$ and $X_1, \dots, X_\ell \in T_g G$,

$$\begin{aligned} S^L_g(\lambda^1, \dots, \lambda^k, X_1, \dots, X_\ell) &= ((L_g)_* S^L)_g(\lambda^1, \dots, \lambda^k, X_1, \dots, X_\ell) \\ &= S^L_e(L_g^* \lambda^1, \dots, L_g^* \lambda^k, L_g^* X_1, \dots, L_g^* X_\ell) \\ &= S_e(L_g^* \lambda^1, \dots, L_g^* \lambda^k, TL_{g^{-1}}(X_1), \dots, TL_{g^{-1}}(X_\ell)). \end{aligned}$$

To prove the existence of S^L, it now suffices to prove that the tensor field defined on G in terms of S_e via this last expression really is left-invariant; this is a straightforward exercise. The existence and uniqueness of S^R is proved similarly, using right instead of left translation. $\square$

EXERCISE 37.14. Show that for any $X \in T_e G$, the unique left-invariant vector field $X^L \in \mathfrak{X}(G)$ and right-invariant vector field $X^R \in \mathfrak{X}(G)$ satisfying $X^L(e) = X^R(e) = X$ are given by

$$X^L(g) = TL_g(X), \qquad X^R(g) = TR_g(X).$$

EXERCISE 37.15. A manifold is called **parallelizable** (*parallelisierbar*) if its tangent bundle is a trivial bundle, or equivalently, if its tangent bundle admits a global frame. Show that every Lie group is parallelizable.

Proposition 37.13 shows that one should not generally expect nontrivial *bi-invariant* tensor fields to exist, at least not if G is non-abelian (in which case right and left translation are equivalent). In general, a bi-invariant tensor field $S \in \Gamma(T^k_\ell G)$ with a given value S_e at the identity element will exist if and only if the unique left- and right-invariant tensors S^L and S^R with that value happen to be identical, which in general they will not be. We will see later that there are a few important exceptions to this.

Using the notation $i(g) = g^{-1}$ for the inversion map $G \xrightarrow{i} G$, the next exercise provides a useful tool for understanding the relationship between left- and right-invariant tensors:

Exercise 37.16. For a tensor field $S \in \Gamma(T^k_\ell G)$ on a Lie group G, prove that S is left-invariant if and only if i^*S is right-invariant.
Hint: Rewrite the compositions $i \circ L_g$ and $L_g \circ i$ for arbitrary $g \in G$ in terms of right translation.

37.3. The Lie algebra of a Lie group. Left- and right-invariant vector fields play a special role in the theory of Lie groups. We will use the following notation for them:

$$\mathfrak{X}^L(G) := \{\text{left-invariant vector fields}\} \subset \mathfrak{X}(G), \quad \mathfrak{X}^R(G) := \{\text{right-invariant vector fields}\} \subset \mathfrak{X}(G),$$

and for each $X \in T_eG$, we let

$$X^L \in \mathfrak{X}^L(G), \qquad X^R \in \mathfrak{X}^R(G)$$

denote the unique invariant vector fields satisfying $X^L(e) = X^R(e) = X$, as guaranteed by Proposition 37.13. The evaluation map $\mathfrak{X}(G) \to T_eG : X \mapsto X(e)$ restricts to each of $\mathfrak{X}^L(G)$ and $\mathfrak{X}^R(G)$ as an isomorphism, with inverse given by $X \mapsto X^L$ or $X \mapsto X^R$ respectively, thus $\mathfrak{X}^L(G)$ and $\mathfrak{X}^R(G)$ are both finite-dimensional subspaces of $\mathfrak{X}(G)$. They are also Lie subalgebras, meaning they are closed under the Lie bracket of vector fields: e.g. using the obvious analogue of Exercise 6.5 for pullbacks instead of pushforwards, one finds

$$X, Y \in \mathfrak{X}^L(G) \quad \Rightarrow \quad L_g^*[X,Y] = [L_g^*X, L_g^*Y] = [X,Y] \ \forall g \in G \quad \Rightarrow \quad [X,Y] \in \mathfrak{X}^L(G),$$

and similarly for $X, Y \in \mathfrak{X}^R(G)$.

Exercise 37.17. Show that every left- or right-invariant vector field on G has a global flow.
Hint: If $X \in \mathfrak{X}^L(G)$ and $\varphi^t_X(e)$ is defined for all $t \in (-\epsilon, \epsilon)$, show that $\varphi^t_X : G \to G$ is globally defined for all t in this range. Use left translation to find flow lines through arbitrary points.

Recall from Lecture 5 in the first semester: a **Lie algebra** (*Lie-Algebra*) in general is a vector space V that is endowed with an antisymmetric bilinear "bracket" operation $[\cdot,\cdot] : V \times V \to V$ satisfying the Jacobi identity

$$[v,[w,u]] + [w,[u,v]] + [u,[v,w]] = 0 \qquad \text{for all } v,w,u \in V.$$

Given two Lie algebras V and W, a **Lie algebra homomorphism** $A : V \to W$ is a linear map that preserves the bracket structures, meaning

$$[Au, Av] = [u,v] \qquad \text{for all } u, v \in V,$$

and we call it a **Lie algebra isomorphism** if it is also bijective. A linear subspace $W \subset V$ of a Lie algebra V is a **Lie subalgebra** if $[u,v] \in W$ for all $u, v \in W$, in which case W itself becomes a Lie algebra for which the inclusion map $W \hookrightarrow V$ is an injective Lie algebra homomorphism.

The only interesting concrete example of a Lie algebra we have considered previously in this course is the space of vector fields $\mathfrak{X}(M)$ on a manifold M, but restricting to invariant vector fields on a Lie group now gives us finite-dimensional examples as well.

Definition 37.18. Given an n-dimensional Lie group G, the **Lie algebra associated to G** is the n-dimensional vector space

$$\mathfrak{g} := T_eG,$$

endowed with the unique bracket $[\cdot,\cdot]$ for which the map $T_eG \to \mathfrak{X}^L(G) : X \mapsto X^L$ is a Lie algebra isomorphism; that is,

$$[X,Y] := [X^L, Y^L](e) \in \mathfrak{g} \qquad \text{for } X, Y \in \mathfrak{g}.$$

Notation. The convention for denoting the Lie algebra associated to a Lie group is to use the same letter or letters but in lowercase Fraktur. So for instance, the Lie algebra of a Lie group H is called $\mathfrak{h}$, the Lie algebras of $\mathrm{GL}(n,\mathbb{F})$ and $\mathrm{SO}(n)$ are called $\mathfrak{gl}(n,\mathbb{F})$ and $\mathfrak{so}(n)$ respectively, and so forth.

The Lie algebra $\mathfrak{g}$ comes with a canonical Lie algebra isomorphism to the Lie algebra $\mathfrak{X}^L(G)$ of left-invariant vector fields, and some books therefore also give $\mathfrak{X}^L(G)$ itself as an equivalent definition of $\mathfrak{g}$. This perspective is useful to keep in mind, though for our purposes, it will usually be most convenient to regard the elements of $\mathfrak{g}$ as actual tangent vectors to G at e rather than vector fields.

You may now be wondering: why do we formulate Definition 37.18 in terms of left-invariant rather than right-invariant vector fields, and does it make a difference? The following exercise shows that, yes, it makes a slight difference, but the difference is a trivial matter of inserting a minus sign into the definition of $[\ ,\ ]$ on $\mathfrak{g}$. With this understood, the choice to define $\mathfrak{g}$ in terms of $\mathfrak{X}^L(G)$ instead of $\mathfrak{X}^R(G)$ is an arbitrary convention without any deep meaning, though fortunately, this convention is observed consistently throughout the literature.

Exercise 37.19. For a Lie group G with inversion map $i : G \to G$, prove:

(a) For each $X \in T_eG$, $X^R = -i^*X^L$.
Hint: Use Exercises 37.10 and 37.16.

(b) Defining a second bracket on $\mathfrak{g} = T_eG$ in terms of right-invariant vector fields by $[X,Y]' := [X^R, Y^R](e)$ gives $[X,Y]' = -[X,Y]$ for all $X, Y \in \mathfrak{g}$. In particular, the map $X \mapsto -X$ is a Lie algebra isomorphism from $(\mathfrak{g}, [\ ,\ ])$ to $(\mathfrak{g}, [\ ,\ ]')$.

Exercise 37.20. What can you deduce from Exercise 37.19 about the Lie algebra $\mathfrak{g}$ if the Lie group G is abelian?

37.4. The exponential map. The exponential map was previously defined in terms of geodesics on a pseudo-Riemannian manifold, but in Lie group theory, the same terminology and notation is used for a map

$$\exp : \mathfrak{g} \to G$$

that bears some formal similarity to our previous notion while having nothing intrinsically to do with metrics or geodesics.

Lemma 37.21. *For a smooth path $\gamma : \mathbb{R} \to G$ with $\gamma(0) = e$, the following conditions are equivalent:*

(i) $\gamma(s+t) = \gamma(s)\gamma(t)$ for all $s, t \in \mathbb{R}$;
(ii) γ is a flow line of a left-invariant vector field;
(iii) γ is a flow line of a right-invariant vector field.

Proof. Write $X := \dot\gamma(0) \in T_eG$, and suppose the first condition holds. Then for each $t \in \mathbb{R}$,

$$\dot\gamma(t) = \frac{d}{ds}\gamma(t+s)\bigg|_{s=0} = \frac{d}{ds}L_{\gamma(t)}(\gamma(s))\bigg|_{s=0} = TL_{\gamma(t)}(X) = X^L(\gamma(t))$$

by Exercise 37.14, showing that γ is a flow line of $X^L \in \mathfrak{X}^L(G)$. A similar computation using $\gamma(s+t) = R_{\gamma(t)}(\gamma(s))$ shows that γ is also a flow line of $X^R \in \mathfrak{X}^R(G)$. Conversely, if $\gamma(t) = \varphi^t_{X^L}(e)$, then for any fixed $s \in \mathbb{R}$, we can compare the two paths $\alpha, \beta : \mathbb{R} \to G$ defined by

$$\alpha(t) := \gamma(s+t), \qquad \beta(t) := \gamma(s)\gamma(t) = L_{\gamma(s)}(\gamma(t)),$$

which have the same starting point $\alpha(0) = \beta(0) = \gamma(s)$. Clearly $\dot\alpha(t) = X^L(\alpha(t))$, and since X^L is left-invariant, we also have

$$\dot\beta(t) = TL_{\gamma(s)}(\dot\gamma(t)) = TL_{\gamma(s)}(X^L(\gamma(t))) = X^L(\gamma(s)\gamma(t)) = X^L(\beta(t)),$$

hence α and β are two flow lines of X^L with the same initial value, and are therefore identical. There is again an analogous argument if γ is a flow line of X^R. □

Lemma 37.21 implies that for each $X \in \mathfrak{g} = T_e G$, there exists a unique Lie group homomorphism

$$\gamma_X : \mathbb{R} \to G, \qquad \text{such that} \qquad \dot{\gamma}_X(0) = X,$$

namely the flow line of X^L starting at $e \in G$ (or the corresponding flow line of X^R, which happens to be the same). We can thus define the **exponential map** (*Exponentialabbildung*) of G by

$$\exp : \mathfrak{g} \to G : X \mapsto \gamma_X(1).$$

The proposition implies the formula

$$\exp(X) = \varphi^1_{X^L}(e) = \varphi^1_{X^R}(e), \tag{37.1}$$

which shows that exp is a smooth map. Moreover:

Exercise 37.22. Assume G is a Lie group.

(a) Show that for each $X \in \mathfrak{g}$, the unique Lie group homomorphism $\gamma : \mathbb{R} \to G$ with $\dot{\gamma}(0) = X$ is given by $\gamma(t) = \exp(tX)$.
(b) Generalize (37.1) to the formula

$$\exp(tX) = \varphi^t_{X^L}(e) = \varphi^t_{X^R}(e) \qquad \text{for } X \in \mathfrak{g},\ t \in \mathbb{R}. \tag{37.2}$$

(c) Deduce from the inverse function theorem that $\exp : \mathfrak{g} \to G$ maps a neighborhood of 0 in $\mathfrak{g}$ diffeomorphically onto a neighborhood of e in G.
(d) Generalize (37.2) one step further by showing that for all $g \in G$, $t \in \mathbb{R}$ and $X \in \mathfrak{g}$,

$$\varphi^t_{X^L}(g) = g \exp(tX), \qquad \text{and} \qquad \varphi^t_{X^R}(g) = \exp(tX) g. \tag{37.3}$$

Hint: For a left-invariant vector field, every left translation $L_g : G \to G$ sends flow lines to flow lines.

The Lie group homomorphism $\mathbb{R} \to G : t \mapsto \exp(tX)$ is also called the **1-parameter subgroup** generated by $X \in \mathfrak{g}$. This term is a bit misleading since the map $\mathbb{R} \to G$ need not be injective in general, and its image (even in the injective case) might fail to be an embedded submanifold and thus a Lie subgroup. Nonetheless, the terminology is standard.

With the exponential map as a tool, we can now write down a more direct and revealing formula for the Lie bracket on $\mathfrak{g}$. To start with, suppose M is a manifold, with vector fields $X, Y \in \mathfrak{X}(M)$, a function $f \in C^\infty(M)$ and a point $p \in M$. Recall from Lecture 6 in the first semester that the Lie bracket $[X, Y]$ is the same as the Lie derivative of Y with respect to X, so

$$[X,Y](p) = \mathcal{L}_X Y(p) = \frac{d}{dt} \left((\varphi^t_X)^* Y \right)(p) \bigg|_{t=0} = \frac{d}{dt} T\varphi^{-t}_X \left(Y(\varphi^t_X(p)) \right) \bigg|_{t=0}.$$

To express this in a more useful form, we can use the flow of Y to write $T\varphi^{-t}_X(Y(\varphi^t_X(p))) = \partial_s \left(\varphi^{-t}_X \circ \varphi^s_Y \circ \varphi^t_X(p) \right)\big|_{s=0}$, implying

$$[X,Y](p) = \partial_t \Big(\partial_s \big(\varphi^{-t}_X \circ \varphi^s_Y \circ \varphi^t_X(p) \big)\big|_{s=0} \Big) \Big|_{t=0} \in T_p M, \tag{37.4}$$

in which the expression being differentiated with respect to t is a family of vectors in the fixed tangent space $T_p M$, and thus has a well-defined derivative in $T_p M$. If we now use this to compute $[X, Y] = [X^L, Y^L](e)$ for some elements $X, Y \in \mathfrak{g}$ in the Lie algebra of a Lie group G, applying (37.3) gives

$$[X,Y] = \partial_t \Big(\partial_s \big(\exp(tX) \exp(sY) \exp(-tX) \big)\big|_{s=0} \Big) \Big|_{t=0}. \tag{37.5}$$

Here again, the expression being differentiated with respect to t is a family of tangent vectors in $T_eG = \mathfrak{g}$ and therefore has a well-defined derivative in $\mathfrak{g}$. The formula (37.5) gives an interpretation of the Lie bracket on $\mathfrak{g}$ as a measurement of the failure of elements of G near the identity to commute with each other. In particular, it immediately implies something that was hinted at in Exercise 37.20: if G is abelian, then the bracket on $\mathfrak{g}$ vanishes. This statement also has a converse, though to state it properly, we need to make a distinction between connected and disconnected Lie groups.

Given any topological group G, it is straightforward to check that the path-component

$$G_0 \subset G$$

of G containing the identity element $e \in G$ is a subgroup, i.e. if $g, h \in G$ both admit continuous paths to e, then so do their inverses and their product. This subgroup is called the **identity component** of G. If G is a Lie group, then since the identity component is an open subset, it is also a Lie group, and its Lie algebra is the same as that of G.

LEMMA 37.23. *In a path-connected topological group G, any open neighborhood $\mathcal{U} \subset G$ of $e \in G$ generates G, i.e. every $g \in G$ is a product of finitely-many elements in $\mathcal{U}$.*

EXERCISE 37.24. Prove Lemma 37.23 by showing that any continuous path $\gamma : [0, 1] \to G$ admits a partition $0 = t_0 < t_1 < \ldots < t_{N-1} < t_N = 1$ such that $\gamma(t_j)\gamma(t_{j-1})^{-1} \in \mathcal{U}$ for all $j = 1, \ldots, N$.

THEOREM 37.25. *A connected Lie group G is abelian if and only if the bracket on its Lie algebra $\mathfrak{g}$ is zero.*

PROOF. If G is abelian then (37.5) implies $[X, Y] = 0$ for all $X, Y \in \mathfrak{g}$. Conversely, if the bracket on $\mathfrak{g}$ vanishes, it means that all left-invariant vector fields on G commute with each other, and by the fundamental theorem regarding Lie brackets and commuting flows (see §6.4 from last semester), it follows via (37.3) that for any $X, Y \in \mathfrak{g}$ and $s, t \in \mathbb{R}$,

$$\exp(sX)\exp(tY) = \varphi^t_{Y^L} \circ \varphi^s_{X_L}(e) = \varphi^s_{X^L} \circ \varphi^t_{Y^L}(e) = \exp(tY)\exp(sX).$$

This proves that all elements in some neighborhood of $e \in G$ commute with each other; by Lemma 37.23, these elements also generate G, so the result follows. □

In algebraic terms, a Lie algebra V is called **abelian** if its bracket is zero, i.e. $[u, v] = 0$ for all $u, v \in V$, so Theorem 37.25 can be rephrased as the statement that a connected Lie group is abelian if and only if its Lie algebra is abelian. The next example shows that the restriction of this result to *connected* Lie groups cannot generally be relaxed.

EXAMPLE 37.26. The connected Lie group $\mathrm{SO}(2)$ consists of rotations of $\mathbb{R}^2$, all of which commute with each other, hence $\mathrm{SO}(2)$ is abelian. It is also 1-dimensional, thus its 1-dimensional Lie algebra $\mathfrak{so}(2)$ is automatically abelian due to the antisymmetry of the bracket. On the other hand, $\mathrm{SO}(2)$ is the identity component of the disconnected Lie group $\mathrm{O}(2)$, whose Lie algebra is also $\mathfrak{so}(2)$, but $\mathrm{O}(2)$ is not abelian: rotations and reflections on $\mathbb{R}^2$ do not always commute with each other.

EXERCISE 37.27. Prove that if $X, Y \in \mathfrak{g}$ satisfy $[X, Y] = 0$, then $\exp(X+Y) = \exp(X)\exp(Y) = \exp(Y)\exp(X)$.
Hint: Use flows of left-invariant vector fields to prove $\exp(sX)\exp(tY) = \exp(tY)\exp(sX)$ for all $s, t \in \mathbb{R}$. Then show that $t \mapsto \exp(tX)\exp(tY)$ is a flow line of $(X + Y)^L$.

The next result allows us to interpret the Lie bracket on $\mathfrak{g}$ as a "linearization" of the group structure of G.

THEOREM 37.28. *For any Lie group homomorphism $\Phi : G \to H$, the derivative of Φ at the identity element defines a Lie algebra homomorphism*

$$\Phi_* := T_e\Phi : \mathfrak{g} \to \mathfrak{h}.$$

PROOF. We need to prove $[\Phi_* X, \Phi_* Y] = \Phi_*[X, Y]$ for all $X, Y \in \mathfrak{g}$. Observe first that for any $X \in \mathfrak{g}$, the map $\gamma(t) := \Phi(\exp(tX))$ is a composition of two Lie group homomorphisms and is therefore the unique 1-parameter subgroup $\mathbb{R} \to H$ satisfying $\dot{\gamma}(0) = \Phi_* X$, implying the relation

$$\exp(t\Phi_* X) = \Phi(\exp(tX)).$$

In light of this, (37.5) implies for every $X, Y \in \mathfrak{g}$,

$$\begin{aligned}
[\Phi_* X, \Phi_* Y] &= \partial_t \Big(\partial_s \big(\exp(t\Phi_* X) \exp(s\Phi_* Y) \exp(-t\Phi_* X) \big)\big|_{s=0} \Big)\Big|_{t=0} \\
&= \partial_t \Big(\partial_s \big(\Phi(\exp(tX))\Phi(\exp(sY))\Phi(\exp(-tX)) \big)\big|_{s=0} \Big)\Big|_{t=0} \\
&= \partial_t \Big(\partial_s \big(\Phi(\exp(tX) \exp(sY) \exp(-tX)) \big)\big|_{s=0} \Big)\Big|_{t=0} \\
&= \partial_t \Big(\Phi_* \, \partial_s \big(\exp(tX) \exp(sY) \exp(-tX) \big)\big|_{s=0} \Big)\Big|_{t=0} \\
&= \Phi_* \, \partial_t \Big(\partial_s \big(\exp(tX) \exp(sY) \exp(-tX) \big)\big|_{s=0} \Big)\Big|_{t=0} = \Phi_*[X, Y].
\end{aligned}$$

□

COROLLARY 37.29. *For any Lie subgroup $H \subset G$, $\mathfrak{h} = T_e H$ is a Lie subalgebra of $\mathfrak{g} = T_e G$.* □

EXERCISE 37.30. Another equivalent characterization of the Lie bracket on $\mathfrak{g}$ is given by the formula

$$df([X, Y]) = \partial_s \partial_t \left(f(\exp(sX) \exp(tY)) - f(\exp(tY) \exp(sX)) \right)\big|_{s=t=0} \quad \text{for} \quad f \in C^\infty(G). \tag{37.6}$$

(a) Prove (37.6).
(b) Use (37.6) to give alternative proofs of Theorems 37.25 and 37.28.

37.5. Matrix groups. In many applications it is useful to observe that the definitions of the Lie bracket and exponential map can be simplified when G is a Lie subgroup of $\mathrm{GL}(n, \mathbb{F})$ for some $n \geqslant 0$ and $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$. Let us first note that since $\mathrm{GL}(n, \mathbb{F})$ itself is an open subset in the vector space $\mathbb{F}^{n \times n}$ of all n-by-n matrices, its tangent space $T_{\mathbf{A}} \mathrm{GL}(n, \mathbb{F})$ at every point $\mathbf{A} \in \mathrm{GL}(n, \mathbb{F})$ is canonically identified with $\mathbb{F}^{n \times n}$, and for a Lie subgroup $G \subset \mathrm{GL}(n, \mathbb{F})$ and $\mathbf{A} \in G$, $T_{\mathbf{A}} G$ is then a linear subspace of $\mathbb{F}^{n \times n}$. This implies that vector fields on G are also functions $G \to \mathbb{F}^{n \times n}$, and the Lie algebra $\mathfrak{g}$ of G is also a linear subspace of

$$\mathfrak{gl}(n, \mathbb{F}) = \mathbb{F}^{n \times n}.$$

Note that in the case $\mathbb{F} = \mathbb{C}$, we are still regarding $\mathrm{GL}(n, \mathbb{C})$ as a *real* smooth manifold, and $\mathfrak{g} \subset \mathbb{C}^{n \times n}$ will in general be a *real*-linear subspace, but need not be complex linear.

LEMMA 37.31. *If G is a Lie subgroup of $\mathrm{GL}(n, \mathbb{F})$, then for any $\mathbf{X} \in \mathfrak{g} \subset \mathbb{F}^{n \times n}$, the unique left-invariant vector field $X^L \in \mathfrak{X}^L(G)$ with $X^L(\mathbb{1}) = \mathbf{X}$ is given by*

$$X^L(\mathbf{A}) = \mathbf{A}\mathbf{X}.$$

PROOF. This follows from Exercise 37.14 since for each $\mathbf{A} \in G$, the left-translation diffeomorphism $L_{\mathbf{A}} : G \to G$ is the restriction to $G \subset \mathrm{GL}(n, \mathbb{F}) \subset \mathbb{F}^{n \times n}$ of a linear map $\mathbb{F}^{n \times n} \to \mathbb{F}^{n \times n}$. □

THEOREM 37.32. *On a Lie subgroup $G \subset \mathrm{GL}(n, \mathbb{F})$, the exponential map* $\exp : \mathfrak{g} \to G$ *is given by the matrix exponential*

$$\exp(\mathbf{X}) = e^{\mathbf{X}} := \sum_{k=0}^{\infty} \frac{1}{k!} \mathbf{X}^k \in \mathbb{F}^{n \times n}.$$

PROOF. For $\mathbf{X} \in \mathfrak{g}$ and $t \in \mathbb{R}$, writing $X^L \in \mathfrak{X}^L(G)$ for the left-invariant vector field in Lemma 37.31, the path $\gamma(t) := e^{t\mathbf{X}} \in \mathbb{F}^{n \times n}$ satisfies $\dot{\gamma}(t) = e^{t\mathbf{X}}\mathbf{X} = X^L(\gamma(t))$ and is therefore the unique flow line of X^L with $\gamma(0) = \mathbb{1}$ and $\dot{\gamma}(0) = \mathbf{X}$. According to (37.2), that is the same thing as $\exp(t\mathbf{X})$. □

The following non-obvious consequence of Theorem 37.32 is quite useful:

COROLLARY 37.33. *If G is a Lie subgroup of $\mathrm{GL}(n, \mathbb{F})$ and $\mathbf{X} \in \mathfrak{g} \subset \mathbb{F}^{n \times n}$ belongs to its Lie algebra, then $e^{\mathbf{X}} \in G$.* □

EXAMPLE 37.34. The Lie algebra of $\mathrm{SO}(n) \subset \mathrm{GL}(n, \mathbb{R})$ is the space $\mathfrak{so}(n) \subset \mathbb{R}^{n \times n}$ of antisymmetric matrices, so Corollary 37.33 implies that $e^{\mathbf{A}}$ is orthogonal whenever $\mathbf{A} \in \mathbb{R}^{n \times n}$ is antisymmetric.

THEOREM 37.35. *For a Lie subgroup $G \subset \mathrm{GL}(n, \mathbb{F})$, the bracket on the Lie algebra $\mathfrak{g} \subset \mathbb{F}^{n \times n}$ is given by the matrix commutator*

$$[\mathbf{X}, \mathbf{Y}] = \mathbf{XY} - \mathbf{YX} \in \mathfrak{g} \subset \mathbb{F}^{n \times n}.$$

PROOF. In this situation we can view the right hand side of (37.5) as a mixed second partial derivative of a well-defined $\mathbb{F}^{n \times n}$-valued function of s and t, and expanding by powers of s and t gives

$$\begin{aligned}
[\mathbf{X}, \mathbf{Y}] &= \partial_t \partial_s \big(e^{t\mathbf{X}} e^{s\mathbf{Y}} e^{-t\mathbf{X}} \big)\big|_{s=t=0} \\
&= \partial_t \partial_s \big(\big(\mathbb{1} + t\mathbf{X} + O(t^2)\big) \big(\mathbb{1} + s\mathbf{Y} + O(s^2)\big) \big(\mathbb{1} - t\mathbf{X} + O(t^2)\big)\big|_{s=t=0} \\
&= \partial_t \partial_s \big(\mathbb{1} + s\mathbf{Y} + st(\mathbf{XY} - \mathbf{YX}) + \ldots \big)\big|_{s=t=0} = \mathbf{XY} - \mathbf{YX},
\end{aligned}$$

where the dots in the last line represent a sum of terms that are at least quadratic in either s or t, and thus do not contribute to the relevant derivatve. □

As a non-obvious corollary, one obtains many interesting linear subspaces of the matrix algebra $\mathbb{F}^{n \times n}$ that are closed under the commutator bracket, e.g. this holds for real antisymmetric matrices ($\mathfrak{so}(n)$), complex anti-Hermitian matrices ($\mathfrak{u}(n)$), traceless matrices ($\mathfrak{sl}(n, \mathbb{F})$), and in general any space that arises as the tangent space at $\mathbb{1}$ to a smooth matrix group.

EXERCISE 37.36. Give an alternative proof of Theorem 37.35 using (37.6).

38. Bi-invariance

Recall that a tensor field on a Lie group G is called *bi-invariant* if it is both left- and right-invariant. If G is abelian, then left-invariance and right-invariance are equivalent notions, thus nontrivial bi-invariant tensors always exist, but in the non-abelian case, one should not generally expect to find any. As was mentioned in the previous lecture, this rule has some notable exceptions, and it is now time to discuss them. A useful by-product of this discussion will be a highly non-obvious result about the exponential map, to be proved in the next lecture, implying for instance that every matrix in $\mathrm{SO}(n)$ can be written as $e^{\mathbf{A}}$ for some antisymmetric matrix $\mathbf{A}$.

38.1. The adjoint representation. Given a Lie group G, each element $X \in \mathfrak{g}$ in its Lie algebra corresponds to a specific left-invariant vector field $X^L \in \mathfrak{X}^L(G)$ and right-invariant vector field $X^R \in \mathfrak{X}^R(G)$, given by

$$X^L(g) = TL_g(X), \qquad X^R(g) = TR_g(X).$$

For each $g \in G$, there is then a unique element $\mathrm{Ad}_g(X) := Y \in \mathfrak{g}$ such that $Y^R(g) = X^L(g)$, thus defining a map

$$\mathrm{Ad}_g : \mathfrak{g} \to \mathfrak{g}.$$

We obtain a formula for $\mathrm{Ad}_g(X)$ by writing $Y^R(g) = TR_g(\mathrm{Ad}_g(X)) = X^L(g) = TL_g(X)$, thus

$$\mathrm{Ad}_g(X) = TR_{g^{-1}} \circ TL_g(X) = T(R_{g^{-1}} \circ L_g)(X) =: TC_g(X),$$

where the last expression refers to the derivative at $e \in G$ of the **conjugation** map

$$C_g := R_{g^{-1}} \circ L_g : G \to G : h \mapsto ghg^{-1}.$$

Since every $X \in \mathfrak{g}$ can be written as the derivative of $t \mapsto \exp(tX) \in G$ at $t = 0$, a more direct formula for Ad_g takes the form

$$\mathrm{Ad}_g(X) = \frac{d}{dt}\left(g \exp(tX) g^{-1}\right)\Big|_{t=0}. \tag{38.1}$$

The maps C_g and Ad_g have some algebraic properties that are useful to note. First, for each individual $g \in G$, $C_g : G \to G$ is a Lie group isomorphism, since

$$C_g(ab) = gabg^{-1} = gag^{-1}gbg^{-1} = C_g(a)C_g(b)$$

and $C_{g^{-1}} = C_g^{-1}$. We can thus define a map

$$C : G \to \mathrm{Aut}(G) : g \mapsto C_g,$$

where $\mathrm{Aut}(G)$ denotes the group of Lie group isomorphisms $G \to G$. Second, the map $C : G \to \mathrm{Aut}(G)$ defined in this way is also a group homomorphism, since

$$C_{gh}(a) = gha(gh)^{-1} = ghah^{-1}g^{-1} = C_g(hah^{-1}) = C_g \circ C_h(a).$$

As a consequence, the linear maps $\mathrm{Ad}_g : \mathfrak{g} \to \mathfrak{g}$ can also be packaged together as a group homomorphism

$$\mathrm{Ad} : G \to \mathrm{GL}(\mathfrak{g}) : g \mapsto \mathrm{Ad}_g,$$

where $\mathrm{GL}(\mathfrak{g})$ denotes the group of invertible linear transformations $\mathfrak{g} \to \mathfrak{g}$, and Ad satisfies

$$\mathrm{Ad}_{gh}(X) = TC_{gh}(X) = T(C_g \circ C_h)(X) = TC_g \circ TC_h(X) = \mathrm{Ad}_g \mathrm{Ad}_h(X).$$

Since $\mathfrak{g}$ is a real vector space of some finite dimension $n \geqslant 0$, $\mathrm{GL}(\mathfrak{g})$ is a Lie group isomorphic to $\mathrm{GL}(n, \mathbb{R})$, and $\mathrm{Ad} : G \to \mathrm{GL}(\mathfrak{g})$ is then a Lie group homomorphism. It is known as the **adjoint representation** (*adjungierte Darstellung*) of G.

This would be a good moment for a brief digression on representation theory. For a general vector space V over $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, let us denote

$$\mathfrak{gl}(V) := \left\{A : V \to V \;\middle|\; A \text{ is } \mathbb{F}\text{-linear}\right\}, \qquad \mathrm{GL}(V) := \left\{A \in \mathfrak{gl}(V) \;\middle|\; A \text{ is a bijection}\right\}.$$

The latter is a group with respect to composition of maps, and we can make $\mathfrak{gl}(V)$ into a Lie algebra by endowing it with the commutator bracket

$$[A, B] := AB - BA \in \mathfrak{gl}(V) \qquad \text{for } A, B \in \mathfrak{gl}(V).$$

If V has finite dimension $n \geqslant 0$, then $\mathrm{GL}(V)$ is an open subset of the finite-dimensional vector space $\mathfrak{gl}(V)$ and thus carries a natural smooth structure that makes it a Lie group; moreover, its Lie algebra is $\mathfrak{gl}(V)$ with the commutator bracket, as one sees by choosing a basis and applying Theorem 37.35.

DEFINITION 38.1. Given a vector space V over $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ and a group G, a **representation** (*Darstellung*) of G on V is a group homomorphism

$$\rho : G \to \mathrm{GL}(V).$$

If V is endowed with an inner product, then the representation ρ is called **orthogonal** (for $\mathbb{F} = \mathbb{R}$) or **unitary** (for $\mathbb{F} = \mathbb{C}$) if $\rho(g) : V \to V$ preserves the inner product for every $g \in G$. If G is a Lie group and V is finite dimensional, we generally also require $\rho : G \to \mathrm{GL}(V)$ to be a smooth map, and thus a Lie group homomorphism.

Similarly, a **representation of a Lie algebra** $\mathfrak{g}$ on V is a Lie algebra homomorphism

$$\rho : \mathfrak{g} \to \mathfrak{gl}(V),$$

where $\mathfrak{gl}(V)$ is endowed with the commutator bracket.

In representation theory, one most frequently considers representations on $\mathbb{F}^n$, which assign to each element $g \in G$ a matrix $\rho(g) \in \mathrm{GL}(n, \mathbb{F})$ so that the group multiplication law becomes matrix multiplication. The adjoint representation $\mathrm{Ad} : G \to \mathrm{GL}(\mathfrak{g})$ can be understood from this perspective as well after choosing a basis of $\mathfrak{g}$, though in general the choice of basis is not canonical. By Theorem 37.28, every finite-dimensional Lie group representation $\rho : G \to \mathrm{GL}(V)$ can be differentiated at $e \in G$ to define a Lie algebra representation $\rho_* : \mathfrak{g} \to \mathfrak{gl}(V)$. Applying this to $\mathrm{Ad} : G \to \mathrm{GL}(\mathfrak{g})$ thus gives a Lie algebra representation

$$\mathrm{ad} : \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g}) : X \mapsto \mathrm{ad}_X .$$

An explicit formula for $\mathrm{ad}_X : \mathfrak{g} \to \mathfrak{g}$ can be deduced from (38.1) and (37.5), namely

$$\mathrm{ad}_X Y = \frac{d}{ds} \mathrm{Ad}_{\exp(sX)}(Y)\Big|_{s=0} = \partial_s \Big(\partial_t \big(\exp(sX) \exp(tY) \exp(-sX) \big)\big|_{t=0} \Big)\Big|_{s=0} = [X, Y]. \tag{38.2}$$

The fact that ad is a Lie algebra homomorphism thus means that for all $X, Y, Z \in \mathfrak{g}$,

$$\mathrm{ad}_{[X,Y]} Z = [[X, Y], Z] = \mathrm{ad}_X \mathrm{ad}_Y Z - \mathrm{ad}_Y \mathrm{ad}_X Z = [X, [Y, Z]] - [Y, [X, Z]],$$

which is equivalent to the Jacobi identity.

Let us say that an element $X \in \mathfrak{g}$ is Ad-**invariant** if $\mathrm{Ad}_g(X) = X$ for all $g \in G$. Returning to the origin of this discussion, the definition of $\mathrm{Ad} : G \to \mathrm{GL}(\mathfrak{g})$ immediately gives rise to the following application:

PROPOSITION 38.2. *For $X \in \mathfrak{g}$, the associated left-invariant vector field $X^L \in \mathfrak{X}^L(G)$ is also bi-invariant if and only if X is* Ad*-invariant.* □

EXERCISE 38.3. For a Lie group G and two elements $X, Y \in \mathfrak{g}$ in its Lie algebra, prove

$$\begin{aligned} \mathrm{Ad}_{\exp(X)}(Y) = e^{\mathrm{ad}_X}(Y) &:= \left(\mathbb{1} + \mathrm{ad}_X + \frac{1}{2!}(\mathrm{ad}_X)^2 + \frac{1}{3!}(\mathrm{ad}_X)^3 + \ldots \right) Y \\ &= Y + [X, Y] + \frac{1}{2!}[X, [X, Y]] + \frac{1}{3!}[X, [X, [X, Y]]] + \ldots . \end{aligned}$$

38.2. Bi-invariant metrics. If G is abelian, then the adjoint representation is trivial, meaning $\mathrm{Ad}_g \in \mathrm{GL}(\mathfrak{g})$ is the identity map for every $g \in G$, and every $X \in \mathfrak{g}$ is therefore Ad-invariant; in light of Proposition 38.2, this is consistent with our initial observation that all left-invariant vector fields in this case are also bi-invariant. For a non-abelian group, there is generally no reason to expect the existence of any nontrivial Ad-invariant elements in $\mathfrak{g}$. One gets better results however by focusing on certain other algebraic structures that can be imposed on $\mathfrak{g}$, for instance, inner products. Any nondegenerate symmetric bilinear form $\langle\ ,\ \rangle$ on $\mathfrak{g}$ gives rise to unique left-invariant and right-invariant pseudo-Riemannian metrics $\langle\ ,\ \rangle^L, \langle\ ,\ \rangle^R \in \Gamma(T_2^0 G)$ on G, characterized by the relations

$$\langle X, Y\rangle = \langle X^L(g), Y^L(g)\rangle^L = \langle X^R(g), Y^R(g)\rangle^R \qquad \text{for all } g \in G,\ X, Y \in \mathfrak{g},$$

and these metrics are positive if and only if $\langle\ ,\ \rangle$ is a positive inner product on $\mathfrak{g}$. In light of the relation $(\mathrm{Ad}_g X)^R(g) = X^L(g)$, the two metrics $\langle\ ,\ \rangle^L$ and $\langle\ ,\ \rangle^R$ are then identical at some point $g \in G$ if and only if $\langle X, Y\rangle = \langle \mathrm{Ad}_g(X), \mathrm{Ad}_g(Y)\rangle$ for all $X, Y \in \mathfrak{g}$. We say that a scalar-valued bilinear form $\langle\ ,\ \rangle$ on $\mathfrak{g}$ is **Ad-invariant** if $\langle \mathrm{Ad}_g(X), \mathrm{Ad}_g(Y)\rangle = \langle X, Y\rangle$ holds for all $g \in G$ and $X, Y \in \mathfrak{g}$. This discussion proves:

PROPOSITION 38.4. *A left- or right-invariant pseudo-Riemannian metric on G is bi-invariant if and only if its restriction to a bilinear form on $T_e G = \mathfrak{g}$ is* Ad*-invariant.* □

This result implies that a Lie group G admits a bi-invariant Riemannian metric if and only if its Lie algebra $\mathfrak{g}$ admits an Ad-invariant inner product. We'll come back to the existence question for Ad-invariant inner products in §38.4 below.

EXERCISE 38.5. For $\mathbf{X}, \mathbf{Y} \in \mathbb{C}^{n\times n}$, define the symmetric bilinear pairing

$$\langle \mathbf{X}, \mathbf{Y}\rangle := -\operatorname{tr}(\mathbf{XY}) \in \mathbb{C}.$$

For any Lie subgroup $G \subset \mathrm{GL}(n,\mathbb{C})$, restricting $\langle\ ,\ \rangle$ to the real subspace $\mathfrak{g} \subset \mathfrak{gl}(n,\mathbb{C}) = \mathbb{C}^{n\times n}$ defines a symmetric complex-valued real-bilinear form on $\mathfrak{g}$. (Note that since $\mathrm{GL}(n,\mathbb{R})$ is naturally a subgroup of $\mathrm{GL}(n,\mathbb{C})$, this includes the Lie subgroups of $\mathrm{GL}(n,\mathbb{R})$, on which the pairing will automatically be real valued.) Prove:

(a) For every Lie subgroup $G \subset \mathrm{GL}(n,\mathbb{C})$, the pairing $\langle\ ,\ \rangle$ on $\mathfrak{g}$ is Ad-invariant.
(b) For $G = \mathrm{SO}(n)$ or $\mathrm{O}(n)$, $\langle\ ,\ \rangle$ defines an Ad-invariant (positive) inner product on $\mathfrak{so}(n)$.
(c) For $G = \mathrm{U}(n)$ or $\mathrm{SU}(n)$, $\langle\ ,\ \rangle$ is real-valued on $\mathfrak{u}(n)$ or $\mathfrak{su}(n)$ respectively and also defines an Ad-invariant inner product.[87]
Hint: Compare $\operatorname{tr}(\mathbf{XY})$ *with* $\operatorname{tr}\left((\mathbf{XY})^\dagger\right)$.
(d) For $n = 2$ and $G = \mathrm{SL}(2,\mathbb{R})$ or $\mathrm{GL}(2,\mathbb{R})$, $\langle\ ,\ \rangle$ defines a nondegenerate symmetric bilinear form on $\mathfrak{sl}(2,\mathbb{R})$ or $\mathfrak{gl}(2,\mathbb{R})$ respectively. What is its signature?
Hint: There might be cleverer methods, but it isn't too hard to guess an explicit orthonormal basis of $\mathfrak{sl}(2,\mathbb{R})$ with respect to $\langle\ ,\ \rangle$, and then extend it to an orthonormal basis of $\mathfrak{gl}(2,\mathbb{R})$.

38.3. Haar measures and the modular function. We next consider how to define integrals of functions on an n-dimensional Lie group G for some $n \in \mathbb{N}$. As shown in Exercise 37.15, G is parallelizable, and therefore also orientable, so it admits volume forms. The most natural choice of volume form will of course be one that is left- and/or right-invariant, and Proposition 37.13 also guarantees a kind of uniqueness for these: for every $\omega \in \Lambda^n T_e^* G$, there are unique left-invariant and right-invariant n-forms $\omega^L, \omega^R \in \Omega^n(G)$ that match ω at $e \in G$, and both will necessarily be volume forms if $\omega \neq 0$. If we assume the latter, orient G so that $\omega^L > 0$ and then define volumes of measurable regions $\mathcal{U} \subset G$ by

$$\mu(\mathcal{U}) := \int_{\mathcal{U}} \omega^L$$

the result is a measure μ on G that is left-invariant in the sense that for each measurable subset $\mathcal{U} \subset G$,

$$\mu(L_g(\mathcal{U})) = \mu(\mathcal{U}) \qquad \text{for all } g \in G,$$

[87]This is a good moment to emphasize that while $\mathfrak{u}(n)$ and $\mathfrak{su}(n)$ naturally live inside $\mathfrak{gl}(n,\mathbb{C}) = \mathbb{C}^{n\times n}$, they are not complex but only *real* subspaces, so inner products on these subspaces are required to be $\mathbb{R}$-bilinear, not sesquilinear.

and the integrals of functions f on G therefore satisfy

$$\int_G (f \circ L_g)\, d\mu = \int_G f\, d\mu \qquad \text{for all } g \in G.$$

This follows easily from the left-invariance of ω^L since $\int_G (f \circ L_g)\, d\mu = \int_G (f \circ L_g)\, \omega^L = \int_G L_g^*(f\, \omega^L) = \int_{L_g(G)} f\, \omega^L = \int_G f\, \omega^L = \int_G f\, d\mu$; note that in defining the orientation of G so that $\omega^L > 0$, we have ensured that the diffeomorphism $L_g : G \to G$ is orientation preserving for every $g \in G$. The left-invariant measure μ defined on G in this way is called a **left Haar measure**. It is not unique, but since $\dim \Lambda^n T_e^* G = 1$, all possible choices of μ are related to each other by multiplication with a positive constant, and if G is compact, one can single out a canonical choice by requiring

$$\mu(G) = \int_G \omega^L = 1.$$

In the compact case we refer to "the" left Haar measure as the particular choice of μ that satisfies this condition. Note that there are *two* possible choices of the underlying left-invariant volume form ω^L used to define μ; one can replace ω^L with $-\omega^L$ without changing the value of integrals such as $\int_{\mathcal{U}} f\, d\mu := \int_{\mathcal{U}} f\, \omega^L$, as doing so also reverses the orientation of G and thus inserts a sign to cancel the change in ω^L.

For everything said above about left Haar measures, one could equally well replace "left" with "right" and obtain a similarly sensible notion, namely a **right Haar measure**, which satisfies

$$\int_G (f \circ R_g)\, d\mu = \int_G f\, d\mu \qquad \text{for all } g \in G,$$

and which can again be fixed uniquely via the normalization condition $\mu(G) = 1$ if G is compact. If a left Haar measure is also a right Haar measure, it is simply called a **Haar measure**, and we will show in a moment that this always holds if G is compact. When there is no danger of confusion, it is common to write the integral of a function f on G with respect to a Haar measure as

$$\int_G f(g)\, dg := \int_G f\, d\mu.$$

The question now is this: when does a bi-invariant Haar measure exist, or equivalently, when is a left Haar measure also a right Haar measure?

The naive way to think about this question turns out to be wrong: one would hope at first to solve the problem by finding a bi-invariant volume form, or equivalently, finding conditions under which a left-invariant volume form must also be right-invariant. The reason this won't work as often as one might like has to do with orientations: every Lie group is orientable, but not all choices of orientation are equally good, e.g. not all are invariant under both right and left translations. Let us say that an orientation of G is **left-invariant** if the diffeomorphisms $L_g : G \to G$ are all orientation preserving; the notions of **right-invariant** and **bi-invariant orientations** can be defined similarly. On a *connected* Lie group, every orientation is bi-invariant, because the existence of continuous paths from arbitrary $g \in G$ to $e \in G$ means that all of the diffeomorphisms $L_g, R_g : G \to G$ can be deformed through continuous families of diffeomorphisms to the identity map, which is orientation preserving. But on a disconnected group, orientations can be switched independently on separate connected components, and this can affect whether they are left- and/or right-invariant. Clearly, a bi-invariant volume form determines a bi-invariant orientation. It turns out however that some of the simplest interesting Lie groups do not admit any bi-invariant orientation:

EXERCISE 38.6. On $\mathrm{O}(2)$, let o^L and o^R denote choices of left- and right-invariant orientations respectively. Show that if o^L and o^R match on the identity component $\mathrm{SO}(2) \subset \mathrm{O}(2)$, then they differ on the component containing the reflection $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$.

The conundrum of invariant orientations is the secret reason why we have been focusing our attention on invariant *measures*[88] rather than volume forms: a measure defined via a volume form can be bi-invariant even if the volume form itself is not. Two smooth volume forms $\omega, \omega' \in \Omega^n(G)$ define the same measure on G if and only if $\omega = \pm\omega'$, where the sign $\pm$ will be locally constant but may differ on different connected components of G. Given a measure defined via the volume form ω, a diffeomorphism $\varphi : G \to G$ will thus preserve this measure if and only if $\varphi^*\omega = \pm\omega$, where again the sign $\pm$ is locally constant but may vary.

Assume now that G is a Lie group, pick a nonzero element $\omega \in \Lambda^n T_e^* G$ and let $\omega^L, \omega^R \in \Omega^n(G)$ denote the unique left-invariant and right-invariant volume forms respectively such that $\omega^L_e = \omega^R_e = \omega$. This means by definition that for every $g \in G$, and $X_1, \dots, X_n \in \mathfrak{g}$,

$$\omega(X_1, \dots, X_n) = \omega^L_g(X_1^L(g), \dots, X_n^L(g)) = \omega^R_g(X_1^R(g), \dots, X_n^R(g)),$$

and the definition of the adjoint representation also implies

$$\omega(\mathrm{Ad}_g(X_1), \dots, \mathrm{Ad}_g(X_n)) = \omega^R_g(X_1^L(g), \dots, X_n^L(g)).$$

Since ω is a top-dimensional form on $\mathfrak{g}$, a relationship between the left hand sides of these two equations is given by

$$\omega(\mathrm{Ad}_g(X_1), \dots, \mathrm{Ad}_g(X_n)) = \det(\mathrm{Ad}_g) \cdot \omega(X_1, \dots, X_n),$$

where $\det(\mathrm{Ad}_g) \in \mathbb{R}$ is by definition the determinant of the matrix representing $\mathrm{Ad}_g : \mathfrak{g} \to \mathfrak{g}$ in any choice of basis for $\mathfrak{g}$. The resulting relation between ω^L and ω^R is

$$\omega^R_g = \det(\mathrm{Ad}_g) \cdot \omega^L_g, \tag{38.3}$$

and we conclude that the volume forms ω^L and ω^R define the same measure on G if and only if $\det(\mathrm{Ad}_g) = \pm 1$ for every $g \in G$. Notice that if this is true for one choice of nonzero $\omega \in \Lambda^n T_e^* G$, then it will be true for all of them, as they are all related to each other by nonzero constant factors. It is traditional to express this condition in terms of the so-called **modular function**

$$\Delta : G \to \mathbb{R}_{>0} : g \mapsto |\det(\mathrm{Ad}_g)|,$$

which is the composition of the two Lie group homomorphisms $\mathrm{Ad} : G \to \mathrm{GL}(\mathfrak{g})$ and $|\det| : \mathrm{GL}(\mathfrak{g}) \to \mathbb{R}_{>0}$ and is therefore also a Lie group homomorphism; here the set of positive real numbers $\mathbb{R}_{>0}$ is regarded as a Lie group with respect to multiplication. The Lie group G is called **unimodular** if its modular function is identically equal to 1. The preceding discussion proves:

THEOREM 38.7. *The following conditions on a Lie group G are equivalent:*

(i) G admits a bi-invariant Haar measure.
(ii) Every left or right Haar measure on G is bi-invariant.
(iii) G is unimodular.

□

To make this result truly useful, we have:

THEOREM 38.8. *Every compact Lie group is unimodular.*

[88] One could equivalently frame the discussion in terms of invariant volume *elements*, which do not require any choice of orientation in order to define integrals. Two volume forms determine the same volume element if and only if they differ at most by a sign at each point; see §11.4 in the notes from last semester.

Proof. The image of the modular function $\Delta : G \to \mathbb{R}_{>0}$ is a subgroup of $\mathbb{R}_{>0}$, and it is also compact if G is compact, since Δ is continuous. But the only compact subgroup of $\mathbb{R}_{>0}$ is the trivial one. $\square$

Exercise 38.9. Prove the following alternative characterization of the modular function: for any left-invariant volume form $\omega^L \in \Omega^n(G)$ and every $g \in G$, $\omega^L \equiv \det(\mathrm{Ad}_g) \cdot R_g^* \omega^L$.
Hint: Show first that $R_g^*\omega^L$ is also left-invariant for each $g \in G$.

Exercise 38.10. Prove that every Lie group admitting a bi-invariant pseudo-Riemannian metric is also unimodular.
Hint: Every matrix in $\mathrm{O}(k,\ell)$ has determinant ± 1.

Remark 38.11. Exercises 38.10 and 38.5(d) together imply that $\mathrm{SL}(2,\mathbb{R})$ and $\mathrm{GL}(2,\mathbb{R})$ are unimodular. There are in fact plenty of Lie groups that are noncompact and non-abelian but nonetheless unimodular!

Exercise 38.12. A popular example of a Lie group that is *not* unimodular is the following connected subgroup of the affine group on $\mathbb{R}$ (see Example 37.9):

$$\mathrm{Aff}_+(\mathbb{R}) := \left\{ \varphi \in \mathrm{Diff}(\mathbb{R}) \;\middle|\; \varphi(t) = at + b \text{ for some } a > 0 \text{ and } b \in \mathbb{R} \right\}.$$

There is a global chart (x,y) identifying $\mathrm{Aff}_+(\mathbb{R})$ with the upper half-plane $\{y > 0\} \subset \mathbb{R}^2$ such that a point (x,y) is identified with the affine transformation $t \mapsto yt + x$. The identity $\mathrm{Id} \in \mathrm{Aff}_+(\mathbb{R})$ thus has coordinates $(x,y) = (0,1)$.

(a) Find the unique functions $f^L, f^R : \mathrm{Aff}_+(\mathbb{R}) \to (0,\infty)$ such that $f^L(\mathrm{Id}) = f^R(\mathrm{Id}) = 1$ and the volume forms $f^L\, dx \wedge dy, f^R\, dx \wedge dy \in \Omega^2(\mathrm{Aff}_+(\mathbb{R}))$ are left-invariant and right-invariant respectively.
(b) Show that the modular function $\Delta : \mathrm{Aff}_+(\mathbb{R}) \to \mathbb{R}_{>0}$ is given by $\Delta(x,y) = y$.

38.4. Applications of the Haar measure. The most popular application of Haar integrals is an averaging trick that underlies several fundamental results in representation theory, starting with the following:

Theorem 38.13. *Suppose G is a compact Lie group and $\rho : G \to \mathrm{GL}(V)$ is a representation of G on a vector space V over $\mathbb{F} \in \{\mathbb{R},\mathbb{C}\}$. Then V can be equipped with a positive inner product that makes ρ orthogonal (if $\mathbb{F} = \mathbb{R}$) or unitary (if $\mathbb{F} = \mathbb{C}$).*

Proof. Choose any inner product $(\ ,\ )$ on V, along with a right Haar measure on G, and define a new inner product $\langle\ ,\ \rangle$ on V by

$$\langle v, w\rangle := \int_G (\rho(g)v, \rho(g)w)\, dg.$$

It is easy to check that this is an inner product; in particular, $\langle v, v\rangle > 0$ for all $v \neq 0 \in V$ since the integrand is then positive for all $g \in G$.[89] For every $h \in G$, one then has

$$\begin{aligned}\langle \rho(h)v, \rho(h)w\rangle &= \int_G (\rho(g)\rho(h)v, \rho(g)\rho(h)w)\, dg = \int_G (\rho(gh)v, \rho(gh)w)\, dg = \int_G (\rho(g)v, \rho(g)w)\, dg \\ &= \langle v, w\rangle,\end{aligned}$$

where we've used the right-invariance relation $\int_G f(gh)\, dg = \int_G f(g)\, dg$. $\square$

We have an immediate application of this theorem that has nothing intrinsically to do with representation theory: applying it to the adjoint representation and recalling Proposition 38.4 gives

[89]Positivity of the inner product is explicitly required here—one cannot do this trick in general with indefinite nondegenerate bilinear forms.

Corollary 38.14. *For every compact Lie group G, the Lie algebra $\mathfrak{g}$ admits an* Ad*-invariant positive inner product, and G therefore admits a bi-invariant Riemannian metric.* □

Remark 38.15. Corollary 38.14 obviously also holds for groups that are not necessarily compact but abelian, e.g. vector spaces with their additive structure, and therefore also for any group that is the Cartesian product of a compact Lie group with a vector space. A theorem of Milnor [**Mil76**] states that *every* Lie group admitting a bi-invariant Riemannian metric is in fact of this form.

Exercise 38.16. A Lie algebra $\mathfrak{g}$ is called **simple** if it contains no nontrivial proper Lie subalgebra $\mathfrak{h} \subset \mathfrak{g}$ such that $[X, Y] \in \mathfrak{h}$ for every $X \in \mathfrak{g}$ and $Y \in \mathfrak{h}$. In particular, if G is a Lie group whose Lie algebra $\mathfrak{g}$ is simple, then no nontrivial proper subspace of $\mathfrak{g}$ is invariant under the map $\mathrm{ad}_X : \mathfrak{g} \to \mathfrak{g}$ for every $X \in \mathfrak{g}$. Prove:

(a) If a Lie group G contains a normal Lie subgroup $H \subset G$ with $0 < \dim H < \dim G$, then its Lie algebra is not simple.
(b) The hypothesis of part (a) applies e.g. to $\mathrm{GL}(n, \mathbb{F})$ and $\mathrm{U}(n)$ for all $n \geqslant 2$.
(c) $\mathfrak{so}(3)$ and $\mathfrak{su}(2)$ are simple.
(d) If G is a Lie group admitting a bi-invariant Riemannian metric and its Lie algebra $\mathfrak{g}$ is simple, then its bi-invariant metric is unique up to positive scaling.
Hint: If $\langle\ ,\ \rangle$ and $\langle\ ,\ \rangle'$ are two Ad*-invariant inner products on $\mathfrak{g}$, then $\langle X, Y\rangle' = \langle X, AY\rangle$ for a linear map $A : \mathfrak{g} \to \mathfrak{g}$ that is symmetric with respect to $\langle\ ,\ \rangle$ and commutes with Ad_g for every $g \in G$. Deduce from the latter that for every $X \in \mathfrak{g}$, $\mathrm{ad}_X : \mathfrak{g} \to \mathfrak{g}$ preserves each eigenspace of A, and conclude that there can only be one eigenspace.*
(e) If G is a connected compact Lie group whose Lie algebra is not simple, then G admits two bi-invariant Riemannian metrics that are not scalar multiples of each other.
Hint: If $\mathfrak{h} \subset \mathfrak{g}$ is invariant under ad_X for every $X \in \mathfrak{g}$, show that the same is true for its orthogonal complement $\mathfrak{h}^\perp \subset \mathfrak{g}$ with respect to any Ad*-invariant inner product. Then argue that $\mathfrak{h}$ and $\mathfrak{h}^\perp$ are each invariant under the adjoint representation of G, so you are free to rescale an* Ad*-invariant inner product independently on the two factors.*

You may have noticed that the proof of Theorem 38.13 didn't actually require the bi-invariance of the Haar measure; right-invariance was enough. The next two exercises do make use of bi-invariance.

Exercise 38.17. Prove: on a unimodular Lie group G, any Haar measure satisfies
$$\int_G f(g)\, dg = \int_G f(g^{-1})\, dg$$
for compactly-supported smooth functions f on G.
Hint: If μ^L is a left-invariant volume form and G is oriented so that $\mu^L > 0$, let G' denote the same group but with a possibly different orientation chosen so that the inversion map $i : G' \to G : g \mapsto g^{-1}$ is orientation preserving. What can you now say about $i^*\mu^L$? (See Exercise 37.16.)

Exercise 38.18. Given a Lie group G and representation $\rho : G \to \mathrm{GL}(V)$, consider the linear subspace
$$V^\rho := \left\{ v \in V \;\middle|\; \rho(g)v = v \text{ for all } g \in G \right\}.$$
If G is compact, we can use the Haar measure to define a linear map
$$\Pi_\rho : V \to V : v \mapsto \int_G \rho(g)v\, dg.$$

(a) Show that Π_ρ is the projection of V onto V^ρ along a complementary subspace $W^\rho \subset V$ that is also G-invariant, meaning $\rho(g)(W^\rho) = W^\rho$ for every $g \in G$.

(b) If V carries an inner product such that the representation ρ is orthogonal/unitary, show that Π_ρ is self-adjoint.

(c) Suppose V is finite dimensional, and $\{\rho_\tau : G \to \mathrm{GL}(V)\}_{\tau\in\mathbb{R}^d}$ is a family of representations depending smoothly on the parameter $\tau \in \mathbb{R}^d$ (in other words, the map $\mathbb{R}^d \times G \to \mathrm{GL}(V) : (\tau, g) \mapsto \rho_\tau(g)$ is smooth). Prove that

$$\bigcup_{\tau\in\mathbb{R}^d} \{\tau\} \times V^{\rho_\tau} \subset \mathbb{R}^d \times V$$

is then a smooth subbundle of the trivial vector bundle $\mathbb{R}^d \times V$ over $\mathbb{R}^d$.
Hint: For a fixed $\sigma \in \mathbb{R}^d$ and $\tau \in \mathbb{R}^d$ nearby, what can you say about the linear maps $V^{\rho_\sigma} \oplus W^{\rho_\sigma} \to V : (v, w) \mapsto \Pi_{\rho_\tau} v + (\mathbb{1} - \Pi_{\rho_\tau})w$?

39. Geometry and topology of Lie groups

39.1. Bi-invariant Riemannian geometry. Now that we know bi-invariant Riemannian metrics exist on a substantial subclass of the Lie groups we are interested in, it is worth asking what kinds of *Riemannian* manifolds these Lie groups are. This ties in with another important question: if G is a Lie group carrying a pseudo-Riemannian metric, then the symbol exp has two possible interpretations, one based on geodesics and the other based on 1-parameter subgroups $\mathbb{R} \to G$. It would be nice to know whether these two versions of the exponential map are the same thing. In general, they will not be unless there is a condition relating the metric to the Lie group structure. Left-invariance or right-invariance would be natural conditions to impose, but we will see in Exercise 39.13 that neither on its own is sufficent. It turns out that together, they are:

Theorem 39.1. *On a Lie group G endowed with a bi-invariant pseudo-Riemannian metric, every flow line of a left- or right-invariant vector field is also a geodesic, and conversely, every geodesic is also a flow line of both a left-invariant and a right-invariant vector field.*

Before proving this statement, let's examine some of its consequences. One of them derives from Exercise 37.17:

Corollary 39.2. *Every Lie group with a bi-invariant pseudo-Riemannian metric is geodesically complete.* □

Remark 39.3. It's worth noting that for metrics with positive signature, geodesic completeness also follows from left- or right-invariance alone; bi-invariance is not required. This is because if $\langle\ ,\ \rangle$ is (say) left-invariant, then the diffeomorphisms $L_g : G \to G$ are isometries for every g, and as a consequence, G admits an isometry carrying any point to any other point, implying that its injectivity radius is constant. As noted in Exercise 34.23, $\mathrm{inj}(M, g) > 0$ implies that (M, g) is geodesically complete.

Now suppose G is connected and carries a bi-invariant Riemannian metric $\langle\ ,\ \rangle$. Geodesic completeness implies via the Hopf-Rinow theorem that any two points in G are connected by a geodesic: in particular, there exist geodesics connecting $e \in G$ to any other point $g \in G$, and as flow lines of left-invariant vector fields, such geodesics take the form $\exp(tX)$ for $X \in \mathfrak{g}$. This implies:

Corollary 39.4. *For any connected Lie group G admitting a bi-invariant Riemannian metric, the map $\exp : \mathfrak{g} \to G$ is surjective.* □

By Corollary 38.14, this result applies to every *compact* connected Lie group: popular examples include $\mathrm{SO}(n)$, $\mathrm{U}(n)$ and $\mathrm{SU}(n)$. The next example shows that it is not true for arbitrary connected Lie groups.

Example 39.5. Prove:

(a) If G is a Lie group and $g \in G$ is in the image of $\exp : \mathfrak{g} \to G$, then $g = h^2$ for some $h \in G$.

(b) The matrices $\begin{pmatrix} -1 & 0 \\ 0 & -2 \end{pmatrix}$ and $\begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix}$ both lie in the identity component of $\mathrm{GL}(2, \mathbb{R})$ but not in the image of $\exp : \mathfrak{gl}(2, \mathbb{R}) \to \mathrm{GL}(2, \mathbb{R})$.
Hint: What kind of spectra could their square roots have?

(c) If $\mathbf{A} \in \mathfrak{sl}(2, \mathbb{R})$ has two negative eigenvalues not equal to -1, then it is not in the image of $\exp : \mathfrak{sl}(2, \mathbb{R}) \to \mathrm{SL}(2, \mathbb{R})$.

Remark 39.6. The result of Exercise 38.5(d) implies that $\mathrm{GL}(2, \mathbb{R})$ and $\mathrm{SL}(2, \mathbb{R})$ both admit bi-invariant pseudo-Riemannian metrics of Lorentzian signature. This illustrates yet another failure of the Hopf-Rinow theorem in the pseudo-Riemannian setting: existence of an *indefinite* bi-invariant pseudo-Riemannian metric on a connected Lie group does not suffice to make $\exp : \mathfrak{g} \to G$ surjective.

Remark 39.7. On the other hand, one can use the Jordan canonical form to show that $\exp : \mathfrak{gl}(n, \mathbb{C}) \to \mathrm{GL}(n, \mathbb{C})$ is surjective for every $n \in \mathbb{N}$, in spite of noncompactness (see [War83, Chapter 3, Exercise 15]).

The proof of Theorem 39.1 hinges on a formula for covariant derivatives of left- or right-invariant vector fields in the presence of a bi-invariant metric. Let us first recall a relation that was derived in the course of proving the existence and uniqueness of the Levi-Cività connection last semester (see Equation (22.2)):

Lemma 39.8 (Koszul formula). *On any pseudo-Riemannian manifold (M, g) with Levi-Cività connection ∇, the relation*

$$\begin{aligned} 2\langle \nabla_X Y, Z \rangle = & \mathcal{L}_X \langle Y, Z \rangle + \mathcal{L}_Y \langle Z, X \rangle - \mathcal{L}_Z \langle X, Y \rangle \\ & + \langle [X, Y], Z \rangle - \langle [X, Z], Y \rangle - \langle [Y, Z], X \rangle \end{aligned}$$

holds for all $X, Y, Z \in \mathfrak{X}(M)$. □

Lemma 39.9. *If $\langle\ ,\ \rangle$ is the restriction to $\mathfrak{g} = T_e G$ of a bi-invariant pseudo-Riemannian metric, then for every $X \in \mathfrak{g}$, the map $\mathrm{ad}_X : \mathfrak{g} \to \mathfrak{g}$ is antisymmetric with respect to this pairing.*

Proof. By Proposition 38.4, $\mathrm{Ad}_g : \mathfrak{g} \to \mathfrak{g}$ is orthogonal for every $g \in G$, and differentiating the relation $\langle \mathrm{Ad}_{\exp(tX)}(Y), \mathrm{Ad}_{\exp(tX)}(Z) \rangle = \langle Y, Z \rangle$ at $t = 0$ gives $\langle \mathrm{ad}_X Y, Z \rangle + \langle Y, \mathrm{ad}_X Z \rangle = 0$. □

Exercise 39.10. Lemma 39.9 has a converse of sorts: if G is connected and $\langle\ ,\ \rangle$ is a (possibly indefinite but nondegenerate) inner product on $\mathfrak{g}$ for which $\mathrm{ad}_X : \mathfrak{g} \to \mathfrak{g}$ is antisymmetric for every $X \in \mathfrak{g}$, then $\langle\ ,\ \rangle$ is Ad-invariant. Prove it!
Hint: Think about the image of the Lie group homomorphism $\mathrm{Ad} : G \to \mathrm{GL}(\mathfrak{g})$.

Proposition 39.11. *If $\langle\ ,\ \rangle$ is a bi-invariant pseudo-Riemannian metric on G with Levi-Cività connection ∇ and $X, Y \in \mathfrak{X}(G)$ are either both left-invariant or both right-invariant, then $\nabla_X Y = \frac{1}{2}[X, Y]$.*

Proof. Assume $X, Y \in \mathfrak{X}^L(G)$, and choose a third left-invariant vector field $Z \in \mathfrak{X}^L(G)$. Since $\langle\ ,\ \rangle$ is also left-invariant, the pairings $\langle X, Y \rangle$, $\langle Y, Z \rangle$ and $\langle X, Z \rangle$ are all constant functions on G, and using (38.2) and Lemma 39.9, the Koszul formula in Lemma 39.8 becomes

$$\begin{aligned} 2\langle \nabla_X Y, Z \rangle &= \langle [X, Y], Z \rangle - \langle [X, Z], Y \rangle - \langle [Y, Z], X \rangle \\ &= \langle [X, Y], Z \rangle + \langle \mathrm{ad}_Z X, Y \rangle + \langle \mathrm{ad}_Z Y, X \rangle = \langle [X, Y], Z \rangle. \end{aligned}$$

This implies the result for all left-invariant vector fields, since $Z \in \mathfrak{X}^L(G)$ can be chosen arbitrarily. The same argument works for right-invariant vector fields. □

Proof of Theorem 39.1. Proposition 39.11 implies $\nabla_X X = 0$ whenever $X \in \mathfrak{X}(G)$ is either left- or right-invariant, thus all of its flow lines are geodesics. Conversely, every geodesic γ satisfies $\dot{\gamma}(0) = X^L(\gamma(0)) = Y^R(\gamma(0))$ for unique choices of invariant vector fields $X^L \in \mathfrak{X}^L(G)$ and $Y^R \in \mathfrak{X}^R(G)$, and γ therefore matches the geodesics that are defined as flow lines of these vector fields through $\gamma(0)$. $\square$

Exercise 39.12. Suppose G is a Lie group with a bi-invariant Riemannian metric $\langle\ ,\ \rangle$, and ∇ is its Levi-Cività connection.

(a) Deduce from Proposition 39.11 that for any left-invariant vector fields $X, Y, Z \in \mathfrak{X}^L(G)$, the Riemann tensor of ∇ satisfies

$$R(X, Y)Z = \frac{1}{4}[Z, [X, Y]].$$

(b) Prove that the sectional curvature of $(G, \langle\ ,\ \rangle)$ is everywhere nonnegative: more precisely, if $X, Y \in \mathfrak{g}$ are orthonormal and $P \subset T_g G$ is spanned by the corresponding left-invariant vector fields $X^L(g), Y^L(g)$ at some point $g \in G$, then

$$K_S(P) = \frac{1}{4} |[X, Y]|^2 .$$

Exercise 39.13. Suppose G is a connected Lie group equipped with a left-invariant pseudo-Riemannian metric $\langle\ ,\ \rangle$ for which every Lie group homomorphism $\mathbb{R} \to G$ is also a geodesic. Prove that $\mathrm{ad}_X : \mathfrak{g} \to \mathfrak{g}$ is antisymmetric with respect to $\langle\ ,\ \rangle$ for every $X \in \mathfrak{g}$. It then follows from Exercise 39.10 that $\langle\ ,\ \rangle$ must in fact by bi-invariant. Having proved this, derive the same conclusion again starting from the assumption that $\langle\ ,\ \rangle$ is right- (but not necessarily left-) invariant.
Hint: Start by using the Koszul formula to deduce something about $\langle \mathrm{ad}_Z X, X\rangle$ for every $X, Z \in \mathfrak{g}$.

Exercise 39.14. Show that for a left- or right-invariant pseudo-Riemannian metric $\langle\ ,\ \rangle$ on a Lie group G, the inversion map $i : G \to G : g \mapsto g^{-1}$ is an isometry if and only if $\langle\ ,\ \rangle$ is bi-invariant.

39.2. The examples SO(3) and SU(2). The remainder of this lecture will focus largely on topological properties of Lie groups, but we begin the discussion with something very concrete. The groups SO(3) and SU(2) have a number of special properties that make them simultaneously a good illustration of the general theory and interesting objects of study in their own right. Their close relationship to each other is important to understand if you want to grasp the big picture on the interplay between Lie groups and their Lie algebras: in short, their Lie algebras are isomorphic, which tells you that SO(3) and SU(2) should appear isomorphic in a neighborhood of the identity, but despite both being compact and connected, they are not isomorphic groups, nor are they homeomorphic manifolds. The would-be diffeomorphism that one attempts to construct starting from a Lie algebra isomorphism $\mathfrak{su}(2) \xrightarrow{\cong} \mathfrak{so}(3)$ turns out instead to be a two-sheeted covering map $\mathrm{SU}(2) \to \mathrm{SO}(3)$, thus presenting SU(2) as the universal cover of SO(3). This will ensure the relevance of SU(2) when we later talk about spin structures and Dirac operators on oriented Riemannian manifolds. We will also see in §39.3 that the existence of the cover $\mathrm{SU}(2) \to \mathrm{SO}(3)$ is a special case of a general theorem.

As a warmup, let us take a look at the 1-dimensional, compact, connected and *abelian* Lie group SO(2):

Exercise 39.15. Show that the matrix $\mathbf{J}_0 := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ spans $\mathfrak{so}(2) \subset \mathbb{R}^{2\times 2}$, and every $\mathbf{A} \in \mathrm{SO}(2)$ can be written as $e^{\theta \mathbf{J}_0}$ for some (non-unique) $\theta \in \mathbb{R}$. (Give a direct proof of this, without appealing to Corollary 39.4.)

One-dimensional Lie algebras like $\mathfrak{so}(2)$ are not very interesting, since they are always abelian. The Lie algebra $\mathfrak{so}(3)$ is somewhat more exciting: it consists of all real antisymmetric 3-by-3 matrices, and is thus three-dimensional, with a natural basis given by the matrices

$$\mathbf{J}_1 := \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad \mathbf{J}_2 := \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \qquad \text{and} \qquad \mathbf{J}_3 := \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{39.1}$$

I call these basis vectors "natural" because they have straightforward geometric interpretations: writing $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3 \in \mathbb{R}^3$ for the standard basis of $\mathbb{R}^3$, the linear transformation $\mathbf{J}_1 : \mathbb{R}^3 \to \mathbb{R}^3$ annihilates the span of $\mathbf{e}_1$ while rotating its orthogonal complement 90 degrees counterclockwise with respect to the ordered basis $(\mathbf{e}_2, \mathbf{e}_3)$, and $\mathbf{J}_2$ and $\mathbf{J}_3$ admit similar characterizations after making cyclic perturbations of the numbers $1, 2, 3$. The commutators of these matrices are easily computed: we have

$$[\mathbf{J}_1, \mathbf{J}_2] = \mathbf{J}_3, \qquad [\mathbf{J}_2, \mathbf{J}_3] = \mathbf{J}_1, \qquad \text{and} \qquad [\mathbf{J}_3, \mathbf{J}_1] = \mathbf{J}_2,$$

and the Lie bracket on $\mathfrak{so}(3)$ is fully determined by these relations due to bilinearity and antisymmetry. Now, even if you have never studied Lie groups and Lie algebras before this course, I guarantee that you have seen this particular Lie algebra before, namely in classical 3-dimensional vector calculus. The **cross product** on $\mathbb{R}^3$ can be characterized as the unique antisymmetric bilinear map $\mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3 : (\mathbf{v}, \mathbf{w}) \mapsto \mathbf{v} \times \mathbf{w}$ such that

$$\mathbf{e}_1 \times \mathbf{e}_2 = \mathbf{e}_3, \qquad \mathbf{e}_2 \times \mathbf{e}_3 = \mathbf{e}_1, \qquad \text{and} \qquad \mathbf{e}_3 \times \mathbf{e}_1 = \mathbf{e}_2.$$

The next exercise clarifies in what sense the elements of $\mathrm{SO}(3)$ define "rotations" on $\mathbb{R}^3$, and in so doing, gives a new interpretation of the cross product on $\mathbb{R}^3$: $\mathbf{v} \times \mathbf{w}$ measures the degree to which rotations about $\mathbf{v}$ fail to commute with rotations about $\mathbf{w}$.

EXERCISE 39.16. Define $\Phi : \mathbb{R}^3 \to \mathfrak{so}(3)$ as the unique linear map with $\Phi(\mathbf{e}_i) = \mathbf{J}_i$ for $i = 1, 2, 3$.

(a) Prove that $(\mathbb{R}^3, \times)$ is a Lie algebra and $\Phi : \mathbb{R}^3 \to \mathfrak{so}(3)$ is a Lie algebra isomorphism.
(b) For $\mathbf{v} \in \mathbb{R}^3$, show that the transformation $\Phi(\mathbf{v}) : \mathbb{R}^3 \to \mathbb{R}^3$ is given by $\Phi(\mathbf{v})\mathbf{w} = \mathbf{v} \times \mathbf{w}$. *Hint: Verify that the formula holds when* $\mathbf{v}$ *and* $\mathbf{w}$ *are standard basis vectors.*
(c) Prove that for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^3$, the cross product and Euclidean inner product $\langle\ ,\ \rangle$ satisfy

$$\begin{aligned} \big(\mathbf{u} \times (\mathbf{v} \times \mathbf{w})\big) + \big(\mathbf{v} \times (\mathbf{w} \times \mathbf{u})\big) + \big(\mathbf{w} \times (\mathbf{u} \times \mathbf{v})\big) &= 0, \\ \langle \mathbf{u} \times \mathbf{v}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{u} \times \mathbf{w} \rangle &= 0 \end{aligned}$$

Hint: Don't try too hard! Just use the results of parts (a) and (b).
(d) Use your previous knowledge of the cross product to prove that for any $\mathbf{v} \in \mathbb{R}^3$ with $|\mathbf{v}| = 1$, the transformation $\Phi(\mathbf{v}) : \mathbb{R}^3 \to \mathbb{R}^3$ annihilates the subspace spanned by $\mathbf{v}$ and rotates its orthogonal complement by a right angle.
(e) For any $\mathbf{v} \in \mathbb{R}^3$ with $|\mathbf{v}| = 1$ and $\theta \in \mathbb{R}$, show that $e^{\theta\Phi(\mathbf{v})} \in \mathrm{SO}(3)$ is a rotation of angle θ about the subspace spanned by $\mathbf{v}$. Conclude that for $\mathbf{v} \in \mathbb{R}^3$, $e^{\Phi(\mathbf{v})} = \mathbb{1}$ if and only if $|\mathbf{v}| \in 2\pi\mathbb{Z}$.
Hint: An intelligent choice of basis reduces this to Exercise 39.15.
(f) Show that every $\mathbf{A} \in \mathrm{SO}(3)$ has 1 as an eigenvalue, and the dimension of the corresponding eigenspace is 1 unless $\mathbf{A} = \mathbb{1}$.
(g) Show that whenever $\mathbf{A} \in \mathrm{SO}(3)$ has a 1-dimensional eigenspace $\ell \subset \mathbb{R}^3$ with eigenvalue 1, $\mathbf{A}$ defines a rotation on the orthogonal complement of ℓ, and $\mathbf{A} = e^{\Phi(\mathbf{v})}$ for some $\mathbf{v} \in \ell$.

Now let's consider the Lie algebra of $\mathrm{SU}(2)$. It consists of all complex anti-Hermitian 2-by-2 matrices that are also traceless: note that in contrast to the real case, being traceless does not follow from the anti-Hermitian requirement, but is an extra condition. (This corresponds to the fact that $\mathrm{SO}(n)$ is a connected component of $\mathrm{O}(n)$ and thus has the same Lie algebra, while $\mathrm{SU}(n)$ is by contrast a codimension-one submanifold of $\mathrm{U}(n)$; see Exercise 4.25 from last semester.) It would be hard to claim that $\mathfrak{su}(2)$ has a "canonical" basis, but it does have one that is traditionally favored: it is written in terms of the so-called *Pauli matrices*

$$\boldsymbol{\sigma}_1 := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \boldsymbol{\sigma}_2 := \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad \boldsymbol{\sigma}_3 := \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

These have their origin in quantum mechanics, where they are interpreted as the Hermitian operators (on a complex 2-dimensional Hilbert space) representing the "spin" of a fermionic particle about the axes spanned by $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ respectively in $\mathbb{R}^3$. Since they are traceless and Hermitian, multiplying them by i makes them traceless and *anti*-Hermitian, thus forming a basis of $\mathfrak{su}(2)$. If you now compute their commutators, you may soon suspect that you should have first multiplied them all by $-\frac{1}{2}$, because doing so gives rise to the familiar relations

$$\left[-\frac{i}{2}\boldsymbol{\sigma}_1, -\frac{i}{2}\boldsymbol{\sigma}_2\right] = -\frac{i}{2}\boldsymbol{\sigma}_3, \qquad \left[-\frac{i}{2}\boldsymbol{\sigma}_2, -\frac{i}{2}\boldsymbol{\sigma}_3\right] = -\frac{i}{2}\boldsymbol{\sigma}_1, \qquad \text{and} \qquad \left[-\frac{i}{2}\boldsymbol{\sigma}_3, -\frac{i}{2}\boldsymbol{\sigma}_1\right] = -\frac{i}{2}\boldsymbol{\sigma}_2,$$

This proves:

PROPOSITION 39.17. *The unique linear map $\mathfrak{su}(2) \to \mathfrak{so}(3)$ that sends $-\frac{i}{2}\boldsymbol{\sigma}_k \mapsto \mathbf{J}_k$ for each $k = 1, 2, 3$ is a Lie algebra isomorphism.* □

With an isomorphism $\mathfrak{su}(2) \to \mathfrak{so}(3)$ in hand, we can now observe something rather funny. Suppose $\Psi : \mathrm{SO}(3) \to \mathrm{SU}(2)$ is a Lie group homomorphism whose derivative $\Psi_* : \mathfrak{so}(3) \to \mathfrak{su}(2)$ at the identity is the isomorphism in Proposition 39.17. In particular, $\Psi_* \mathbf{J}_3 = -\frac{i}{2}\boldsymbol{\sigma}_3 = \begin{pmatrix} -i/2 & 0 \\ 0 & i/2 \end{pmatrix}$, thus Ψ must give a relation between the corresponding 1-parameter subgroups, namely

$$\Psi(e^{t\mathbf{J}_3}) = e^{-it\boldsymbol{\sigma}_3/2} = \begin{pmatrix} e^{-it/2} & 0 \\ 0 & e^{it/2} \end{pmatrix}.$$

But this is absurd: the left hand side of this expression is a 2π-periodic function of t, and the right hand side is not—its minimal period is 4π. In particular, while $e^{t\mathbf{J}_3} \in \mathrm{SO}(3)$ traverses the family of rotations about $\mathbf{e}_3$ exactly once as t goes from 0 to 2π, $e^{-it\boldsymbol{\sigma}_3/2} \in \mathrm{SU}(2)$ traverses an embedded path from $\mathbb{1}$ to $-\mathbb{1}$, making Ψ at best a "double-valued" map. One concludes that there is no Lie group homomorphism $\mathrm{SO}(3) \to \mathrm{SU}(2)$ whose derivative is the isomorphism in Proposition 39.17. In fact, Exercise 39.21 below will show that there are no nontrivial Lie group homomorphisms $\mathrm{SO}(3) \to \mathrm{SU}(2)$ at all.

On the other hand, one gets better results by looking for a Lie group homomorphism $\Psi : \mathrm{SU}(2) \to \mathrm{SO}(3)$: if we assume $\Psi_* : \mathfrak{su}(2) \to \mathfrak{so}(3)$ is the isomorphism in Proposition 39.17, the discussion above implies

$$\Psi(e^{-it\boldsymbol{\sigma}_3/2}) = e^{t\mathbf{J}_3}.$$

This does not lead to a contradiction, but it is also clear that Ψ in this situation cannot be a homeomorphism, as composing the embedding $\mathbb{R}/4\pi\mathbb{Z} \to \mathrm{SU}(2) : t \mapsto e^{-it\boldsymbol{\sigma}_3}$ with Ψ produces a double covering of the aforementioned loop of rotations in $\mathrm{SO}(3)$. A hint of what's really going on here is provided by the following exercise.

EXERCISE 39.18. Prove that if G and H are connected Lie groups and $\Phi : G \to H$ is a Lie group homomorphism for which the induced Lie algebra homomorphism $\Phi_* : \mathfrak{g} \to \mathfrak{h}$ is an isomorphism, then Φ is a covering map.
Hint: For inspiration, see the proof of Lemma 36.19.

As it turns out, there is an easy trick for producing concrete examples of covering maps $\mathrm{SU}(2) \to \mathrm{SO}(3)$. It is essentially the adjoint representation:

EXERCISE 39.19. Recall that the **center** of a group G is the subgroup consisting of all elements that commute with everything in G.

(a) Show that the center of $\mathrm{SU}(2)$ contains only the two elements $\mathbb{1}$ and $-\mathbb{1}$.
(b) Show that $\mathrm{SU}(2)$ is diffeomorphic to S^3.
Hint: What does the image of the map $\mathrm{SU}(2) \to \mathbb{C}^2 : \mathbf{A} \mapsto \mathbf{A}\mathbf{v}$ *look like for a fixed unit vector* $\mathbf{v} \in \mathbb{C}^2$*?*

Now fix an Ad-invariant (positive) inner product $\langle\ ,\ \rangle$ on $\mathfrak{su}(2)$; its existence is guaranteed by Corollary 38.14 since $\mathrm{SU}(2)$ is compact, and Exercise 38.5 gives an explicit example. We can then denote by $\mathrm{O}(\mathfrak{su}(2)) \subset \mathrm{GL}(\mathfrak{su}(2))$ the group of linear transformations on $\mathfrak{su}(2)$ that preserve $\langle\ ,\ \rangle$, and define $\mathrm{SO}(\mathfrak{su}(2)) \subset \mathrm{O}(\mathfrak{su}(2))$ as the identity component. The dimension of $\mathfrak{su}(2)$ is three, thus any choice of orthonormal basis for $\mathfrak{su}(2)$ determines a Lie group isomorphism of $\mathrm{SO}(\mathfrak{su}(2))$ with $\mathrm{SO}(3)$. Since $\langle\ ,\ \rangle$ is Ad-invariant and $\mathrm{SU}(2)$ is connected, the adjoint representation $\mathrm{Ad} : \mathrm{SU}(2) \to \mathrm{GL}(\mathfrak{su}(2))$ now takes values in $\mathrm{SO}(\mathfrak{su}(2))$ and thus defines a Lie group homomorphism

$$\mathrm{Ad} : \mathrm{SU}(2) \to \mathrm{SO}(\mathfrak{su}(2)) \cong \mathrm{SO}(3).$$

(c) Show that the Lie algebra homomorphism $\mathrm{ad} : \mathfrak{su}(2) \to \mathfrak{so}(\mathfrak{su}(2))$ is an isomorphism.
(d) It follows now from Exercise 39.18 that $\mathrm{Ad} : \mathrm{SU}(2) \to \mathrm{SO}(\mathfrak{su}(2))$ is a covering map. Show that its degree is 2, i.e. every element of $\mathrm{SO}(\mathfrak{su}(2))$ has exactly two elements of $\mathrm{SU}(2)$ in its preimage. Show moreover that for any nontrivial $\mathbf{A} \in \mathfrak{su}(2)$, there exists a unique $\tau > 0$ such that the path $[0,\tau] \to \mathrm{SU}(2) : t \mapsto e^{t\mathbf{A}}$ defines an embedded path from $\mathbb{1}$ to $-\mathbb{1}$, while $[0,\tau] \to \mathrm{SO}(\mathfrak{su}(2)) : t \mapsto \mathrm{Ad}_{e^{\tau \mathbf{A}}}$ is a closed loop containing all the rotations about a fixed 1-dimensional subspace in $\mathfrak{su}(2) \cong \mathbb{R}^3$.
(e) Find a diffeomorphism between $\mathrm{SO}(3)$ and $\mathbb{RP}^3$.

REMARK 39.20. There are also more geometrically intuitive ways to see the diffeomorphism $\mathrm{SO}(3) \cong \mathbb{RP}^3$ than via the covering map constructed in Exercise 39.19. See for example the explanation at https://en.wikipedia.org/wiki/3D_rotation_group#Topology.

Let's round out this discussion with some observations about the special properties of $\mathrm{SO}(3)$, some of which we will see have important implications for the geometry of Riemannian 3-manifolds.

EXERCISE 39.21. Recall from Exercise 39.16 the Lie algebra isomorphism $\Phi : (\mathbb{R}^3, \times) \to \mathfrak{so}(3)$ given by $\Phi(\mathbf{v})\mathbf{w} = \mathbf{v} \times \mathbf{w}$. Prove:

(a) Every nontrivial proper Lie subalgebra of $\mathfrak{so}(3)$ is 1-dimensional.
(b) Every nontrivial Lie algebra homomorphism $\mathfrak{so}(3) \to \mathfrak{so}(3)$ is an isomorphism.
Hint: What could the dimensions of its kernel and image be?
(c) The transformation $\mathbf{A} : \mathbb{R}^3 \to \mathbb{R}^3$ is a Lie algebra isomorphism with respect to $\times$ whenever $\mathbf{A} \in \mathrm{SO}(3)$.
(d) Every Lie algebra isomorphism of $(\mathbb{R}^3, \times)$ to itself comes from an element of $\mathrm{SO}(3)$ as described in part (c).
Hint: Show first that if $\psi : \mathbb{R}^3 \to \mathbb{R}^3$ *is a Lie algebra isomorphism and* $\mathbf{v}, \mathbf{w} \in \mathbb{R}^3$ *are orthogonal, then* $\psi(\mathbf{v})$ *and* $\psi(\mathbf{w})$ *are also orthogonal. You can deduce from this via linear algebra (cf. the proof of Lemma 24.5 from last semester) that* ψ *is a positive multiple of an orthogonal transformation.*
(e) For any $\mathbf{A} \in \mathrm{SO}(3)$, $\Phi\mathbf{A}\Phi^{-1} = \mathrm{Ad}_{\mathbf{A}} : \mathfrak{so}(3) \to \mathfrak{so}(3)$, i.e. the isomorphism $\Phi : \mathbb{R}^3 \to \mathfrak{so}(3)$ identifies the natural action of $\mathrm{SO}(3)$ on $\mathbb{R}^3$ with the adjoint representation of $\mathrm{SO}(3)$.
(f) Every connected nontrivial proper Lie subgroup of $\mathrm{SO}(3)$ is of the form $\{e^{t\mathbf{A}} \in \mathrm{SO}(3) \mid t \in \mathbb{R}\}$ for some $\mathbf{A} \in \mathfrak{so}(3)$, and is thus isomorphic to $\mathrm{SO}(2) \cong S^1$.
(g) If $\Psi : \mathrm{SO}(3) \to \mathrm{SO}(3)$ is a nontrivial Lie group homomorphism, then Ψ is in fact a Lie group isomorphism of the form $\Psi(\mathbf{B}) = \mathbf{A}\mathbf{B}\mathbf{A}^{-1}$ for some $\mathbf{A} \in \mathrm{SO}(3)$.

Hint: Since $\exp : \mathfrak{so}(3) \to \mathrm{SO}(3)$ *is surjective,* Ψ *is uniquely determined by the relation* $\Psi(e^{t\mathbf{B}}) = e^{t\Psi_*\mathbf{B}}$ *for* $\mathbf{B} \in \mathfrak{so}(3)$.

(h) There exist nontrivial Lie group homomorphisms $\mathrm{SO}(2) \to \mathrm{SO}(2)$ that are not isomorphisms.

(i) There exist no nontrivial Lie group homomorphisms $\mathrm{SO}(3) \to \mathrm{SU}(2)$.
Hint: This is easy if you know enough about covering space theory, but you can also do without that. Deduce from the results above that if such a homomorphism exists, then the induced Lie algebra homomorphism $\mathfrak{so}(3) \to \mathfrak{su}(2)$ *cannot differ very much from the isomorphism in Proposition 39.17. Then use the exponential map.*

Exercise 39.22. Suppose G is a Lie group with an Ad-invariant positive inner product $\langle\ ,\ \rangle$ on its Lie algebra $\mathfrak{g}$.

(a) Show that for any $X, Y \in \mathfrak{g}$, $[X, Y]$ is orthogonal to both X and Y.
Hint: Use the formula $\mathrm{ad}_X Y = [X, Y]$ *and the fact that* $\mathrm{ad}_X : \mathfrak{g} \to \mathfrak{g}$ *is antisymmetric.*

(b) Show that if $\mathfrak{g}$ is not abelian and $\dim \mathfrak{g} = 3$, then there exists a constant $\lambda > 0$ and an orthonormal basis $e_1, e_2, e_3 \in \mathfrak{g}$ satisfying

$$[e_1, e_2] = \lambda e_3, \qquad [e_2, e_3] = \lambda e_1 \quad \text{and} \quad [e_3, e_1] = \lambda e_2.$$

(c) Under the assumptions of part (b), deduce that the Lie algebra $\mathfrak{g}$ is isomorphic to $\mathfrak{so}(3)$.

(d) Prove that for every compact connected non-abelian Lie group G with $\dim G = 3$, there exists a Lie group homomorphism $G \to \mathrm{SO}(3)$ that is a covering map.
Remark: If you know enough covering space theory, you may realize that this implies G *is isomorphic to either* $\mathrm{SO}(3)$ *or* $\mathrm{SU}(2)$*. The crucial observation here is that* $\mathrm{SO}(3) \cong \mathbb{RP}^3$ *has fundamental group* $\mathbb{Z}_2$ *and* $\mathrm{SU}(2) \cong S^3$ *is simply connected, thus the covering map* $\mathrm{SU}(2) \to \mathrm{SO}(3)$ *in Exercise 39.19 is the universal cover. It follows that if the cover* $G \to \mathrm{SO}(3)$ *is not an isomorphism, the scenario in Exercise 39.19 is the only other possibility.*

(e) Find an example of a connected non-abelian (but not compact!) 3-dimensional Lie group whose Lie algebra is not isomorphic to $\mathfrak{so}(3)$.

39.3. Simply connected groups. From a topological perspective, there is a simple reason why our attempt in the previous subsection to lift the Lie algebra isomorphism $\mathfrak{so}(3) \cong \mathfrak{su}(2)$ to a Lie group homomorphism $\mathrm{SO}(3) \to \mathrm{SU}(2)$ was doomed to failure: by Exercise 39.18, such a homomorphism would necessarily be a covering map, and you cannot cover a simply connected manifold like $\mathrm{SU}(2) \cong S^3$ with one like $\mathrm{SO}(3) \cong \mathbb{RP}^3$ that is not simply connected. The next result clarifies, on the other hand, why our attempt to lift the isomorphism $\mathfrak{su}(2) \to \mathfrak{so}(3)$ to a homomorphism $\mathrm{SU}(2) \to \mathrm{SO}(3)$ was guaranteed to succeed:

Theorem 39.23. *Suppose* G *and* H *are Lie groups and* $\phi : \mathfrak{g} \to \mathfrak{h}$ *is a Lie algebra homomorphism. If* G *is simply connected, then there exists a unique Lie group homomorphism* $\Phi : G \to H$ *such that* $\Phi_* = \phi$.

Notice that if ϕ is an isomorphism and G and H are *both* simply connected, the theorem can be applied in both directions, giving:

Corollary 39.24. *If* G *and* H *in Theorem 39.23 are both simply connected and* $\phi : \mathfrak{g} \to \mathfrak{h}$ *is a Lie algebra isomorphism, then* $\Phi : G \to H$ *is a Lie group isomorphism. In particular, two simply connected Lie groups are isomorphic if and only if their Lie algebras are isomorphic.* □

Sketch of the proof of Theorem 39.23. The uniqueness of $\Phi : G \to H$ is quite easy: if $\Phi_* = \phi$, then Φ must satisfy $\Phi(\exp(X)) = \exp(\phi(X))$ for every $X \in \mathfrak{g}$, and Φ is thus uniquely determined on the image of $\exp : \mathfrak{g} \to G$, which is an open neighborhood of $e \in G$. By Exercise 37.23, that neighborhood generates G, so it follows that Φ is uniquely determined.

The proof of existence is more involved, so we will give only enough of a sketch to elucidate why the various hypotheses are necessary.

Step 1: We claim that defining Φ on a neighborhood of $e \in G$ by $\Phi(\exp(X)) = \exp(\phi(X))$ gives a so-called *local homomorphism*, meaning that for a sufficiently small neighborhood $\mathcal{U} \subset G$ of e, $a, b, ab \in \mathcal{U}$ implies $\Phi(ab) = \Phi(a)\Phi(b)$. This is where the hypothesis of $\phi : \mathfrak{g} \to \mathfrak{h}$ being a Lie algebra homomorphism is essential. Notice first that if $a, b \in \mathcal{U}$ happen to be of the form $a = \exp(X)$ and $b = \exp(Y)$ with $[X, Y] = 0$, then Exercise 37.27 implies $ab = \exp(X + Y)$, and thus

$$\Phi(ab) = \Phi(\exp(X+Y)) = \exp(\phi(X+Y)) = \exp(\phi(X)+\phi(Y)) = \exp(\phi(X))\exp(\phi(Y)) = \Phi(a)\Phi(b),$$

where in the last step we again made use of Exercise 37.27 and the fact that $[\phi(X), \phi(Y)] = \phi([X, Y]) = 0$. Things are more complicated if $[X, Y] \neq 0$, as one then needs some information about the error term Z in the relation $\exp(X)\exp(Y) = \exp(X + Y + Z)$; here Z can be assumed to depend smoothly on X and Y when they are close enough to 0, and Z will also vanish when $X = Y = 0$. The standard approach now is to apply the so-called Baker-Campbell-Hausdorff formula, which presents Z as a convergent series whose terms are all iterated brackets of X and Y:

$$Z = \frac{1}{2}[X, Y] + \frac{1}{12}[X, [X, Y]] - \frac{1}{12}[Y, [X, Y]] + \dots.$$

The point is that since every term in this formula involves brackets and the bracket is preserved by ϕ, $\phi(Z)$ now satisfies an identical formula as a function of $\phi(X)$ and $\phi(Y)$, and one thus obtains

$$\Phi(ab) = \Phi(\exp(X + Y + Z)) = \exp(\phi(X) + \phi(Y) + \phi(Z)) = \exp(\phi(X))\exp(\phi(Y)) = \Phi(a)\Phi(b).$$

For details on the Baker-Campbell-Hausdorff formula, see e.g. [**Hal15**, Chapter 5] or [**DK00**, §1.7].

Step 2: We next write down a formula for $\Phi(g)$ for arbitrary $g \in G$, though the formula will appear at first to depend on some choices. In particular, choose a continuous path $\gamma : [0, 1] \to G$ from $\gamma(0) = e$ to $\gamma(1) = g$, along with a partition $0 = t_0 < t_1 < \dots < t_{N-1} < t_N = 1$, writing $g_j := \gamma(t_j)$ for $j = 0, \dots, N$ and defining $h_j \in G$ for $j = 1, \dots, N$ by the relation $g_j := h_j g_{j-1}$, so that $g = h_N h_{N-1} \dots h_1$. Assume the partition is chosen to be fine enough so that all of the elements h_j and $h_j h_{j-1}$ belong to the neighborhood $\mathcal{U} \subset G$ of e on which Φ is already defined and known to be a local homomorphism. The correct procedure is then obviously to define

$$\Phi(g) := \Phi(h_N)\Phi(h_{n-1}) \dots \Phi(h_1) \in H, \tag{39.2}$$

and we claim that this definition of $\Phi(g)$ will not change under any sufficiently small perturbation of the path γ from e to g and the partition points $t_1, \dots, t_{N-1}$. Since our definition depends only on the points $g_1, \dots, g_N$ rather than the path γ itself, we can prove this by examining what happens if these points are altered one at a time, e.g. suppose a particular point g_j is replaced by a nearby point g_j', and define $h_j', h_{j+1}' \in G$ by

$$g_j' = h_j' g_{j-1}, \qquad g_{j+1} = h_{j+1}' g_j'.$$

Then $g_{j+1} = h_{j+1}' h_j' g_{j-1} = h_{j+1} h_j g_{j-1}$, implying $h_{j+1}' h_j' = h_{j+1} h_j$, and if g_j' is close enough to g_j, we can assume as a consequence of the local homomorphism condition that $\Phi(h_{j+1}')\Phi(h_j') = \Phi(h_{j+1})\Phi(h_j)$. As a consequence, the total product in (39.2) does not change when h_j and h_{j+1} are replaced by h_j' and h_{j+1}'. Note finally that for similar reasons, the product (39.2) will not change if the partition $t_1, \dots, t_{N-1}$ is made finer via the addition of a new point dividing one of the intervals (t_{j-1}, t_j), so in this way the number of points in the partition can always be increased by 1.

Step 3: Now it is time to make use of simple connectedness. From step 2, we have for each $g \in G$ a definition of $\Phi(g)$ that depends on a choice of continuous path γ from e to g and a fine partition of this path, and the definition is invariant under small changes in this data. But since G is simply connected, any two such paths can be connected by a continuous 1-parameter family of paths from e to g, and after choosing suitable partitions for the paths in this family, the resulting family of definitions of $\Phi(g)$ will be independent of the parameter. This proves that the definition of $\Phi(g)$ resulting from steps 1 and 2 is independent of all choices, thus defining a global map $\Phi : G \to H$. It is now an easy exercise to check that this map is smooth and is a group homomorphism. $\square$

Before moving on from this subject, two other results are worth mentioning. First, simply connected Lie groups are easy to find, in fact:

THEOREM 39.25. *For every connected Lie group G, there exists a simply connected Lie group $\widetilde{G}$ with a Lie group homomorphism $\pi : \widetilde{G} \to G$ that is a covering map.*

The group $\widetilde{G}$ in this theorem is called the **universal cover** of G, and the existence of the covering map $\pi : \widetilde{G} \to G$ implies that its Lie algebra is isomorphic to that of G, thus it follows from Corollary 39.24 above that $\widetilde{G}$ is unique up to a Lie group isomorphism. We saw a nontrivial example in the previous subsection: SU(2) is the universal cover of SO(3). Outside of concrete examples like this one, the group $\widetilde{G}$ can be constructed from an arbitrary connected Lie group G via a general procedure that is standard in topology: one defines $\widetilde{G}$ as the set of equivalence classes of continuous paths

$$\widetilde{G} := \{\gamma : [0,1] \to G \mid \gamma(0) = e\} \big/ \sim,$$

where two paths are considered equivalent if and only if they are homotopic with fixed end points, and the covering map $\pi : \widetilde{G} \to G$ is then given by

$$\pi([\gamma]) := \gamma(1).$$

The obvious group structure to define on $\widetilde{G}$ is

$$[\alpha][\beta] := [\alpha\beta],$$

where $\alpha\beta : [0,1] \to G$ denotes the path $t \mapsto \alpha(t)\beta(t)$, and for this choice, $\pi : \widetilde{G} \to G$ is manifestly a group homomorphism. It is not difficult to check that $\widetilde{G}$ can also be endowed with a smooth structure that makes it a Lie group and the map $\pi : \widetilde{G} \to G$ smooth.

Finally, we complete the correspondence between Lie algebras and simply connected Lie groups: the following result is known as *Lie's third fundamental theorem.*

THEOREM 39.26. *Every finite-dimensional Lie algebra is isomorphic to the Lie algebra of a Lie group (and therefore also of a simply connected Lie group).*

We will not have any further use for Theorem 39.26 in this course, but it is nice to know about since it provides (in conjunction with Theorems 39.23 and 39.25) a natural *bijective* correspondence between finite-dimensional Lie algebras and simply connected Lie groups.

40. Quotient manifolds

Many important examples of smooth manifolds arise naturally as quotients of other manifolds by smooth group actions. Our main goal in this lecture is to establish the basic notions underlying such quotients and determine when they are actually smooth.

40.1. Smooth group actions. A (left) **action** of a group G on a set M is a map

$$G \times M \to M : (g, p) \mapsto gp$$

that satisfies

$$ep = p \text{ for all } p \in M \qquad \text{and} \qquad (gh)p = g(hp) \text{ for all } g, h \in G \text{ and } p \in M.$$

We also say in this case that G **acts from the left** on M, and the notation

$$g \cdot p := gp$$

is sometimes also used. Given a subset $\mathcal{U} \subset M$ and/or elements $g \in G$, $p \in M$, we denote

$$\begin{aligned} g\mathcal{U} &:= \{gp \in M \mid p \in \mathcal{U}\} \subset M, \\ Gp &:= \{gp \in M \mid g \in G\} \subset M, \\ G\mathcal{U} &:= \bigcup_{g \in G} g\mathcal{U} = \bigcup_{p \in \mathcal{U}} Gp \subset M, \end{aligned}$$

and call Gp or $G\mathcal{U}$ the **orbit** of p or $\mathcal{U}$ respectively under the G-action. The set $\mathcal{U}$ is called G**-invariant** if $G\mathcal{U} = \mathcal{U}$, and the point p is **fixed** under the G-action if $Gp = \{p\}$. The set

$$G_p := \{g \in G \mid gp = p\} \subset G$$

is known as the **stabilizer** or **isotropy subgroup** of the point p; note that it is always a subgroup of G, and it equals G if and only if p is fixed under the action. Any two orbits of a group action are either identical or disjoint, and the **quotient** of M by G is defined as the set of all orbits, and denoted by

$$M/G := \{Gp \subset M \mid p \in M\}.$$

To put this another way, there is an equivalence relation defined on M such that $p \sim q$ if and only if p and q belong to the same orbit, and M/G is then the quotient by this equivalence relation. From this perspective, it is common to write elements of the quotient as

$$[p] := Gp \in M/G$$

for each $p \in M$.

Given an action of G on two sets M and N, a map $f : M \to N$ is called G**-equivariant** if

$$f(gp) = gf(p) \qquad \text{for all } g \in G \text{ and } p \in M.$$

Similarly, a function f on M is called G**-invariant** if it satisfies $f(gp) = f(p)$ for all $g \in G$ and $p \in M$.

If G is a topological group and M a topological space, one calls a group action **continuous** if the map $G \times M \to M$ is continuous; similarly, for G a Lie group and M a smooth manifold, the action is said to be **smooth** if the map $G \times M \to M$ is smooth. We say in these cases that G **acts continuously/smoothly** (from the left) on M. In particular, a smooth left action of G on M determines a group homomorphism

$$G \to \mathrm{Diff}(M) : g \mapsto \varphi_g, \qquad \varphi_g(p) := gp.$$

Whenever M has a topology, it is natural to equip the quotient M/G with the **quotient topology**, defined as the largest topology for which the **quotient projection**

$$\pi : M \to M/G : p \mapsto [p]$$

is a continuous map. Concretely, this means that a set $\mathcal{U} \subset M/G$ is open if and only if its orbit $\pi^{-1}(\mathcal{U}) = G\mathcal{U}$ is an open subset of M. It should be emphasized that even if M is a smooth manifold and G acts smoothly, M/G may have horrible topological properties, e.g. it can easily fail to be

Hausdorff (see Example 40.1 below). One of the main themes of this lecture will be to ascertain what conditions on a smooth group action allow us to avoid such undesirable quotients.

The notion of a **right group action** is defined similarly, as a map $M \times G \to M : (p, g) \mapsto pg$ that satisfies the conditions

$$pe = p \text{ for all } p \in M \qquad \text{and} \qquad p(gh) = (pg)h \text{ for all } g, h \in G \text{ and } p \in M.$$

All of the definitions above admit obvious modifications to accommodate right instead of left group actions.[90] While the difference is mostly a matter of bookkeeping, left and right group actions are not completely equivalent notions—the principal difference is that if one uses a right action to define diffeomorphisms $\varphi_g(p) := pg$, then the resulting map $G \to \mathrm{Diff}(M) : g \mapsto \varphi_g$ is not a group homomorphism, but rather an *antihomomorphism*, i.e. it satisfies $\varphi_{gh} = \varphi_h \circ \varphi_g$. From this perspective, the distinction between left and right actions is only truly meaningful when G is nonabelian, and even in the nonabelian case, any left action $G \times M \to M$ can be converted into a right action $M \times G \to M$ by defining $pg := g^{-1}p$, or vice versa. For this reason, all important results about left actions admit minor modifications to produce corresponding results about right actions, so that we will lose no generality by considering only left actions in most of our exposition.

Example 40.1. Any representation $\rho : G \to \mathrm{GL}(V)$ of a Lie group G on a finite-dimensional vector space V defines a smooth left action of G on V by writing $gv := \rho(g)v$ for $g \in G$, $v \in V$. Group actions of this form are sometimes also called **linear** group actions. The quotient of a vector space V by a linear group action is typically not a very nice object, e.g. if the homomorphism $\rho : G \to \mathrm{GL}(V)$ is surjective or contains a large enough subgroup such as $\mathrm{SL}(V)$ (assuming $\dim V \geqslant 2$), then V/G contains only two points, namely $[0]$ and $[v]$ for any $v \neq 0 \in V$, and the quotient topology on V/G has the disturbing property that V/G itself is the only open neighborhood of the point $[0]$, implying that every sequence converges to $[0]$. Many of them also simultaneously converge to $[v]$, thus V/G is not Hausdorff.

Example 40.2. Variations on the following example of a right group action will arise naturally when we discuss principal fiber bundles. If V is an n-dimensional real vector space and M is the set of all vector space isomorphisms $\phi : \mathbb{R}^n \to V$, then the group $\mathrm{GL}(n, \mathbb{R})$ of invertible linear transformations $\mathbb{R}^n \to \mathbb{R}^n$ acts on M from the right by $\phi \cdot A := \phi \circ A$. For this action, every point of M belongs to the same orbit, thus the quotient M/G is a one-point space.

Example 40.3. For every Lie group G, the map $G \times G \to G : (g, h) \mapsto gh$ can be interpreted as a smooth left action or right action of G on itself, depending on whether you choose to view the first or second copy of G as the manifold that is being acted upon; in either case, every point belongs to the same orbit, so G/G is a one-point space. More generally, every Lie subgroup $H \subset G$ naturally has both a smooth left action and a smooth right action on G, defined by

$$H \times G \to G : (h, g) \mapsto hg, \qquad \text{and} \qquad G \times H \to G : (g, h) \mapsto gh$$

respectively. For the left action, G/H is the set of **right cosets** $\{Hg \subset G \mid g \in G\}$, whereas the right action produces the set of **left cosets** $\{gH \subset G \mid g \in G\}$. These two versions of the quotient G/H have a canonical identification with each other whenever H is a *normal* subgroup, in which case G/H also inherits a natural group structure.

Example 40.4. A smooth action of the group $\mathbb{Z}_2 := \mathbb{Z}/2\mathbb{Z}$ on a manifold M is equivalent to a so-called **involution**, i.e. a diffeomorphism $\varphi : M \to M$ satisfying $\varphi \circ \varphi = \mathrm{Id}$, which can be

[90] In situations where M has both a left and a right action by G, it is occasionally convenient to distinguish between the two quotients by writing M/G for the quotient by a right action but $G\backslash M$ for the quotient by a left action. I find the latter notation confusing and will not use it for quotients, but will instead write M/G for both left and right actions.

defined in this case by $\varphi(p) := [1] \cdot p$ using the unique nontrivial element $[1] \in \mathbb{Z}_2$. A familiar example is the map $S^n \to S^n : p \mapsto -p$, representing a smooth action of $\mathbb{Z}_2$ on S^n whose quotient is the manifold $\mathbb{RP}^n$. Alternatively, $\mathbb{RP}^n$ can be defined as $(\mathbb{R}^{n+1}\backslash\{0\})/\mathbb{R}^*$ where the multiplicative abelian group $\mathbb{R}^* := \mathbb{R}\backslash\{0\}$ acts on $\mathbb{R}^{n+1}\backslash\{0\}$ via scalar multiplication.

Example 40.5. The complex analogue of Example 40.4 gives two equivalent presentations of the complex projective space (cf. Exercise 32.11) as quotients:

$$\mathbb{CP}^n = (\mathbb{C}^{n+1}\backslash\{0\})/\mathbb{C}^* = S^{2n-1}/S^1,$$

where again $\mathbb{C}^* := \mathbb{C}\backslash\{0\}$ is a multiplicative abelian group, S^{2n-1} and S^1 are regarded as the unit spheres in $\mathbb{C}^{n+1}$ and $\mathbb{C}$ respectively, and the second presentation is obtained from the first one by observing that any point of $\mathbb{C}^{n+1}\backslash\{0\}$ is in the same $\mathbb{C}^*$-orbit with a unit vector, while two unit vectors belong to the same $\mathbb{C}^*$-orbit if and only if they belong to the same S^1-orbit.

40.2. Fundamental vector fields. A smooth left group action $G \times M \to M$ associates to each point $p \in M$ a smooth map $\psi_p : G \to M$ defined by $\psi_p(g) := gp$. The **fundamental vector field** $X^F \in \mathfrak{X}(M)$ corresponding to a given element $X \in \mathfrak{g}$ is then defined by

$$X^F(p) := T_e\psi_p(X) = \frac{d}{dt}\exp(tX)p\bigg|_{t=0} \in T_pM.$$

For right actions $M \times G \to M$, one can analogously define $X^F(p) := \partial_t p\exp(tX)|_{t=0}$. Denote the space of all fundamental vector fields for a given group action on M by

$$\mathfrak{X}^F(M) \subset \mathfrak{X}(M).$$

It is clearly a finite-dimensional vector space since it comes with a surjective linear map $\mathfrak{g} \to \mathfrak{X}^F(M) : X \mapsto X^F$, and we will see in a moment that it is also a Lie subalgebra of $\mathfrak{X}(M)$.

Example 40.6. If $M = G$ and we consider a Lie subgroup $H \subset G$ acting on G from the left as in Example 40.3, then each $X \in \mathfrak{h} \subset \mathfrak{g}$ produces a fundamental vector field $X^F \in \mathfrak{X}(G)$ given by $X^F(g) = \partial_t \exp(tX)g|_{t=0} = \partial_t R_g(\exp(tX))|_{t=0} = TR_g(X) = X^R(g)$, so $\mathfrak{X}^F(G)$ is the Lie subalgebra of $\mathfrak{X}^R(G)$ corresponding to $\mathfrak{h} \subset \mathfrak{g}$ under the canonical bijection $\mathfrak{g} \to \mathfrak{X}^R(G) : X \mapsto X^R$. Similarly, if we regard H as acting on G from the right, then the fundamental vector fields are left-invariant.

Proposition 40.7. *For any smooth left group action $G \times M \to M$, the action on M of the 1-parameter subgroup generated by an element $X \in \mathfrak{g}$ is determined by the flow of the associated fundamental vector field $X^F \in \mathfrak{X}(M)$, namely by*

$$\exp(tX)p = \varphi^t_{X^F}(p) \qquad \text{for } t \in \mathbb{R},\ p \in M.$$

Proof. Clearly the formula is valid for $t = 0$, so its validity in general follows by computing the derivative of the path $t \mapsto \exp(tX)p \in M$ for arbitrary t:

$$\begin{aligned}\frac{d}{dt}(\exp(tX)p) &= \frac{d}{ds}\exp((s+t)X)p\bigg|_{s=0} = \frac{d}{ds}(\exp(sX)\exp(tX))\,p\bigg|_{s=0} \\ &= \frac{d}{ds}\exp(sX)\,(\exp(tX)p)\bigg|_{s=0} = X^F(\exp(tX)p).\end{aligned}$$

□

Theorem 40.8. *For any smooth left group action $G \times M \to M$, the map $\mathfrak{g} \to \mathfrak{X}(M) : X \mapsto X^F$ is a Lie group anti-homomorphism, i.e. it satisfies*

$$[X, Y]^F = -[X^F, Y^F]$$

for all $X, Y \in \mathfrak{g}$, implying in particular that the space $\mathfrak{X}^F(M)$ of fundamental vector fields is a Lie subalgebra of $\mathfrak{X}(M)$.

PROOF. As a preliminary remark, observe that for any $g \in G$ and $X \in \mathfrak{g}$, the map $\mathbb{R} \to G : t \mapsto g\exp(tX)g^{-1}$ is a Lie group homomorphism whose derivative at $t = 0$ is $\mathrm{Ad}_g(X)$, giving rise to the formula

$$g\exp(tX)g^{-1} = \exp\left(t\,\mathrm{Ad}_g(X)\right).$$

Given $X, Y \in \mathfrak{g}$, we can now use (37.4) and Proposition 40.7 to compute

$$\begin{aligned}
[X^F, Y^F](p) &= \partial_t\Big(\partial_s\big(\varphi^{-t}_{X^F} \circ \varphi^s_{Y^F} \circ \varphi^t_{X^F}(p)\big)\big|_{s=0}\Big)\Big|_{t=0} \\
&= \partial_t\Big(\partial_s\big(\exp(-tX)\exp(sY)\exp(tX)p\big)\big|_{s=0}\Big)\Big|_{t=0} \\
&= \partial_t\Big(\partial_s \exp\left(s\,\mathrm{Ad}_{\exp(-tX)}(Y)\right) p\big|_{s=0}\Big)\Big|_{t=0} = \partial_t\left(\mathrm{Ad}_{\exp(-tX)}(Y)\right)^F(p)\Big|_{t=0} \\
&= \left(\mathrm{ad}_{-X} Y\right)^F(p) = -[X, Y]^F(p).
\end{aligned}$$

□

REMARK 40.9. I'm sure that you, like I, would have liked Theorem 40.8 better without the extra minus sign in the statement. But you can see why it needs to be there if you consider the example of a group G acting from the left on itself: as observed in Example 40.6, the Lie algebra of fundamental vector fields on G is in this case precisely the Lie algebra of *right-invariant* vector fields $\mathfrak{X}^R(G)$, and the natural map $\mathfrak{g} \to \mathfrak{X}^F(G)$ becomes the natural isomorphism $\mathfrak{g} \to \mathfrak{X}^R(G) : X \mapsto X^R$, which we saw in Exercise 37.19 is not a Lie algebra homomorphism unless one changes the sign of the bracket on $\mathfrak{g}$. This is a side-effect of our choice to define the bracket on $\mathfrak{g}$ in terms of left-invariant instead of right-invariant vector fields.

It has an even more unfortunate consequence if one considers the following question: in what sense is $\mathrm{Diff}(M)$ an "infinite-dimensional Lie group", and what then is its Lie algebra? Heuristically, one would like to think of the tangent space $\mathfrak{diff}(M) := T_{\mathrm{Id}}\,\mathrm{Diff}(M)$ as consisting of derivatives at $s = 0$ of smooth families of diffeomorphisms $\{\varphi_s \in \mathrm{Diff}(M)\}_{s\in(-\epsilon,\epsilon)}$ with $\varphi_0 = \mathrm{Id}$, and computing such a derivative always gives rise to a vector field

$$X \in \mathfrak{X}(M), \qquad X(p) := \partial_s\varphi_s(p)|_{s=0} \in T_pM,$$

making it natural to define $\mathfrak{diff}(M) := \mathfrak{X}(M)$. There is an obvious guess for what the bracket on $\mathfrak{diff}(M)$ should be, but Theorem 40.8 tells us that this guess is only correct after changing a sign: indeed, for the obvious left action of $\mathrm{Diff}(M)$ on M defined by $\varphi \cdot p := \varphi(p)$, the natural surjection $\mathfrak{diff}(M) \to \mathfrak{X}^F(M)$ becomes the identity map, which is therefore a Lie algebra anti-homomorphism, meaning that the bracket on $\mathfrak{diff}(M)$ determined by the group structure of $\mathrm{Diff}(M)$ is *minus* the standard Lie bracket on $\mathfrak{X}(M)$. We wouldn't have this extra minus sign if we had chosen to define the Lie algebra of a Lie group in terms of right-invariant instead of left-invariant vector fields, and this strikes me as a strong argument for doing so, but that would of course lead to perilous inconsistencies with the existing literature. At present I am aware of exactly one book that attempts this: see [**Olv86**, Exercise 1.33].

As a coda to this remark, note that if heuristic discussions of $\mathrm{Diff}(M)$ as an infinite-dimensional Lie group make you uneasy, you can nonetheless apply the same reasoning to any finite-dimensional Lie group G that occurs as a subgroup of $\mathrm{Diff}(M)$ acting smoothly on M, e.g. we will see later that isometry groups of pseudo-Riemannian manifolds fit this description. The upshot is that the map $\mathfrak{g} \to \mathfrak{X}^F(M) : X \mapsto X^F$ in this situation gives a natural identification of the Lie algebra $\mathfrak{g}$ with some finite-dimensional Lie subalgebra of $\mathfrak{X}(M)$, but the bracket on this subalgebra is *minus* the standard Lie bracket of vector fields.

EXERCISE 40.10. Work out the analogue of Theorem 40.8 for a smooth right action, and show in particular that the unpleasant minus sign disappears. (This is unfortunately not helpful from the perspective of Remark 40.9, since $\mathrm{Diff}(M)$ cannot really be said to have a canonical *right* action on M.)

40.3. Orbits and slices. Quotients in topology can be notoriously ugly objects: it is easy to cook up simple examples of non-Hausdorff spaces by starting from very nice spaces like $\mathbb{R}^n$ and quotienting them by badly-chosen equivalence relations, e.g. via certain kinds of linear group actions as in Example 40.1. If we want M/G to be a smooth manifold in particular, then clearly some conditions need to be imposed.

Assuming $G \times M \to M : (g, p) \mapsto gp$ is a smooth action, we begin by fixing a point $p \in M$ and examining the smooth map

$$G \to M : g \mapsto gp \tag{40.1}$$

that sends G onto the orbit of p. Before we state some general results, here are a couple of instructive examples to keep in mind.

EXAMPLE 40.11. Writing $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$, any $(a, b) \in \mathbb{R}^2 \backslash \{(0,0)\}$ determines a smooth action of $\mathbb{R}$ on $\mathbb{T}^2$ defined by

$$t \cdot [(\theta, \phi)] := [(\theta + at, \phi + bt)].$$

The properties of this action and the quotient $\mathbb{T}^2/\mathbb{R}$ depend heavily on the choice of the constants a and b. We have the following cases:

(i) If $a/b \in \mathbb{Q}$ or $b = 0$, then the orbits $\mathbb{R}[(x, y)] \subset \mathbb{T}^2$ are all submanifolds diffeomorphic to S^1, while each of the maps (40.1) is a non-injective immersion covering the corresponding orbit infinitely-many times. It is not hard to show that the quotient is then homeomorphic to S^1, and in fact admits a natural smooth structure for which the quotient projection $\mathbb{T}^2 \to \mathbb{T}^2/\mathbb{R} : p \mapsto [p]$ is smooth. One obtains an even nicer picture after observing that there exists a unique constant $c > 0$ (dependent on a and b) such that the maps $\mathbb{R} \to \mathbb{T}^2$ defined by (40.1) are injective on $[0, c)$ but have period c, implying that the action of the subgroup $c\mathbb{Z} \subset \mathbb{R}$ on $\mathbb{T}^2$ is trivial, so the $\mathbb{R}$-action can be replaced by an action of the quotient group $G := \mathbb{R}/c\mathbb{Z} \cong S^1$ without changing the orbits or the quotient. After this modification, (40.1) gives an embedding $G \to \mathbb{T}^2$ and thus an explicit diffeomorphism of each orbit $\mathbb{R}[(x, y)] = G[(x, y)]$ with the group $G = \mathbb{R}/c\mathbb{Z} \cong S^1$.

(ii) If $b \neq 0$ and $a/b \notin \mathbb{Q}$, then each of the maps (40.1) is an injective immersion $\mathbb{R} \to \mathbb{T}^2$ with dense image, making $\mathbb{T}^2/\mathbb{R}$ a horrible non-Hausdorff space in which every sequence converges to every point, because the only open sets are $\varnothing$ and $\mathbb{T}^2/\mathbb{R}$ itself.

EXAMPLE 40.12. Fix two antipodal points $p_\pm \in S^2$ and let $\mathbb{Z}_2$ act on S^2 via the involution that rotates S^2 by 180 degrees around the line in $\mathbb{R}^3$ connecting p_+ to p_-. Equivalently, one can use stereographic projection to identify S^2 with the one-point compactification $\mathbb{R}^2 \cup \{\infty\}$ of $\mathbb{R}^2$ such that $p_+ = 0$ and $p_- = \infty$, and then define the involution in question as the unique continuous extension of the antipodal map $\mathbb{R}^2 \to \mathbb{R}^2 : \mathbf{v} \mapsto -\mathbf{v}$. For this action, the two points p_+ and p_- are special because their stabilizers are $\mathbb{Z}_2$ and their orbits each consist of only one point, whereas every point in $S^2 \backslash \{p_+, p_-\}$ has trivial stabilizer and two points in its orbit. A similar $\mathbb{Z}_2$-action can be defined on S^n for any $n \in \mathbb{N}$ by identifying the latter with $\mathbb{R}^n \cup \{\infty\}$ and letting the nontrivial element of $\mathbb{Z}_2$ act as the unique continuous extension of the antipodal map $\mathbb{R}^n \to \mathbb{R}^n$, which fixes the point at infinity. There are again two special points $p_\pm \in S^n$ that are fixed under this action.

EXERCISE 40.13. Show that for the $\mathbb{Z}_2$-actions on S^2 and S^3 in Example 40.12:

(a) $S^2/\mathbb{Z}_2$ is a topological manifold homeomorphic to S^2.

(b) $(S^3\backslash\{p_+, p_-\})/\mathbb{Z}^2$ admits a smooth manifold structure for which the projection of $S^3\backslash\{p_+, p_-\}$ to the quotient is smooth, but $S^3/\mathbb{Z}_2$ does not. (One can show in fact that $S^n/\mathbb{Z}_2$ is not a topological manifold for any $n \geqslant 3$; see Remark 40.14 below.)
Hint: The points $[p_\pm] \in S^3/\mathbb{Z}_2$ each have arbitrarily small neighborhoods bounded by a surface diffeomorphic to $\mathbb{RP}^2$. Show that if these neighborhoods had smooth structures compatible with the smooth structure of S^3, it would follow that $\mathbb{RP}^2$ is orientable.

Remark 40.14 (for readers who like topology). The fact that the quotient $S^2/\mathbb{Z}_2$ in Exercise 40.13 is a topological manifold should be understood as a by-product of the coincidence that $\mathbb{RP}^1$ and S^1 are homeomorphic. On the flip side, the fact that $\mathbb{RP}^n$ and S^n are not even weakly homotopy equivalent for $n \geqslant 2$ (e.g. their fundamental groups are different since S^n is simply connected and admits a double cover of $\mathbb{RP}^n$) can be used to prove that the two fixed points $[p_\pm] \in S^n/\mathbb{Z}_2$ do not have any neighborhoods homeomorphic to $\mathbb{R}^n$ when $n \geqslant 3$. Indeed, if $[p_+]$ had a compact disk-like neighborhood $A_1 \subset S^n/\mathbb{Z}_2$ with $\partial A_1 \cong S^{n-1}$, then one could find inside it two smaller compact neighborhoods $A_2 \subset B_1 \subset A_1$ such that $\partial A_2 \cong S^{n-1}$, $\partial B_1 \cong \mathbb{RP}^{n-1}$ and $A_1\backslash\mathring{A}_2 \cong [0,1] \times S^{n-1}$. Continuing in this way gives rise to a nested sequence of compact submanifolds $A_1 \supset B_1 \supset A_2 \supset B_2 \supset \ldots$ such that $\partial A_j \cong S^{n-1}$ and $\partial B_j \cong \mathbb{RP}^{n-1}$ while $A_j\backslash\mathring{A}_{j+1} \cong [0,1] \times S^{n-1}$ and $B_j\backslash\mathring{B}_{j+1} \cong [0,1] \times \mathbb{RP}^{n-1}$ for every j. Since inclusion maps such as $S^{n-1} \hookrightarrow [0,1] \times S^{n-1}$ are homotopy equivalences, one can use compositions of inclusion maps in this situation to deduce that $\mathbb{RP}^{n-1}$ and S^{n-1} must be weakly homotopy equivalent. (For a similar argument in more detail, see [**CE12**, Lemma 16.13].) Alternatively, when n is odd, one could appeal to the well-known fact from algebraic topology that $\mathbb{RP}^{n-1}$ is not homeomorphic to the boundary of any compact topological manifold. (This follows from Poincaré duality and the fact that $\chi(\mathbb{RP}^{n-1})$ in these cases is odd, because gluing two copies of a compact n-manifold with boundary $\mathbb{RP}^{n-1}$ along their boundaries would then produce a closed n-manifold whose Euler characteristic cannot be 0.)

Throughout the following, assume $G \times M \to M : (g, p) \mapsto gp$ is a smooth group action.

Lemma 40.15. *For each $p \in M$, an element $X \in \mathfrak{g}$ satisfies $X^F(p) = 0$ if and only if $\exp(tX)$ belongs to the stabilizer G_p for all $t \in \mathbb{R}$.*

Proof. If $\exp(tX) \in G_p$ for all $t \in \mathbb{R}$ then $X^F(p) = \partial_t \exp(tX)p|_{t=0} = \partial_t p|_{t=0} = 0$. Conversely, $X^F(p) = 0$ implies via Proposition 40.7 that $\exp(tX)p = \varphi^t_{X^F}(p) = p$ for all $t \in \mathbb{R}$, so $\exp(tX) \in G_p$. □

One piece of intuition you might gather from Example 40.11 is that the space of orbits M/G is not likely to have a nice structure unless the orbits themselves are well behaved, e.g. we would ideally hope that they are smooth submanifolds, with the map (40.1) as a natural embedding. The following condition is clearly necessary for this.

Definition 40.16. An action of G on M is called **free** if the stabilizer $G_p \subset G$ of every point $p \in M$ is the trivial subgroup. Equivalently, the action is free if for every $g \in G\backslash\{e\}$, the bijection $M \to M : p \mapsto gp$ has no fixed points. (One also says that G **acts without fixed points** on M.)

Lemma 40.17. *If G acts smoothly and freely on M, then for every $p \in M$, the map $\varphi : G \to M : g \mapsto gp$ is an injective immersion, and its derivative at $e \in G$ is given by $\mathfrak{g} \to T_pM : X \mapsto X^F(p)$.*

Proof. If $\varphi(g) = \varphi(h)$ for some $g, h \in G$, then $h^{-1}g \in G_p$, which implies $g = h$ since $G_p \subset G$ is the trivial group, thus φ is injective. The formula $T_e\varphi(X) = X^F(p)$ for the derivative of φ follows directly from the definition of fundamental vector fields, and according to Lemma 40.15,

$X^F(p) = 0$ if and only if $\exp(tX) \in G_p$ for all $t \in \mathbb{R}$, which means $X = 0$ since G_p is trivial. Finally, for any $g \in G$, denote $\psi : G \to M : h \mapsto hgp$ and observe that

$$\psi = \varphi \circ R_g,$$

thus $T_e\psi = T_g\varphi \circ T_e R_g$, and $T_g\varphi$ is therefore injective if and only if $T_e\psi$ is injective. The latter follows from the previous argument since the stabilizer group G_{gp} is also trivial. $\square$

Exercise 40.18. Assume $G \times M \to M$ is a smooth group action.

(a) Show that whenever $p, q \in M$ belong to the same orbit, their stabilizer subgroups $G_p, G_q \subset G$ are conjugate.
(b) Show that the map (40.1) is still an injective immersion if the action is not free but the particular stabilizer G_p is trivial.
(c) Show that the map (40.1) is still an immersion (though not injective) if the stabilizer G_p is nontrivial but discrete.

If the Lie group G is compact, then Lemma 40.17 implies that the orbits of any *free* smooth action of G on a manifold M are smooth submanifolds. This need not hold however if G is not compact, as injective immersions $\varphi : G \to M$ are not always embeddings: the inverse map $M \supset \varphi(G) \to G$ might fail to be continuous, as seen for instance in the case $a/b \notin \mathbb{Q}$ of Example 40.11. With this in mind, there is another topological condition one can usefully impose: a map $f : X \to Y$ between two topological spaces is called **proper** (*eigentlich*) if for every compact set $K \subset Y$, the preimage $f^{-1}(K) \subset X$ is also compact.

Exercise 40.19. Show that for any proper injective continuous map $f : M \to N$ between two metric spaces, the image $f(M) \subset N$ is closed and the inverse $N \supset f(M) \to M$ is continuous. In particular, a smooth proper injective immersion between manifolds is also an embedding, whose image is a closed subset and smooth submanifold.

The map (40.1) from G to M is automatically proper whenever G is compact, but there are also plenty of interesting examples involving noncompact groups for which this map is proper. The next condition ensures that this will hold in a "uniform" manner for all orbits:

Definition 40.20. For a topological group G and topological space M, a continuous action $G \times M \to M$ is called **proper** if the map $G \times M \to M \times M : (g, p) \mapsto (p, gp)$ is proper.

Exercise 40.21. Show that every smooth action of a compact Lie group is proper.

Exercise 40.22. Show that for any smooth and proper action of G on M:

(a) The quotient M/G is Hausdorff.
(b) The stabilizer $G_p \subset G$ is compact for every $p \in M$.
(c) If $p_j \in M$ is a sequence convergent to $p \in M$, then any sequence $g_j \in G_{p_j}$ has a subsequence convergent to an element of G_p.

Corollary 40.23 (of Lemma 40.17). *If G is a Lie group acting smoothly, freely and properly on a manifold M, then for every $p \in M$, the map (40.1) is an embedding and the orbit $Gp \subset M$ is a closed subset and smooth submanifold of M.* $\square$

Exercise 40.24. You may have encountered the following definition in a topology course: a continuous free action $G \times M \to M$ is called **properly discontinuous** (*eigentlich diskontinuierlich*) if every $p \in M$ has a neighborhood $\mathcal{U} \subset M$ such that $g\mathcal{U} \cap \mathcal{U} = \emptyset$ for all $g \in G\backslash\{e\}$. Show that a continuous free action of a Lie group G on a manifold M is properly discontinuous if and only if it is proper and the group G is discrete.[91]

We are now ready to state a result that determines which group actions have the nicest quotients.

THEOREM 40.25 (slice theorem for free and proper actions). *Assume $G \times M \to M$ is a smooth, free and proper action of a Lie group G on a manifold M, and $\Sigma \subset M$ is a smooth submanifold containing a point $p \in \Sigma$ such that*

$$\Sigma \cap Gp = \{p\} \qquad \text{and} \qquad T_p\Sigma \oplus T_p(Gp) = T_pM.$$

Then after replacing Σ if necessary by a sufficiently small open neighborhood of p in Σ, the map

$$\Phi : G \times \Sigma \to M : (g, q) \mapsto gq$$

is a diffeomorphism onto an open and G-invariant neighborhood of Gp in M.

Before proving the theorem, let us explain its main application. Observe that submanifolds $\Sigma \subset M$ satisfying the hypotheses of the theorem clearly exist, due to the fact (from Corollary 40.23) that $Gp \subset M$ is a submanifold: it is easy to write down examples in local coordinates on a neighborhood of $p \in M$ using a slice chart for Gp. A submanifold Σ for which the map Φ in the theorem is a diffeomorphism is sometimes called a **local slice** for the group action, as it has the important property that for some open neighborhood $\mathcal{U} \subset M/G$ of $[p]$ in the space of orbits, Σ intersects every orbit in this set exactly once.

COROLLARY 40.26. *For any smooth, free and proper group action $G \times M \to M$, the quotient M/G admits a smooth manifold structure with*

$$\dim(M/G) = \dim M - \dim G$$

such that the quotient projection $\pi : M \to M/G$ is a smooth submersion.

PROOF. For any $p \in M$, a chart for M/G on a neighborhood of $[p] \in M/G$ can be defined by choosing any local slice $\Sigma \subset M$ through p, as provided by Theorem 40.25, and observing that the restricted quotient map

$$\Sigma \xrightarrow{\pi} M/G \tag{40.2}$$

is a homeomorphism onto a neighborhood of $[p]$, so that composing its inverse with a smooth chart for Σ near p defines a chart for M/G. That (40.2) is a homeomorphism onto an open set follows as an easy exercise from the fact that $\Phi : G \times \Sigma \to M$ is an equivariant diffeomorphism onto a G-invariant neighborhood of Gp. To see that any two charts on M/G constructed in this way from smooth local slices $\Sigma_1, \Sigma_2 \subset M$ are smoothly compatible, let us first simplify the notation by assuming after shrinking Σ_1 and Σ_2 that their projections cover the same region

$$\mathcal{V} := \pi(\Sigma_1) = \pi(\Sigma_2) \subset M/G,$$

in which case they also have identical orbits $G\Sigma_1 = G\Sigma_2 = \pi^{-1}(\mathcal{V}) \subset M$. Theorem 40.25 now gives rise to a G-equivariant diffeomorphism

$$\begin{tikzcd} G \times \Sigma_1 \arrow[r, "\Phi"] \arrow[rr, bend right, "\Psi"'] & \pi^{-1}(\mathcal{V}) \arrow[r, "\Phi^{-1}"] & G \times \Sigma_2, \end{tikzcd}$$

[91] I am intentionally avoiding giving a definition of the term "properly discontinuous" for group actions that are not free. There exist various inequivalent versions of such a definition in the literature, but they are only needed in situations that are more specialized than we will consider here.

where equivariance in this case means that Ψ takes the form

$$\Psi(g, q) = (gf(q), \varphi(q))$$

for smooth maps $f : \Sigma_1 \to G$ and $\varphi : \Sigma_1 \to \Sigma_2$, and the latter is necessarily a diffeomorphism. The transition map for our two charts is then a composition of $\varphi : \Sigma_1 \to \Sigma_2$ with charts (or their inverses) for Σ_1 and Σ_2, thus it is smooth.

That $\pi : M \to M/G$ is a submersion with respect to any of these charts follows from the fact that the obvious projection $G \times \Sigma \to \Sigma$ is a submersion. $\square$

EXERCISE 40.27. In the setting of Corollary 40.26, show:

(a) A map $f : M/G \to N$ to another smooth manifold is smooth if and only if the composition $f \circ \pi : M \to N$ is smooth.
(b) The derivative at $p \in M$ of the quotient projection $\pi : M \to M/G$ descends to a vector space isomorphism

$$T_pM/T_p(Gp) \xrightarrow{\cong} T_{[p]}(M/G).$$

Hint: You do not need to know how the charts on M/G are constructed—all you actually need to know is that $\pi : M \to M/G$ is a submersion. This also shows that the submersion condition uniquely determines the smooth structure of M/G.

EXERCISE 40.28. Assume $G \times M \to M$ and $H \times N \to N$ are a pair of smooth, free and proper group actions. A map $F : M \to N$ is said to **descend** to the quotient if there exists a map $f : M/G \to N/H$ such that $f([p]) = [F(p)]$ for every $p \in M$. Prove:

(a) If $G = H$ and the map $F : M \to N$ is G-equivariant, then it descends to the quotient.
(b) If $F : M \to N$ is smooth and descends to the quotient, then the induced map $f : M/G \to N/H$ is also smooth.

EXERCISE 40.29. Check that the group actions in Examples 40.4 and 40.5 whose quotients are projective spaces are free and proper, and that the resulting smooth structures on $\mathbb{RP}^n$ and $\mathbb{CP}^n$ match the smooth structures defined via explicit charts as in Exercise 32.11.
Hint: Thanks to Exercise 40.27, you can use the submersion property to characterize the smooth structure on any quotient.

PROOF OF THEOREM 40.25. We observe first that the orbit tangent spaces $T_q(Gq) \subset T_qM$ form a smooth subbundle of TM, as one can see by choosing any basis $X_1, \dots, X_n \in \mathfrak{g}$ and using the values of the corresponding fundamental vector fields $X_1^F(q), \dots, X_n^F(q) \in T_qM$ as a basis of $T_q(Gq)$ at each point. It follows that the condition $T_p\Sigma \oplus T_p(Gp) = T_pM$ is open, i.e. if it holds at the given point $p \in \Sigma$, then it also holds for all nearby points $q \in \Sigma$. We are thus free to assume after shrinking Σ if necessary that $T_q\Sigma \oplus T_q(Gq) = T_qM$ for all $q \in \Sigma$, and with this assumption in place, we claim that the map $\Phi : G \times \Sigma \to M : (g, q) \mapsto gq$ is a local diffeomorphism. Indeed, at any point of the form $(e, q) \in G \times \Sigma$, the derivative $T_{(e,q)}\Phi : T_{(e,q)}(G \times \Sigma) = \mathfrak{g} \times T_q\Sigma \to T_qM$ can be written as

$$T_{(e,q)}\Phi(X, v) = X^F(q) + v \qquad \text{for } X \in \mathfrak{g},\ v \in T_q\Sigma,$$

and this map is an isomorphism since $\mathfrak{g} \to T_q(Gq) : X \mapsto X^F(q)$ is an isomorphism and $T_q(Gq) \oplus T_q\Sigma = T_qM$. To understand $T_{(g,q)}\Phi$ at an arbitrary point $(g, q) \in G \times \Sigma$, we can now make use of the observation that for the obvious G-action on $G \times \Sigma$ defined by $g \cdot (h, p) := (gh, p)$, Φ is equivariant, i.e.

$$\Phi\big(g \cdot (h, p)\big) = \Phi(gh, p) = ghp = g \cdot \Phi(h, p).$$

For a given $g \in G$, let $\varphi_g : M \to M$ and $\Psi_g : G \times \Sigma \to G \times \Sigma$ denote the diffeomorphisms defined via the two G-actions: then $\Phi \circ \Psi_g = \varphi_g \circ \Phi$, and thus

$$T_{(e,q)}(\Phi \circ \Psi_g) = T_{(g,q)}\Phi \circ T_{(e,q)}\Psi_g = T_q\varphi_g \circ T_{(e,q)}\Phi.$$

Since $T_{(e,q)}\Psi_g$, $T_q\varphi_g$ and $T_{(e,q)}\Phi$ are all isomorphisms, it follows that $T_{(g,q)}\Phi$ is as well, proving the claim.

Next, we claim that after shrinking Σ further if necessary, Φ is injective. If not, then there exist sequences (g_j, p_j) and (h_j, q_j) in $G \times \Sigma$ such that $p_j, q_j \to p$, $(g_j, p_j) \neq (h_j, q_j)$ and $g_j p_j = h_j q_j$ for all j. The latter implies $q_j = h_j^{-1} g_j p_j$ and thus

$$\Phi(e, q_j) = q_j = a_j p_j = \Phi(a_j, p_j) \qquad \text{for } a_j := h_j^{-1} g_j \in G,$$

and since p_j and q_j both converge, the properness of the action then implies that a_j converges to some element $g \in G$ after restricting to a subsequence. Continuity now implies $gp = p$, thus $g = e$ since the action is free. Since Φ maps a neighborhood of $(e, p) \in G \times \Sigma$ diffeomorphically onto a neighborhood of $p \in M$, it follows that $(a_j, p_j) = (e, q_j)$ for all j sufficiently large, which means $(g_j, p_j) = (h_j, q_j)$ and is thus a contradiction, proving the claim.

Finally, having restricted Σ to make Φ an injective local diffeomorphism, note that after shrinking Σ further, we are free to assume that it has compact closure $\bar{\Sigma} \subset M$ and that Φ is still injective on $G \times \bar{\Sigma}$. The properness of the action then implies that the map $\Phi|_{G\times\bar{\Sigma}} : G \times \bar{\Sigma} \to M$ is proper, so by Exercise 40.19, its inverse is also continuous. $\square$

EXERCISE 40.30. Prove a converse to the slice theorem, i.e. assuming $G \times M \to M$ is a smooth group action for which local slices as described in Theorem 40.25 exist through any point, prove that the action must be free and proper. *Hint: Given a manifold Σ, what can you say about the action of G on $G \times \Sigma$ defined by $g \cdot (h, p) := (gh, p)$?*

EXERCISE 40.31. Given a pair of smooth, free and proper group actions $G \times M \to M$ and $H \times N \to N$, prove:

(a) The Lie group $G \times H$ acts smoothly, freely and properly on $M \times N$ by $(g, h) \cdot (p, q) := (gp, hq)$.
(b) The map $(M \times N)/(G \times H) \to (M/G) \times (N/H) : [(p, q)] \mapsto ([p], [q])$ is well defined and gives a diffeomorphism with respect to the smooth structures on quotients arising from Corollary 40.26.

As mentioned in §40.1, all results stated in this lecture for left actions have more-or-less obvious analogues for right actions. The following exercise uses a right action so that it can subsequently define a *left* action on the resulting quotient.

EXERCISE 40.32. Assume that G is a Lie group with a Lie subgroup $H \subset G$. Prove:

(a) The left and right actions of H on G defined by

$$H \times G \to G : (h, g) \mapsto hg \qquad \text{and} \qquad G \times H \to G : (g, h) \mapsto gh$$

are smooth, free and proper.
(b) Defining G/H as the set of left cosets gH (i.e. the quotient by the right action in part (a)) and endowing it with the smooth structure arising from Corollary 40.26, the map

$$G \times (G/H) \to G/H : (g, aH) \mapsto gaH$$

defines a smooth left action of G on G/H.
Hint: Exercise 40.31 makes $G \times (G/H)$ diffeomorphic to the quotient of $G \times G$ by a free and proper action of some product subgroup. You can therefore use Exercise 40.27 to check the smoothness of a map defined on $G \times (G/H)$.
(c) If the subgroup $H \subset G$ is normal, then G/H has a natural Lie group structure for which the quotient projection $G \to G/H$ is a Lie group homomorphism.

EXERCISE 40.33. A mild generalization of Theorem 40.25 is sometimes needed in the study of *moduli spaces*, i.e. sets of geometric objects (such as algebraic curves, or solutions to elliptic PDEs) which can be described via finitely-many parameters. Assume $G \times M \to M$ is a smooth group action that is proper and not necessarily free, but instead has **finite isotropy**, meaning that the stabilizer $G_p \subset G$ is finite for every $p \in M$.

(a) Show that for each $p \in M$, the orbit $Gp \subset M$ is a smooth submanifold. Unlike the case of a free action, Gp will not generally be diffeomorphic to G. What instead?
(b) Show that p admits a G_p-invariant neighborhood $\mathcal{U} \subset M$ with a G_p-invariant Riemannian metric, i.e. the diffeomorphisms $\mathcal{U} \to \mathcal{U} : q \mapsto gq$ are isometries for each $g \in G_p$.
Hint: Start with any metric, then act on it with every element of G_p and take an average.
(c) Construct a submanifold $\Sigma \subset M$ that satisfies $\Sigma \cap Gp = \{p\}$ and $T_p\Sigma \oplus T_p(Gp) = T_pM$ and, additionally, is invariant under the action of G_p. *Hint: Use geodesics.*
(d) Show that after possibly shrinking Σ to a smaller neighborhood of p, the map $\Phi : G \times \Sigma \to M : (g, q) \mapsto gq$ can be assumed to be a local diffeomorphism satisfying

$$g(\Sigma) \cap \Sigma = \varnothing \qquad \text{for all } g \in G \backslash G_p.$$

(e) Deduce that the map $\Sigma \to M/G : p \mapsto [p]$ descends to the quotient Σ/G_p as a homeomorphism onto a neighborhood of $[p]$ in M/G.
(f) Writing $n := \dim M - \dim G$, conclude that the topological space M/G has the following local structure generalizing the notion of a manifold: every point has a neighborhood homeomorphic to the quotient of an open subset of $\mathbb{R}^n$ by a smooth finite group action. A space with this local structure is called an n-dimensional **orbifold**.
(g) Show that every $p \in M$ has a neighborhood $\mathcal{U} \subset M$ such that $|G_q| \leqslant |G_p|$ for all $q \in \mathcal{U}$. In particular, the set of points with trivial stabilizers is an open G-invariant subset $M^* \subset M$, and M^*/G is an n-dimensional manifold.

41. Closed subgroups and their quotients

We have two more pieces of general Lie group theory to cover before moving on to the next topic. After proving the closed subgroup theorem in §41.1, we will briefly discuss homogeneous spaces in §41.2, which form a large class of manifolds with smooth group actions, including important examples such as the Grassmannians and Stiefel manifolds that describe sets of linear subspaces and frames on those subspaces. These will be discussed in §41.3, and will also arise naturally in our introduction to fiber bundles in the next lecture.

41.1. The closed subgroup theorem. Subgroups of Lie groups do not have to be *Lie subgroups* in general: for example, $\mathbb{R}$ is a Lie group with respect to addition, and the rational numbers $\mathbb{Q} \subset \mathbb{R}$ form a subgroup, but clearly not a submanifold. The main result of this section will make it easy to recognize when a subgroup of a Lie group is also a submanifold. It will imply in particular that for a smooth group action on a manifold, stabilizer subgroups are *always* Lie subgroups, and can thus be used to form quotient manifolds as in Exercise 40.32.

The proof of the theorem below would be somewhat easier if one could assume in every Lie group G that the relation $\exp(X+Y) = \exp(X)\exp(Y)$ holds for all $X, Y \in \mathfrak{g}$, but of course that is not true in general if $[X, Y] \neq 0$. We therefore need the following lemma as a tool for quantifying the extent to which this relation fails.

LEMMA 41.1. *For any Lie group G with Lie algebra $\mathfrak{g}$, there exists a neighborhood $\mathcal{O} \subset \mathfrak{g} \times \mathfrak{g}$ of $(0,0)$ and a smooth function $R : \mathcal{O} \to \operatorname{Hom}(\mathfrak{g} \times \mathfrak{g}, \mathfrak{g})$ such that for all $(X, Y) \in \mathcal{O}$,*

$$\exp(X)\exp(Y) = \exp\big(X + Y + R(X,Y)(X,Y)\big), \qquad \text{and} \qquad R(0,0) = 0.$$

Proof. Since $\exp : \mathfrak{g} \to G$ maps a neighborhood of $0 \in \mathfrak{g}$ diffeomorphically onto a neighborhood of $e \in G$, there exists for any sufficiently small neighborhood $\mathcal{O} \subset \mathfrak{g} \times \mathfrak{g}$ of $(0,0)$ a unique smooth function $F : \mathcal{O} \to \mathfrak{g}$ taking values in a similarly small neighborhood of 0 such that

$$\exp\left(X + Y + F(X,Y)\right) = \exp(X)\exp(Y).$$

Moreover, F satisfies $F(X,0) = 0$ for all X and $F(0,Y) = 0$ for all Y, implying that both F and its first derivative vanish at the point $(0,0)$. The fundamental theorem of calculus thus implies

$$F(X,Y) = \int_0^1 \frac{d}{d\tau} F(\tau X, \tau Y)\, d\tau = \left(\int_0^1 DF(\tau X, \tau Y)\, d\tau\right)(X,Y),$$

and we define $R(X,Y) \in \operatorname{Hom}(\mathfrak{g} \times \mathfrak{g}, \mathfrak{g})$ as the integral in parentheses. □

Theorem 41.2. *Suppose G is a Lie group and $H \subset G$ is a subgroup. Then H is a Lie subgroup*[92] *of G if and only if it is a closed subset.*

Proof. Assume first that $H \subset G$ is a Lie subgroup, meaning that it is both a subgroup and a submanifold, thus we can find a neighborhood $\mathcal{U} \subset G$ of $e \in G$ admitting a slice chart for H, making $H \cap \mathcal{U}$ a closed subset of $\mathcal{U}$. Now suppose $h_k \in H$ is a sequence converging in G to an element $g \in G$. Then since h_k is a Cauchy sequence for a suitable choice of metric, we can assume $h_k h_j^{-1} \in \mathcal{U}$ whenever j and k are both sufficiently large. Fixing j and letting $k \to \infty$, we have $h_k h_j^{-1} \to g h_j^{-1} \in \mathcal{U}$, but since $h_k h_j^{-1}$ belongs to the closed subset $H \cap \mathcal{U} \subset \mathcal{U}$, this implies $g h_j^{-1}$ is also in H, and therefore so is g. This completes the proof that $H \subset G$ is closed.

The converse is the truly surprising part: we need to show that if $H \subset G$ is both a subgroup and a closed subset, then it is a submanifold, i.e. every $h \in H$ has a neighborhood in G admitting a slice chart for H. It suffices in fact to prove this for $h = e$, because if $G \supset \mathcal{U} \xrightarrow{x} \mathbb{R}^n$ is a slice chart with $e \in \mathcal{U}$, then for any $h \in H$, the map

$$G \supset L_h(\mathcal{U}) \xrightarrow{x \circ L_{h^{-1}}} \mathbb{R}^n$$

will be a slice chart with $h \in L_h(\mathcal{U})$. (Here the slice condition follows from the observation that $g \in H \cap \mathcal{U}$ if and only if $L_h(g) \in H \cap L_h(\mathcal{U})$.)

With this understood, our goal now is to construct a slice chart for H on a neighborhood of the identity, or equivalently, to find a vector space V, subspace $W \subset V$, neighborhood $\mathcal{O} \subset V$ of 0 and smooth map $\varphi : \mathcal{O} \to G$ that is a diffeomorphism onto a neighborhood of e such that $\varphi^{-1}(H) = W \cap \mathcal{O}$. A good candidate for φ is the exponential map, so V will be the Lie algebra $\mathfrak{g}$, and we expect the subspace $W \subset V$ to be the Lie algebra of the subgroup H. With this expectation in mind, we start by identifying a candidate for the latter: let

$$\mathfrak{h} := \left\{ X \in \mathfrak{g} \;\middle|\; \exp(tX) \in H \text{ for all } t \in \mathbb{R} \right\}.$$

We will see below that $\mathfrak{h} \subset \mathfrak{g}$ is a linear subspace; we can already see clearly that it is closed under scalar multiplication.

Choose a norm $\|\cdot\|$ on $\mathfrak{g}$ and denote the unit sphere in any linear subspace $V \subset \mathfrak{g}$ by

$$S(V) := \left\{ X \in V \;\middle|\; \|X\| = 1 \right\}.$$

Here is the first application of the hypothesis that $H \subset G$ is closed: we claim that if $X_k \in \mathfrak{g}\backslash\{0\}$ is any sequence converging to 0 such that $\exp(X_k) \in H$ for all k and the sequence $X_k/\|X_k\| \in S(\mathfrak{g})$ converges to an element $X \in S(\mathfrak{g})$, then $X \in \mathfrak{h}$. Indeed, for any $t \in \mathbb{R}$, the assumption that $X_k \to 0$

[92]Recall that the definition of the term "Lie subgroup" we are using is stricter than what appears in some textbooks: for us, a Lie subgroup is always also a smoothly *embedded* submanifold (cf. Remark 37.3).

makes it possible to choose a sequence $n_k \in \mathbb{Z}$ with $|n_k| \to \infty$ such that $n_k X_k \to tX$. Since $H \subset G$ is a subgroup and $t \mapsto \exp(tY)$ is a group homomorphism $\mathbb{R} \to G$ for each $Y \in \mathfrak{g}$, it follows that

$$H \ni (\exp(X_k))^{n_k} = \exp(n_k X_k) \to \exp(tX),$$

and since $H \subset G$ is closed, this proves $\exp(tX) \in H$.

We claim next that $\mathfrak{h} \subset \mathfrak{g}$ is closed under vector addition and is thus a linear subspace. For this there is a similar trick as in the previous claim: given $X, Y \in \mathfrak{h}$ and $t \in \mathbb{R}$, we plug the product $\exp(tX/k)\exp(tY/k) \in H$ for $k \in \mathbb{N}$ into Lemma 41.1, take its kth power and let $k \to \infty$, obtaining

$$\begin{aligned} H \ni (\exp(tX/k)\exp(tY/k))^k &= \left(\exp\left(\frac{t}{k}(X+Y) + \frac{t^2}{k^2}R(tX/k, tY/k)(X,Y)\right)\right)^k \\ &= \exp\left(t(X+Y) + \frac{t^2}{k}R(tX/k, tY/k)(X,Y)\right) \to \exp(t(X+Y)). \end{aligned}$$

The assumption that $H \subset G$ is closed thus implies $\exp(t(X+Y)) \in H$ as claimed.

We will now be done if we can show that some neighborhood of e in H is contained in $\exp(\mathfrak{h})$. To see this, suppose $h_k \in H$ is a sequence with $h_k \to e$ but $h_k \notin \exp(\mathfrak{h})$. Choosing a linear subspace $\mathfrak{h}^\perp \subset \mathfrak{g}$ complementary to $\mathfrak{h}$, we can if necessary discard finitely many terms lying outside a small neighborhood of e in order to write

$$h_k = \exp(X_k + Y_k),$$

for unique sequences $X_k \in \mathfrak{h}$ and $Y_k \in \mathfrak{h}^\perp$ that both converge to 0. By assumption $Y_k \neq 0$ for every k, and since $S(\mathfrak{h}^\perp)$ is compact, we can then assume after restricting to a subsequence that $Y_k/\|Y_k\|$ converges to some element $Y \in S(\mathfrak{h}^\perp)$. But consider the sequence

$$g_k := \exp(-X_k)h_k = \exp(-X_k)\exp(X_k + Y_k) \in H,$$

which also converges to e and can thus be written as $g_k = \exp(Z_k)$ for a unique sequence $Z_k \in \mathfrak{g}$ that converges to 0 and satisfies $Z_k \neq 0$ for all k. Restricting to a further subsequence, we can assume the sequence $Z_k/\|Z_k\| \in S(\mathfrak{g})$ has a limit $Z \in S(\mathfrak{g})$, which must lie in $S(\mathfrak{h})$ according to the first claim above. Using Lemma 41.1, we have

$$\exp(X_k + Y_k) = h_k = \exp(X_k)g_k = \exp(X_k)\exp(Z_k) = \exp\left(X_k + Z_k + R(X_k, Z_k)(X_k, Z_k)\right),$$

implying

$$Y_k = Z_k + R(X_k, Z_k)(X_k, Z_k)$$

for all k. Since $R(X_k, Z_k) \to R(0,0) = 0$, this relation implies $Y = Z$, which is impossible since $S(\mathfrak{h}) \cap S(\mathfrak{h}^\perp) = \emptyset$. $\square$

COROLLARY 41.3 (using Lemma 40.15). *For any smooth group action $G \times M \to M$ and any point $p \in M$, the stabilizer $G_p \subset G$ is a Lie subgroup, and its Lie algebra $\mathfrak{g}_p \subset \mathfrak{g}$ is the space of all $X \in \mathfrak{g}$ for which the corresponding fundamental vector field $X^F \in \mathfrak{X}(M)$ vanishes at p.* $\square$

41.2. Homogeneous spaces. For a Lie group G with closed subgroup $H \subset G$, Exercise 40.32 shows that the smooth actions of H on G from the left or the right are free and proper, so the resulting quotients G/H always have natural smooth structures, though they are not generally groups unless the subgroup H is normal. Manifolds that arise in this way as quotients of Lie groups are called **homogeneous spaces**. The reason for this term has to do with a special class of group actions, namely those for which every point is in the same orbit with every other point.

DEFINITION 41.4. A group action is **transitive** if it has only one orbit.

Example 41.5. For any smooth manifold M, the topological group $\mathrm{Diff}(M)$ has a natural left action on M defined by $\varphi \cdot p := \varphi(p)$. If M is connected, then this action is transitive because for any two points $p, q \in M$, there exists a diffeomorphism $\varphi : M \to M$ with $\varphi(p) = q$. (Hint: construct φ as the flow of a compactly-supported vector field.)

Example 41.6. For a pseudo-Riemannian manifold (M, g), we will see later in this course that the group of isometries $\mathrm{Isom}(M, g) \subset \mathrm{Diff}(M)$ has a natural Lie group structure for which the action $\mathrm{Isom}(M, g) \times M \to M : (\varphi, p) \mapsto \varphi(p)$ is smooth. We say that (M, g) is **homogeneous** if this action is transitive. Homogeneous manifolds form a special class of pseudo-Riemannian manifolds that can be said to "look the same" (in suitable coordinates) at any point.

Example 41.7. The natural action of a Lie group G on itself from the left or the right is obviously transitive. More generally, we recall from Exercise 40.32 that if we form a homogeneous space G/H via a closed subgroup $H \subset G$ acting on G from the *right*, then G acts smoothly on G/H from the *left* by

$$G \times G/H \to G/H : (g, aH) \mapsto gaH. \tag{41.1}$$

The transitivity of the action of G on itself implies immediately that this action on the quotient is also transitive.

Exercise 41.8. Show that for the natural action of $\mathrm{SO}(3)$ on $S^2 \subset \mathbb{R}^3$ by linear transformations restricted to the unit sphere, the only Lie subgroup of $\mathrm{SO}(3)$ that acts transitively on S^2 is $\mathrm{SO}(3)$ itself. *Hint: See Exercise 39.21.*

Exercise 41.9. Show that the analogue of Exercise 41.8 does not hold for $\mathrm{SO}(4)$, i.e. there exists a proper Lie subgroup $G \subset \mathrm{SO}(4)$ that acts transitively on the unit sphere $S^3 \subset \mathbb{R}^4$.
Hint: Identify $\mathbb{R}^4$ with $\mathbb{C}^2$ so that $\mathrm{GL}(2, \mathbb{C})$ becomes a subgroup of $\mathrm{GL}(4, \mathbb{R})$, and show that from this perspective, $\mathrm{SO}(4) \cap \mathrm{GL}(2, \mathbb{C}) = \mathrm{U}(2)$.

The next result says that every smooth transitive group action is equivalent to (41.1).

Theorem 41.10. *For any smooth transitive left group action $G \times M \to M$ and any point $p \in M$, there exists a G-equivariant diffeomorphism*

$$G/G_p \to M : [g] \mapsto gp,$$

where the G-action on G/G_p is defined as in (41.1).

Proof. Denote the map in question by $\psi : G/G_p \to M$. It is easy to check that ψ is well defined, injective and G-equivariant, and its smoothness follows from Exercise 40.27 and the fact that the map $G \to M : g \mapsto gp$ is smooth. Transitivity is then equivalent to the condition that ψ is surjective, and it will therefore be a diffeomorphism if and only if its derivative at every point is an isomorphism. Let us first examine whether this is true at a single point, namely at $[e] \in G/G_p$. By Exercise 40.27, the tangent space $T_{[e]}(G/G_p)$ is naturally isomorphic to the quotient vector space $T_eG/T_e(G_p) = \mathfrak{g}/\mathfrak{g}_p$, and under this identification, $T_{[e]}\psi : T_{[e]}(G/G_p) \to T_pM$ becomes

$$\mathfrak{g}/\mathfrak{g}_p \to T_pM : [X] \mapsto X^F(p),$$

which is well defined and injective since by Lemma 40.15, $X^F(p) = 0$ if and only if $X \in \mathfrak{g}_p$. By equivariance and transitivity, the same then holds at every point of G/G_p, meaning that ψ is an immersion, and it will therefore be a diffeomorphism if and only if the dimensions of G/G_p and M are the same. The alternative would be $\dim(G/G_p) < \dim M$, making ψ a *surjective* smooth map from a manifold to another manifold of strictly larger dimension, and there are various ways to see that such objects never exist.[93] The quickest is perhaps to quote Sard's theorem (see e.g. [Mil97]), which states that for a smooth map $f : M \to N$, almost every point $q \in N$ is a regular value of f, meaning that $T_p f : T_p M \to T_q N$ is surjective for every $p \in f^{-1}(q)$. The only way for this to hold when $\dim M < \dim N$ is if almost every $q \in N$ lies outside the image of f, so f cannot be surjective. $\square$

41.3. Grassmann and Stiefel manifolds. We have previously seen two examples of real Grassmann manifolds (or **Grassmannians** for short): one of them is $\mathrm{Gr}_1(\mathbb{R}^{n+1}) := \mathbb{RP}^n$, in its interpretation as the space of all 1-dimensional subspaces of $\mathbb{R}^{n+1}$. Another appeared when we described the sectional curvature of a Riemannian manifold (M, g) as a real-valued function K_S on the set $\mathrm{Gr}_2(TM) := \bigcup_{p\in M} \mathrm{Gr}_2(T_p M)$, where $\mathrm{Gr}_2(T_p M)$ denotes the set of all 2-dimensional subspaces of $T_p M$. More generally, if V is an n-dimensional vector space over $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ and $0 \leqslant k \leqslant n$, we define

$$\mathrm{Gr}_k(V) := \left\{ W \subset V \;\middle|\; W \text{ a linear subspace with } \dim W = k \right\}.$$

Here the word "linear subspace" means real or complex subspaces, depending on the field $\mathbb{F}$; taking $\mathbb{F} = \mathbb{C}$, another example that we've seen before is $\mathbb{CP}^n = \mathrm{Gr}_1(\mathbb{C}^{n+1})$.

It is intuitively clear what kind of topology we'd like to define on $\mathrm{Gr}_k(V)$, though writing it down precisely takes a bit more thought. One good approach is to think in terms of bases: if $v_1, \dots, v_k \in V$ is a basis of $W \in \mathrm{Gr}_k(V)$, then elements $W' \in \mathrm{Gr}_k(V)$ should be considered "close to" W if they are spanned by bases $v_1', \dots, v_k' \in V$ such that v_i' is close to v_i for each $i = 1, \dots, k$. This suggests presenting $\mathrm{Gr}_k(V)$ as a quotient of another object, namely the set of all **k-frames** in V,

$$\mathrm{St}_k(V) := \left\{ (v_1, \dots, v_k) \in V^{\times k} \;\middle|\; v_1, \dots, v_k \text{ are linearly independent} \right\}.$$

As an open subset of the vector space $V^{\times k} = V \times \ldots \times V$, $\mathrm{St}_k(V)$ is clearly a smooth manifold, and it will often be useful to note that it has a natural diffeomorphism with the open subset of $\mathrm{Hom}(\mathbb{F}^k, V)$ consisting of injective linear maps $\mathbb{F}^k \to V$, i.e. we associate to each such map $\phi : \mathbb{F}^k \to V$ the k-frame $(\phi(\mathbf{e}_1), \dots, \phi(\mathbf{e}_k))$, where $\mathbf{e}_1, \dots, \mathbf{e}_k$ is the standard basis of $\mathbb{F}^k$. Now observe that there is a natural surjective map

$$\pi : \mathrm{St}_k(V) \to \mathrm{Gr}_k(V)$$

associating to each k-frame $(v_1, \dots, v_k)$ the k-dimensional subspace that it spans, or equivalently, the image of the corresponding injective linear map $\phi : \mathbb{F}^k \to V$. This is in fact the quotient projection for a smooth right action of the group $\mathrm{GL}(k, \mathbb{F})$ on $\mathrm{St}_k(V)$, defined by composing injective maps $\phi : \mathbb{F}^k \to V$ with invertible transformations $\mathbf{A} : \mathbb{F}^k \to \mathbb{F}^k$,

$$\mathrm{St}_k(V) \times \mathrm{GL}(k, \mathbb{F}) \to \mathrm{St}_k(V) : (\phi, \mathbf{A}) \mapsto \phi\mathbf{A} := \phi \circ \mathbf{A}.$$

It is straightforward to check that this group action is smooth, free and proper, thus it endows

$$\mathrm{Gr}_k(V) \cong \mathrm{St}_k(V) / \mathrm{GL}(k, \mathbb{F})$$

with the structure of a smooth manifold. Its dimension in the real case is $\dim \mathrm{St}_k(V) - \dim \mathrm{GL}(k, \mathbb{F}) = nk - k^2 = k(n-k)$, and in the complex case, one can check that $\mathrm{Gr}_k(V)$ is also a complex manifold, with $k(n-k)$ as its complex dimension.

[93] They all depend in some way on the assumption that manifolds are separable and metrizable, or in topological terms, they satisfy the second countability axiom. Otherwise, one could call $\coprod_{t\in\mathbb{R}} \{t\} \times \mathbb{R}^{n-1}$ a disconnected $(n-1)$-manifold and regard its obvious inclusion into $\mathbb{R}^n$ as a surjective smooth map. But thanks to separability, manifolds are not allowed to have uncountably many connected components.

One important feature of $\mathrm{Gr}_k(V)$ that is not so obvious from the presentation above is that, unlike $\mathrm{St}_k(V)$, $\mathrm{Gr}_k(V)$ is compact. One way to see this is by choosing an inner product $\langle\ ,\ \rangle$ on V and replacing $\mathrm{St}_k(V)$ with the set of **orthonormal k-frames**

$$\mathrm{St}_k^{\mathrm{O}}(V) := \left\{(v_1, \dots, v_k) \in \mathrm{St}_k(V) \;\middle|\; \langle v_i, v_j \rangle = \delta_{ij} \text{ for all } i, j \in \{1, \dots, k\}\right\}.$$

In the case $V = \mathbb{F}^n$, this is traditionally called a **Stiefel manifold** and denoted by $V_k(\mathbb{F}^n)$ or $V_{n,k}$, though I will stick with the more verbose notation $\mathrm{St}_k^{\mathrm{O}}(\mathbb{F}^n)$ for now. It is manifestly compact, since it is a closed subset of the k-fold product of the unit sphere in V. To see that it is also a smooth submanifold of $\mathrm{St}_k(V)$, let us consider the Lie group

$$\mathrm{O}(V) := \left\{A \in \mathrm{GL}(V) \;\middle|\; \langle Av, Aw \rangle = \langle v, w \rangle \text{ for all } v, w \in V\right\},$$

which is isomorphic to $\mathrm{O}(n)$ if $\mathbb{F} = \mathbb{R}$ and $\mathrm{U}(n)$ if $\mathbb{F} = \mathbb{C}$. This group has an obvious smooth left action on $\mathrm{St}_k(V)$ that preserves $\mathrm{St}_k^{\mathrm{O}}(V)$, i.e. an orthogonal transformation $A \in \mathrm{O}(V)$ sends a k-frame $(v_1, \dots, v_k)$ to the k-frame $(Av_1, \dots, Av_k)$, or equivalently, it sends an injective linear map $\phi : \mathbb{F}^k \to V$ to the injective linear map $A \circ \phi : \mathbb{F}^k \to V$. Since every orthonormal k-frame can be completed to an orthonormal basis of V, the action of $\mathrm{O}(V)$ on $\mathrm{St}_k^{\mathrm{O}}(V)$ is also transitive, which is a strong hint that $\mathrm{St}_k^{\mathrm{O}}(V)$ should be not only a smooth manifold, but also a homogeneous space. To make this precise, we need to choose a reference k-frame $(e_1, \dots, e_k) \in \mathrm{St}_k^{\mathrm{O}}(V)$ and examine its stabilizer under the $\mathrm{O}(V)$-action: this will be the set of all orthogonal transformations $V \to V$ that fix the subspace $W \subset V$ spanned by $e_1, \dots, e_k$, thus it is equivalent to the group $\mathrm{O}(W^\perp)$ of orthogonal transformations on the orthogonal complement $W^\perp \subset V$. Theorem 41.10 then suggests considering the smooth map

$$\mathrm{O}(V)/\,\mathrm{O}(W^\perp) \to \mathrm{St}_k(V) : [A] \mapsto (Ae_1, \dots, Ae_k), \tag{41.2}$$

where the homogeneous space $\mathrm{O}(V)/\,\mathrm{O}(W^\perp)$ is defined via the right action of the closed subgroup $\mathrm{O}(W^\perp) := \left\{A \in \mathrm{O}(V) \;\middle|\; A|_W = \mathbb{1}_W\right\}$ on $\mathrm{O}(V)$.

Exercise 41.11. Check that the map (41.2) is an embedding, and its image is $\mathrm{St}_k^{\mathrm{O}}(V)$, thus proving that $\mathrm{St}_k^{\mathrm{O}}(V)$ is a smooth submanifold of $\mathrm{St}_k(V)$ diffeomorphic to $\mathrm{O}(V)/\,\mathrm{O}(W^\perp)$.

To recover the Grassmannian $\mathrm{Gr}_k(V)$ from the Stiefel manifold $\mathrm{St}_k^{\mathrm{O}}(V)$, we observe that the restriction of the natural surjection $\mathrm{St}_k(V) \to \mathrm{Gr}_k(V)$ to $\mathrm{St}_k^{\mathrm{O}}(V)$ is also surjective since every subspace admits an orthonormal basis, and it can also be interpreted as a quotient projection, where the group acting on $\mathrm{St}_k^{\mathrm{O}}(V)$ from the right is now

$$\mathrm{O}(\mathbb{F}^k) = \begin{cases} \mathrm{O}(k) & \text{if } \mathbb{F} = \mathbb{R}, \\ \mathrm{U}(k) & \text{if } \mathbb{F} = \mathbb{C}. \end{cases}$$

The right action of $\mathrm{O}(\mathbb{F}^k)$ on the larger manifold $\mathrm{St}_k(V)$ is manifestly smooth, so its restriction to $\mathrm{St}_k^{\mathrm{O}}(V)$ is also smooth due to the fact that the latter is a smooth submanifold. This action is also free, and trivially proper since $\mathrm{O}(\mathbb{F}^k)$ is compact, so the quotient is a smooth manifold, and we thus obtain a new presentation of $\mathrm{Gr}_k(V)$ as

$$\mathrm{Gr}_k(V) = \mathrm{St}_k^{\mathrm{O}}(V)/\,\mathrm{O}(\mathbb{F}^k) \cong \left(\mathrm{O}(V)/\,\mathrm{O}(W^\perp)\right) \Big/ \,\mathrm{O}(\mathbb{F}^k). \tag{41.3}$$

Two important observations about this presentation: first, since the inclusion $\mathrm{St}_k^{\mathrm{O}}(V) \hookrightarrow \mathrm{St}_k(V)$ is smooth and descends to a well-defined bijection of quotients $\mathrm{St}_k^{\mathrm{O}}(V)/\,\mathrm{O}(\mathbb{F}^k) \to \mathrm{St}_k(V)/\,\mathrm{GL}(k, \mathbb{F})$, Exercise 40.28 implies that this bijection is also smooth—one can check that it is additionally a local diffeomorphism, and therefore a diffeomorphism, proving that our two ways of defining the smooth structure on $\mathrm{Gr}_k(V)$ give the same result. Second, the double quotient on the right hand side of (41.3) involves two right actions, but this picture can be simplified. Indeed, our diffeomorphism of $\mathrm{St}_k^{\mathrm{O}}(V)$ with $\mathrm{O}(V)/\mathrm{O}(W^\perp)$ depended essentially on two choices: a k-dimensional subspace $W \subset V$, and then an orthonormal basis $(e_1, \dots, e_k)$ of this subspace, or equivalently, an orthogonal isomorphism $\mathbb{F}^k \to W$. The latter gives rise to an identification of the group $\mathrm{O}(\mathbb{F}^k)$ with the subgroup $\mathrm{O}(W) \subset \mathrm{O}(V)$ consisting of orthogonal transformations on V that act trivially on $W^\perp$, and this subgroup commutes with $\mathrm{O}(W^\perp)$. The result is that instead of a double quotient, it is equivalent to divide $\mathrm{O}(V)$ by the product of these two subgroups, giving an explicit presentation of $\mathrm{Gr}_k(V)$ as a homogeneous space,

$$\mathrm{Gr}_k(V) \cong \mathrm{O}(V)/\left(\mathrm{O}(W) \times \mathrm{O}(W^\perp)\right). \tag{41.4}$$

The subgroup in this denominator is equivalently $H := \{A \in \mathrm{O}(V) \mid A(W) = W\}$, and an explicit bijection can be defined by sending the coset $AH \in \mathrm{O}(V)/H$ to $A(W) \in \mathrm{Gr}_k(V)$. In particular, this gives

$$\mathrm{Gr}_k(\mathbb{R}^n) \cong \mathrm{O}(n)/(\mathrm{O}(k) \times \mathrm{O}(n-k)), \qquad \mathrm{Gr}_k(\mathbb{C}^n) \cong \mathrm{U}(n)/(\mathrm{U}(k) \times \mathrm{U}(n-k)), \tag{41.5}$$

where the subgroup in each case consists of all matrices that admit block decompositions $\begin{pmatrix} \mathbf{A} & 0 \\ 0 & \mathbf{B} \end{pmatrix}$ determined by a k-by-k block $\mathbf{A}$ and an $(n-k)$-by-$(n-k)$ block $\mathbf{B}$.

Exercise 41.12. As a sanity check, compute the dimensions of the two homogeneous spaces in (41.5), and make sure they match our previous computation of $\dim \mathrm{Gr}_k(V)$.

A minor enhancement of $\mathrm{Gr}_k(V)$ is also worth mentioning: when V is a real n-dimensional vector space, we denote by

$$\widetilde{\mathrm{Gr}}_k(V) \xrightarrow{\Pi} \mathrm{Gr}_k(V)$$

the natural two-to-one cover in which each element of $\widetilde{\mathrm{Gr}}_k(V)$ is a subspace $W \in \mathrm{Gr}_k(V)$ endowed with the extra data of an orientation, and the projection Π is defined by forgetting the orientation. The projection $\pi : \mathrm{St}_k^{\mathrm{O}}(V) \to \mathrm{Gr}_k(V)$ factors through this and another projection $\widetilde{\pi} : \mathrm{St}_k^{\mathrm{O}}(V) \to \widetilde{\mathrm{Gr}}_k(V)$ since each k-frame naturally determines an orientation of the subspace that it spans. I will leave it as an exercise for the reader to show that when $k < n$, the natural smooth structure on $\widetilde{\mathrm{Gr}}_k(V)$ is obtained by replacing each of the Lie groups in (41.4) with their identity components, that is,

$$\widetilde{\mathrm{Gr}}_k(V) \cong \mathrm{SO}(V)/\left(\mathrm{SO}(W) \times \mathrm{SO}(W^\perp)\right),$$

and the inclusion map $\mathrm{SO}(V) \hookrightarrow \mathrm{O}(V)$ then descends to the quotients to define the double cover $\Pi : \widetilde{\mathrm{Gr}}_k(V) \to \mathrm{Gr}_k(V)$ as a smooth map. (This correspondence fails in the case $k = n$ because the action of $\mathrm{SO}(V)$ on $\mathrm{St}_n^{\mathrm{O}}(V)$ is not transitive, but this detail is unimportant since $\mathrm{Gr}_n(V)$ and $\widetilde{\mathrm{Gr}}_n(V)$ are not very interesting spaces when $n = \dim V$.)

Exercise 41.13. Prove that for any n-dimensional vector space V over $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ and $0 \leqslant k \leqslant n$,

$$E := \left\{(W, v) \in \mathrm{Gr}_k(V) \times V \;\middle|\; v \in W\right\}$$

defines a smooth subbundle of the trivial vector bundle $\mathrm{Gr}_k(V) \times V \to \mathrm{Gr}_k(V)$. (It is called the **tautological vector bundle** over $\mathrm{Gr}_k(V)$.)

Exercise 41.14. On a real vector space V of dimension $2n$, let

$$\mathcal{J}(V) := \left\{J \in \mathrm{End}(V) \;\middle|\; J^2 = -\mathbb{1}\right\}.$$

The elements $J \in \mathcal{J}(V)$ are called **complex structures** on V, as each one can be used to endow V with the structure of an n-dimensional complex vector space on which scalar multiplication is defined by $(a+ib)v := av + bJv$. Prove that $\mathcal{J}(V)$ is a smooth noncompact submanifold of $\mathrm{End}(V)$ with dimension $2n^2$, and that it has exactly two connected components.

Hint: Find a smooth action of $\mathrm{GL}(V)$ *on* $\mathrm{End}(V)$ *that preserves* $\mathcal{J}(V)$ *and has stabilizer at some point* $J_0 \in \mathcal{J}(V)$ *isomorphic to* $\mathrm{GL}(n, \mathbb{C})$.

42. Fiber bundles

42.1. Examples and main definition. For the next several lectures, we will be considering various flavors of objects that clearly deserve to be called "bundles", even though their "fibers" are not always vector spaces. We will more generally allow fibers of bundles to be smooth manifolds, typically with some additional structure, satisfying the condition that each fiber admits a smooth family of diffeomorphisms (preserving whatever additional structure it has) to all the nearby fibers. Here are some familiar examples with which to motivate the main definition.

EXAMPLE 42.1. For a smooth vector bundle $E \to M$ with a positive bundle metric $\langle\ ,\ \rangle$, the associated **unit sphere bundle** $\pi : SE \to M$ is defined by restricting the projection $E \to M$ to the subset

$$SE := \left\{ v \in E \;\middle|\; \langle v, v\rangle = 1 \right\} \subset E.$$

If E is a real bundle of rank m, then the fibers $SE_p := \pi^{-1}(p) \subset SE$ are all diffeomorphic to the sphere S^{m-1}, and every choice of orthonormal frame for E over an open subset $\mathcal{U} \subset M$ determines a smooth family of diffeomorphisms $SE_p \to S^{m-1}$ for every $p \in \mathcal{U}$. Unit sphere bundles have the nice property that whenever M is compact, SE is also a *compact* manifold; by contrast, the total space of a vector bundle of positive rank is never compact. We made use of the compactness of $S(TM)$ once last semester, when we proved that all compact Riemannian manifolds are geodesically complete (see §23.3).

EXAMPLE 42.2. A close relative of the unit sphere bundle $\pi : SE \to M$ in Example 42.1 is the **unit disk bundle**,

$$\mathbb{D}E := \left\{ v \in E \;\middle|\; \langle v, v\rangle \leqslant 1 \right\} \subset E,$$

which is a smooth manifold with boundary $\partial(\mathbb{D}E) = SE$ whenever M is a manifold without boundary,[94] and it is compact if M is compact. As in Example 42.1, every orthonormal frame for E over a subset $\mathcal{U} \subset M$ defines a smooth family of diffeomorphisms of the fibers $\mathbb{D}E_p$ to the disk $\mathbb{D}^m$ for $p \in \mathcal{U}$.

EXAMPLE 42.3. If $E \to M$ is a vector bundle of rank m over $\mathbb{F}$, then for each $k = 0, \ldots, m$ there are corresponding bundles of Grassmannian and Stiefel manifolds

$$\mathrm{Gr}_k(E) := \bigcup_{p \in M} \mathrm{Gr}_k(E_p), \qquad \mathrm{St}_k(E) := \bigcup_{p \in M} \mathrm{St}_k(E_p).$$

If E is endowed with a positive bundle metric, then we also have the bundle of *orthonormal* k-frames

$$\mathrm{St}_k^{\mathrm{O}}(E) := \bigcup_{p \in M} \mathrm{St}_k^{\mathrm{O}}(E_p),$$

and if $\mathbb{F} = \mathbb{R}$, the bundle of *oriented* k-planes

$$\widetilde{\mathrm{Gr}}_k(E) := \bigcup_{p \in M} \widetilde{\mathrm{Gr}}_k(E_p).$$

Having seen in the previous subsection that all of these objects have fibers that are smooth manifolds, it should be easy to convince yourself that suitable local frames for E always give rise to smooth families of diffeomorphisms identifying fibers with $\mathrm{Gr}_k(\mathbb{F}^m)$, $\mathrm{St}_k(\mathbb{F}^m)$, $\mathrm{St}_k^{\mathrm{O}}(\mathbb{F}^m)$ or $\widetilde{\mathrm{Gr}}_k(\mathbb{R}^m)$ respectively.

[94] If $\partial M \neq \varnothing$, then $\mathbb{D}E$ can be regarded as a smooth manifold "with boundary and corners", a notion that does not quite fit into the definitions in this course, so we are avoiding talking about it.

EXAMPLE 42.4. The following special cases of Example 42.3 deserve more attention: assuming $\mathrm{rank}(E) = m$, we call

$$FE := \mathrm{St}_m(E)$$

the **frame bundle** of E, and if E is a real bundle with a positive bundle metric, there is similarly the **orthonormal frame bundle**

$$F^{\mathrm{O}}E := \mathrm{St}_m^{\mathrm{O}}(E).$$

The fibers of a frame bundle are bases of the fibers of the underlying vector bundle, so in particular, each fiber FE_p of FE is diffeomorphic to the group $\mathrm{GL}(m, \mathbb{F})$, though there is generally no canonical choice of diffeomorphism, nor a natural group structure on FE_p that can be defined without additional choices. What we have instead is a natural smooth right action of $\mathrm{GL}(m, \mathbb{F})$ on FE_p, defined by identifying frames with vector space isomorphisms $\mathbb{F}^m \to E_p$ and composing those isomorphisms with invertible linear transformations $\mathbb{F}^m \to \mathbb{F}^m$. It is easy to check that this group action is free and transitive, so that any map $\mathrm{GL}(m, \mathbb{F}) \to FE_p$ defined by letting the group act on a chosen point in FE_p is a diffeomorphism. Similarly, the fibers $F^{\mathrm{O}}E_p$ of the orthonormal frame bundle are compact manifolds with a natural free and transitive right action of the group

$$\mathrm{O}(\mathbb{F}^m) := \begin{cases} \mathrm{O}(m) & \text{if } \mathbb{F} = \mathbb{R}, \\ \mathrm{U}(m) & \text{if } \mathbb{F} = \mathbb{C}, \end{cases}$$

implying that they are all (non-canonically) diffeomorphic to $\mathrm{O}(\mathbb{F}^m)$.

EXAMPLE 42.5. As an addendum to the previous example, observe that if we endow the tautological k-plane bundle $E \to \mathrm{Gr}_k(V)$ from Exercise 41.13 with the natural bundle metric induced by any choice of inner product on V, then the natural projection $\mathrm{St}_k^{\mathrm{O}}(V) \to \mathrm{Gr}_k(V)$ is precisely the orthonormal frame bundle of E.

All of the above examples are special cases of the following object.

DEFINITION 42.6. A **smooth fiber bundle** $\pi : E \to M$ consists of a pair of smooth manifolds E (the **total space**) and M (the **base**), and a surjective smooth map π that is locally trivializable in the following sense. There exists a smooth manifold F, called the **standard fiber** of E, such that every point $p \in M$ admits a neighborhood $\mathcal{U} \subset M$ with a diffeomorphism

$$\Phi : E|_{\mathcal{U}} := \pi^{-1}(\mathcal{U}) \to \mathcal{U} \times F$$

sending the **fiber** $E_q := \pi^{-1}(q)$ over each point $q \in \mathcal{U}$ to $\{q\} \times F$. The pair $(\mathcal{U}, \Phi)$ is in this case called a **local trivialization** of the bundle $\pi : E \to M$. A smooth **section** of $\pi : E \to M$ is a smooth map $s : M \to E$ such that $\pi \circ s = \mathrm{Id}_M$, and the set of all smooth sections will be denoted by $\Gamma(E)$.

There is of course also a topological analogue of this definition, in which every instance of the word "manifold" is replaced by "topological space" and all maps are required to be continuous instead of smooth. Geometers sometimes use the word **fibration** as a synonym for "fiber bundle", which is slightly unfortunate because "fibration" means something more general in topology, but this rarely causes any actual confusion.

The notion of a **pullback bundle** has a relatively straightforward definition in the world of smooth fiber bundles. It comes from the observation that if $\pi : E \to M$ and $f : N \to M$ are smooth maps and π is a submersion, then by a bit of basic transversality theory (see Exercise 42.7 below), the set

$$f^*E := \left\{ (p, x) \in N \times E \;\middle|\; \pi(x) = f(p) \right\}$$

is a smooth submanifold of $N \times E$, and the smooth map

$$N \times E \supset f^*E \xrightarrow{f^*\pi} N : (p, x) \mapsto p$$

is then also a submersion. If $\pi : E \to M$ is a fiber bundle, then one can define from any local trivialization $\Phi : E|_{\mathcal{U}} \to M \times F$ a smooth map

$$f^*\Phi : (f^*\pi)^{-1}(\mathcal{U}) \to f^{-1}(\mathcal{U}) \times F : (p, x) \mapsto (p, \mathrm{pr}_2 \circ \Phi(x)),$$

where $\mathrm{pr}_2 : \mathcal{U} \times F \to F$ denotes the obvious projection, and $(f^{-1}(\mathcal{U}), f^*\Phi)$ can then be interpreted as a local trivialization of $f^*\pi : f^*E \to N$, making the latter a smooth fiber bundle with fibers $(f^*E)_p = E_{f(p)}$ for every $p \in N$.

Exercise 42.7. Here's a subject that probably should have been covered in last semester's course but wasn't: given a smooth map $f : N \to M$ and a smooth submanifold $Q \subset M$, we say that f is **transverse** to Q and write "$f \pitchfork Q$" if for every $p \in N$ with $q := f(p) \in Q$, we have $(\operatorname{im} T_p f) + T_q Q = T_q M$.

(a) Prove that if $f \pitchfork Q$, then $\Sigma := f^{-1}(Q) \subset N$ is a smooth submanifold with $T_p\Sigma = \left\{ X \in T_p N \mid T_p f(X) \in T_q Q \right\}$ for every $p \in \Sigma$ and $q := f(p)$. What is its dimension?
Hint: This generalizes the version of the implicit function theorem that we have often used to study level sets of smooth maps at regular values, but it also follows from that version of the theorem if you work in suitable coordinates near Q.

(b) Suppose $f : N \to M$ and $\pi : E \to M$ are two smooth maps such that π is a submersion. Prove that the map

$$f \times \pi : N \times E \to M \times M : (p, x) \mapsto (f(p), \pi(x))$$

is then transverse to the so-called **diagonal** submanifold $\Delta := \{(p, p) \mid p \in M\} \subset M \times M$, and thus $\Sigma := (f \times \pi)^{-1}(\Delta)$ is a submanifold of $N \times E$. Prove moreover that the map $\Sigma \to N : (p, x) \mapsto p$ is a submersion.

Let's throw in a word of caution about sections of fiber bundles: since the fibers are no longer required to be vector spaces, $\Gamma(E)$ is at this stage only a set rather than a vector space, and we will see that in many cases of interest, $\Gamma(E) = \varnothing$. This is something that *never* happens with vector bundles, since they all at least admit the zero-section, and one can always use partitions of unity to construct many nontrivial sections. But on general fiber bundles, there is typically no distinguished element of each fiber analogous to the zero element of a vector space, and partitions of unity are of little use when fibers have no linear structure with which to interpolate. In fact, if $E \to M$ is a vector bundle and $FE \to M$ is its frame bundle as defined in Example 42.4, then a section of FE is the same thing as a *global* frame for E, which means $\Gamma(FE) \neq \varnothing$ if and only if the bundle E is trivial. This observation is one of the main advantages of the notion of a frame bundle: it reduces many geometric and topological questions about E to questions about sections of a suitable frame bundle related to E.

42.2. Structure groups. The definition of a fiber bundle in the previous section was a bit too general to be truly useful in typical applications. Most of the actual examples one encounters have more structure than was allowed for in that definition, e.g. the fibers of the unit sphere bundle in Example 42.1 carry natural Riemannian metrics, and fibers of the frame bundles in Example 42.4 come with free and transitive group actions. We shall now introduce a very general framework for encoding these kinds of structure.

For a fiber bundle $\pi : E \to M$ with standard fiber F, we can again use the term **bundle atlas** to mean a collection of local trivializations $\{(\mathcal{U}_\alpha, \Phi_\alpha)\}_{\alpha \in I}$ such that $M = \bigcup_{\alpha \in I} \mathcal{U}_\alpha$. At this stage there is no need to define a notion of "smooth compatibility" since E, M and F were all assumed already to be endowed with smooth structures, and Definition 42.6 included the condition that local trivializations are smooth maps. (We did not originally include this condition in the definition of a smooth vector bundle, but we could have done so without changing anything.) It is instructive nonetheless to consider what the notion of a "transition function" might mean in this context. If $(\mathcal{U}_\alpha, \Phi_\alpha)$ and $(\mathcal{U}_\beta, \Phi_\beta)$ are two local trivializations with a nonempty overlap $\mathcal{U}_\alpha \cap \mathcal{U}_\beta \neq \varnothing$, one obtains a diffeomorphism of the form

$$(\mathcal{U}_\alpha \cap \mathcal{U}_\beta) \times F \xrightarrow{\Phi_\beta \circ \Phi_\alpha^{-1}} (\mathcal{U}_\alpha \cap \mathcal{U}_\beta) \times F : (p, x) \mapsto (p, f(p, x)),$$

where for each $p \in \mathcal{U}_\alpha \cap \mathcal{U}_\beta$, $f(p, \cdot) : F \to F$ is a diffeomorphism, thus defining a map

$$g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{Diff}(F), \qquad g_{\beta\alpha}(p) := f(p, \cdot).$$

If we were dealing with purely topological fiber bundles, in which E, M and F are topological spaces not equipped with smooth structures, then $g_{\beta\alpha}$ would instead take values in the group $\mathrm{Homeo}(F)$, and under some mild topological conditions (cf. §37.1), the latter could be understood as a topological group such that the transition functions are continuous. For a smooth bundle, the map $g_{\beta\alpha}$ will certainly be continuous with respect to the C^∞_{loc}-topology on $\mathrm{Diff}(F)$, but we cannot meagingfully require it to be "smooth" since $\mathrm{Diff}(F)$ is not a Lie group in any sense that would be useful for this discussion. In this regard, the distinction between smooth and topological fiber bundles cannot be expressed purely in terms of the transition functions $g_{\beta\alpha}$, and is thus not precisely analogous to the case of vector bundles. For *most* smooth fiber bundles that we will be interested in, however, there is a reasonable way to define the notion of smooth transition functions: we can require them to take values in a finite-dimensional Lie group G that acts smoothly on the standard fiber F.

Let us consider a concrete example: suppose $\pi : E \to M$ is a smooth real vector bundle of rank m, so the standard fiber in this case is $F = \mathbb{R}^m$, and the diffeomorphisms $g_{\beta\alpha}(p) \in \mathrm{Diff}(\mathbb{R}^m)$ for $p \in \mathcal{U}_\alpha \cap \mathcal{U}_\beta$ are linear maps. We can then view $g_{\beta\alpha}$ as taking values in the Lie group $\mathrm{GL}(m, \mathbb{R})$, and each matrix in $\mathrm{GL}(m, \mathbb{R})$ defines a diffeomorphism $\mathbb{R}^m \to \mathbb{R}^m$ via the canonical smooth left action of $\mathrm{GL}(m, \mathbb{R})$ on $\mathbb{R}^m$. If we now suppose additionally that $\pi : E \to M$ is oriented and endowed with a positive bundle metric, then it also admits a bundle atlas $\{(\mathcal{U}_\alpha, \Phi_\alpha)\}_{\alpha \in I}$ consisting only of trivializations that correspond to positively-oriented orthonormal frames, and the transition functions $g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{GL}(m, \mathbb{R})$ relating any two such trivializations will take values in the Lie subgroup $\mathrm{SO}(m) \subset \mathrm{GL}(m, \mathbb{R})$, which again acts on the standard fiber $\mathbb{R}^m$ in a canonical way. Since each transition function $g_{\beta\alpha}$ is uniquely determined by the two trivializations $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{R}^m$ and $\Phi_\beta : E|_{\mathcal{U}_\beta} \to \mathcal{U}_\beta \times \mathbb{R}^m$, it is straightforward to check that the entire collection of transition functions $\{g_{\beta\alpha}\}_{(\alpha,\beta) \in I \times I}$ automatically satisfies the relations

$$g_{\alpha\alpha} = \mathbb{1} \quad \text{on } \mathcal{U}_\alpha, \qquad \text{and} \qquad g_{\alpha\beta} g_{\beta\gamma} = g_{\alpha\gamma} \quad \text{on } \mathcal{U}_\alpha \cap \mathcal{U}_\beta \cap \mathcal{U}_\gamma \tag{42.1}$$

for all $\alpha, \beta, \gamma \in I$. One subtlety needs to be mentioned here: the reason we know these relations are satisfied is that each orientation-preserving orthogonal transformation $\mathbb{R}^m \to \mathbb{R}^m$ is represented by one and *only one* matrix in $\mathrm{SO}(m)$, so e.g. we conclude from the fact that $g_{\alpha\beta} g_{\beta\gamma} g_{\alpha\gamma}^{-1}$ acts trivially on $\mathbb{R}^m$ that its value is everywhere $\mathbb{1} \in \mathrm{SO}(m)$. But one sometimes also encounters group actions that do not have this property. We say that a smooth left group action $G \times M \to M$ is **effective** if the resulting group homomorphism $G \to \mathrm{Diff}(M)$ is injective, i.e. the only $g \in G$ satisfying $gp = p$ for all $p \in M$ is $g = e$. The canonical action of any Lie subgroup $G \subset \mathrm{GL}(m, \mathbb{F}^m)$ on $\mathbb{F}^m$ is effective, but for instance, one can use the double cover $\mathrm{SU}(2) \to \mathrm{SO}(3)$ discussed in §39.2 to define a non-effective action of $\mathrm{SU}(2)$ on $\mathbb{R}^3$. In such cases, the relations (42.1) will not be satisfied automatically, but we will find that it is useful to require them explicitly in our definitions. This detail will be especially relevant when we discuss spin structures.

Taking the example a step further, suppose $\pi' : E' \to M$ is a second oriented real vector bundle of rank m with a positive bundle metric, and $\{(\mathcal{U}_\alpha, \Phi_\alpha)\}_{\alpha \in I'}$ is a bundle atlas for E' corresponding to positively-oriented orthonormal frames, thus giving rise to $\mathrm{SO}(m)$-valued transition functions $\{g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{SO}(m)\}_{(\alpha,\beta) \in I' \times I'}$. The preferred class of smooth linear bundle maps $\Psi : E \to E'$ then consists of smooth maps Ψ that restrict to the fibers E_p as orientation-preserving orthogonal transformations $E_p \to E'_p$ for all $p \in M$. At the level of local trivializations, such a map Ψ admits the following characterization: for each pair of local trivializations $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{R}^m$ and $\Phi_\beta : E'|_{\mathcal{U}_\beta} \to \mathcal{U}_\beta \times \mathbb{R}^m$ from the given bundle atlases, there exists a unique smooth function $h_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{SO}(m)$ such that

$$\Phi_\beta \circ \Psi \circ \Phi_\alpha^{-1}(p, v) = (p, h_{\beta\alpha}(p)v) \qquad \text{for all } p \in \mathcal{U}_\alpha \cap \mathcal{U}_\beta \text{ and } v \in \mathbb{R}^m,$$

and the resulting collection $\{h_{\beta\alpha}\}_{(\alpha,\beta)\in I\times I'}$ of $\mathrm{SO}(m)$-valued functions then automatically satisfies

$$h_{\alpha\beta} g_{\beta\gamma} = h_{\alpha\gamma} \quad \text{on } \mathcal{U}_\alpha \cap \mathcal{U}_\beta \cap \mathcal{U}_\gamma, \qquad g_{\delta\alpha} h_{\alpha\beta} = h_{\delta\beta} \quad \text{on } \mathcal{U}_\delta \cap \mathcal{U}_\alpha \cap \mathcal{U}_\beta, \tag{42.2}$$

for all $\alpha, \delta \in I'$ and $\beta, \gamma \in I$. Again, these relations are automatic due to the fact that each of the functions $h_{\alpha\beta} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{SO}(m)$ is uniquely determined by the two corresponding trivializations $(\mathcal{U}_\alpha, \Phi_\alpha)$ and $(\mathcal{U}_\beta, \Phi_\beta)$, which is true because $\mathrm{SO}(m)$ acts *effectively* on $\mathbb{R}^m$. If we were dealing with a non-effective group action, the relations (42.2) would not be guaranteed unless they are imposed as an extra condition on isomorphisms between fiber bundles. This is what we will do in the definitions below.

Definition 42.8. Suppose M is a smooth manifold and G is a Lie group. A **system of G-valued transition functions** on M consists of the data $\mathcal{T} = (\{\mathcal{U}_\alpha\}_{\alpha\in I}, \{g_{\beta\alpha}\}_{(\alpha,\beta)\in I\times I})$, where $\{\mathcal{U}_\alpha\}_{\alpha\in I}$ is an open covering of M and $g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to G$ are smooth functions that satisfy

$$g_{\alpha\alpha} = e \quad \text{on } \mathcal{U}_\alpha, \qquad \text{and} \qquad g_{\alpha\beta} g_{\beta\gamma} = g_{\alpha\gamma} \quad \text{on } \mathcal{U}_\alpha \cap \mathcal{U}_\beta \cap \mathcal{U}_\gamma \tag{42.3}$$

for all $\alpha, \beta, \gamma \in I$. If $\mathcal{T}^j := (\{\mathcal{U}_\alpha\}_{\alpha\in I_j}, \{g_{\beta\alpha}\}_{(\alpha,\beta)\in I_j\times I_j})$ for $j = 1, 2$ are two systems of G-valued transition functions on M, then a **morphism** from $\mathcal{T}^1$ to $\mathcal{T}^2$ is a collection $\{h_{\beta\alpha} : \mathcal{U}_\beta \cap \mathcal{U}_\alpha \to G\}_{(\alpha,\beta)\in I_1\times I_2}$ of smooth functions that satisfy

$$h_{\alpha\beta} g_{\beta\gamma} = h_{\alpha\gamma} \quad \text{on } \mathcal{U}_\alpha \cap \mathcal{U}_\beta \cap \mathcal{U}_\gamma, \qquad g_{\delta\alpha} h_{\alpha\beta} = h_{\delta\beta} \quad \text{on } \mathcal{U}_\delta \cap \mathcal{U}_\alpha \cap \mathcal{U}_\beta, \tag{42.4}$$

for all $\alpha, \delta \in I_2$ and $\beta, \gamma \in I_1$.

The second relation in (42.3) is also often written in the form

$$g_{\alpha\beta} g_{\beta\gamma} g_{\gamma\alpha} = e,$$

which follows since $g_{\alpha\gamma} g_{\gamma\alpha} = g_{\alpha\alpha} = e$, implying $g_{\gamma\alpha} = g_{\alpha\gamma}^{-1}$. It is known as the **cocycle condition**. (For the context of this terminology, see Remark 32.6.)

Definition 42.9. Suppose $\pi : E \to M$ is a smooth fiber bundle, G is a Lie group and F is a manifold. A **G-bundle atlas** for $\pi : E \to M$ with standard fiber F is a tuple $\mathcal{A} = (\mathbf{\Phi}, \mathcal{T}, \rho)$ consisting of a bundle atlas $\mathbf{\Phi} = \{\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times F\}_{\alpha\in I}$, a system of G-valued transition functions $\mathcal{T} = (\{\mathcal{U}_\alpha\}_{\alpha\in I}, \{g_{\beta\alpha}\}_{(\alpha,\beta)\in I\times I})$, and a smooth left action $\rho : G \times F \to F : (g, x) \mapsto gx$ such that for every $\alpha, \beta \in I$,

$$\Phi_\beta \circ \Phi_\alpha^{-1}(p, x) = (p, g_{\beta\alpha}(p)x) \qquad \text{for all } p \in \mathcal{U}_\alpha \cap \mathcal{U}_\beta \text{ and } x \in F.$$

Suppose moreover that for $j = 1, 2$, $\pi^j : E^j \to M$ denote two smooth fiber bundles over M, both equipped with G-bundle atlases $\mathcal{A}^j = (\mathbf{\Phi}^j, \mathcal{T}^j, \rho)$ in which the standard fiber F and group action $\rho : G \times F \to F$ are identical. A smooth map $\Psi : E^1 \to E^2$ will then be called a **G-bundle isomorphism** from $(E^1, \mathcal{A}^1)$ to $(E^2, \mathcal{A}^2)$ if there exists a morphism $\{h_{\beta\alpha}\}_{(\alpha,\beta)\in I_1\times I_2}$ from $\mathcal{T}^1$ to $\mathcal{T}^2$ such that for every $\alpha \in I_1$ and $\beta \in I_2$,

$$\Phi_\beta \circ \Psi \circ \Phi_\alpha^{-1}(p, x) = (p, h_{\beta\alpha}(p)x) \qquad \text{for all } p \in \mathcal{U}_\alpha \cap \mathcal{U}_\beta \text{ and } x \in F.$$

Note that the map $\Psi : E^1 \to E^2$ in Definition 42.9 is necessarily invertible, so e.g. in the case of vector bundles, the definition does not account for smooth linear bundle maps that are not bundle isomorphisms. Non-invertible linear bundle maps are in any case not very nice objects—their kernels and images for instance can have varying dimension from point to point, so that they are not always vector bundles. We will not attempt to fit them into the general framework we are developing for fiber bundles.

LEMMA 42.10. *The inverse of any G-bundle isomorphism $\Psi : (E^1, \mathcal{A}^1) \to (E^2, \mathcal{A}^2)$ is a G-bundle isomorphism $\Psi^{-1} : (E^2, \mathcal{A}^2) \to (E^1, \mathcal{A}^1)$, and the composition of two G-bundle isomorphisms $\Psi : (E^1, \mathcal{A}^1) \to (E^2, \mathcal{A}^2)$ and $\Psi' : (E^2, \mathcal{A}^2) \to (E^3, \mathcal{A}^3)$ is a G-bundle isomorphism $\Psi' \circ \Psi : (E^1, \mathcal{A}^1) \to (E^3, \mathcal{A}^3)$.*

PROOF. The statement about the inverse is an easy exercise: one need only define $h_{\alpha\beta}(p) := h_{\beta\alpha}(p)^{-1}$ for $(\alpha, \beta) \in I_1 \times I_2$ and check that the required conditions are satisfied. For the statement about compositions, suppose $\alpha \in I_1$, $\gamma \in I_3$ and $p \in \mathcal{U}_\alpha \cap \mathcal{U}_\gamma$. Since $\{\mathcal{U}_\beta\}_{\beta \in I_2}$ is an open cover of M, we can then find some $\beta \in I_2$ such that a neighborhood of p is contained in $\mathcal{U}_\alpha \cap \mathcal{U}_\beta \cap \mathcal{U}_\gamma$, and on this neighborhood we define

$$h_{\gamma\alpha} := h_{\gamma\beta} h_{\beta\alpha},$$

where $h_{\gamma\beta}$ and $h_{\beta\alpha}$ come from morphisms $\mathcal{T}^2 \to \mathcal{T}^3$ and $\mathcal{T}^1 \to \mathcal{T}^2$ respectively. We claim that the function $h_{\gamma\alpha}$ is then independent of the choice of $\beta \in I_2$ with $p \in \mathcal{U}_\beta$. Indeed, if $\delta \in I_2$ is another such choice, we have

$$h_{\gamma\delta} h_{\delta\alpha} = h_{\gamma\beta} g_{\beta\delta} g_{\delta\beta} h_{\beta\alpha} = h_{\gamma\beta} g_{\beta\beta} h_{\beta\alpha} = h_{\gamma\beta} h_{\beta\alpha}.$$

It follows that $h_{\gamma\alpha}$ can be defined in this manner near every point $p \in \mathcal{U}_\alpha \cap \mathcal{U}_\gamma$ so as to give a well-defined and smooth function $h_{\gamma\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\gamma \to G$. It is straightforward to check that the resulting collection of functions for all $(\alpha, \gamma) \in I_1 \times I_3$ makes $\Psi' \circ \Psi$ into a G-bundle isomorphism. $\square$

Lemma 42.10 makes the equivalence relation in the next definition well defined.

DEFINITION 42.11. Given a smooth fiber bundle $\pi : E \to M$ and Lie group G, a **G-structure** on $\pi : E \to M$ is an equivalence class of G-bundle atlases, where $\mathcal{A}^1$ and $\mathcal{A}^2$ are considered equivalent if and only if the identity map $E \to E$ is a G-bundle isomorphism $(E, \mathcal{A}^1) \to (E, \mathcal{A}^2)$. A fiber bundle endowed with a G-structure is sometimes called a **G-bundle**, or a "fiber bundle with **structure group** G". When a G-structure is given, we will refer to the bundle atlases in its equivalence class as the **G-compatible bundle atlases**, and the trivializations in these bundle atlases as **G-compatible trivializations**.

EXAMPLE 42.12. For any pair of manifolds M, F and any Lie group G, there is a **trivial G-bundle** over M with fiber F, defined as the product $M \times F$ with the obvious projection to M. The G-structure on this bundle is defined via a single global trivialization, thus requiring only one transition function, which is a constant function with value $e \in G$.

DEFINITION 42.13. A G-bundle $\pi : E \to M$ with standard fiber F is **trivial** if it admits a G-bundle isomorphism to the product bundle in Example 42.12.

EXERCISE 42.14. Show that a G-bundle $\pi : E \to M$ is trivial if and only if it admits a G-compatible bundle atlas that contains only one trivialization.

When we introduced smooth structures on vector bundles in Lecture 32, we did not talk about equivalence classes of bundle atlases, though we could have done, i.e. by defining two bundle atlases $\mathcal{A}_1$ and $\mathcal{A}_2$ to be equivalent if and only if every trivialization in $\mathcal{A}_1$ is smoothly compatible with every trivialization in $\mathcal{A}_2$. The reason we did not bother to make that definition at the time is that every equivalence class of smooth vector bundle atlases has a unique *maximal* representative, namely the union of all the atlases in the equivalence class, thus it was quicker but equally valid to say that a smooth structure is the same thing as a maximal smooth bundle atlas. Taking a maximal union of atlases is straightforward in that situation due to the fact that the action of $\mathrm{GL}(m, \mathbb{F})$ on $\mathbb{F}^m$ is effective, so that transition functions are uniquely determined by trivializations. It is less straightforward for general G-bundle atlases, but the next exercise shows that it can be done.

Exercise 42.15. For a smooth fiber bundle $\pi : E \to M$ and Lie group G, one can define a partial order $\prec$ on the set of all G-bundle atlases by inclusion, meaning we write $\mathcal{A}^1 \prec \mathcal{A}^2$ if for $j = 1, 2$, $\mathcal{A}^j$ has the form $(\mathbf{\Phi}^j, \mathcal{T}^j, \rho)$ where $\mathbf{\Phi}^j = \left\{\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times F^j\right\}_{\alpha \in I_j}$ and $\mathcal{T}^j = (\{\mathcal{U}_\alpha\}_{\alpha \in I_j}, \{g_{\alpha\beta}\}_{(\alpha,\beta)\times I_j \times I_j})$ such that I_1 is a subset of I_2. This condition implies in particular that every trivialization and every transition function in $\mathcal{A}^1$ is also in $\mathcal{A}^2$. Prove:

(a) If $\mathcal{A}^1 \prec \mathcal{A}^2$, then $\mathcal{A}^1$ and $\mathcal{A}^2$ are equivalent.
(b) If $\mathcal{A}^1$ and $\mathcal{A}^2$ are equivalent, then there exists another G-bundle atlas $\mathcal{A}^3$ such that $\mathcal{A}^1 \prec \mathcal{A}^3$ and $\mathcal{A}^1 \prec \mathcal{A}^3$.
(c) For any collection of G-bundle atlases $\{\mathcal{A}^j\}_{j \in J}$ that is totally ordered, meaning either $\mathcal{A}^j \prec \mathcal{A}^k$ or $\mathcal{A}^k \prec \mathcal{A}^j$ holds for every $j, k \in J$, there exists a G-bundle atlas $\mathcal{A}$ such that $\mathcal{A}^j \prec \mathcal{A}$ for every $j \in J$.

Readers familiar with Zorn's lemma (see e.g. [**Jän05**, **Kel75**]) will recognize that we have just established its hypotheses. The result is that every equivalence class of G-bundle atlases has a unique maximal representative.

With the new definitions in hand, we next revisit a few familiar examples and introduce some new ones. To start with, we observe that in the language of G-bundles, a vector bundle is nothing other than a fiber bundle whose standard fiber is a vector space and whose structure group acts on it linearly:

Example 42.16. If $\pi : E \to M$ has a G-bundle atlas with $G = \mathrm{GL}(m, \mathbb{F})$ acting on standard fiber $\mathbb{F}^m$ via matrix-vector multiplication, then it is a vector bundle of rank m over the field $\mathbb{F}$. The vector space structure on each fiber can be defined via any choice of local trivialization belonging to the G-structure, and is independent of this choice due to the fact that transition functions are linear maps.[95]

Reducing $\mathrm{GL}(m, \mathbb{F})$ to a smaller Lie subgroup $G \subset \mathrm{GL}(m, \mathbb{F})$ still acting linearly on $\mathbb{F}^m$ corresponds to endowing the fibers E_p of our vector bundle with whatever extra structure is preserved by the group G, and in a way that depends smoothly on the point $p \in M$. The standard examples were already discussed last semester in Lecture 18, and they include:

- $\mathrm{GL}_+(m, \mathbb{R})$: E is a real oriented vector bundle.
- $\mathrm{O}(m)$ or $\mathrm{U}(m)$: E is endowed with a positive bundle metric, and is thus called a **Euclidean** vector bundle in the real case, or a **Hermitian** vector bundle in the complex case.
- $\mathrm{SL}(m, \mathbb{F})$: the fibers of E are endowed with nontrivial top-dimensional alternating $\mathbb{F}$-multilinear forms; we sometimes call this structure a (real or complex) **volume form** on the bundle E.
- $\mathrm{SO}(m)$: E is an oriented Euclidean vector bundle of rank m, or equivalently, a Euclidean vector bundle endowed with a volume form μ such that orthonormal bases $v_1, \ldots, v_m$ satisfy $\mu(v_1, \ldots, v_m) = \pm 1$. (Here the orientation determines the sign, and vice versa.)
- $\mathrm{SU}(m)$: E is a Hermitian vector bundle of rank m endowed with a complex volume form μ such that orthonormal bases $v_1, \ldots, v_m$ satisfy $|\mu(v_1, \ldots, v_m)| = 1$.
- $\mathrm{O}(k, \ell)$: E is a real vector bundle of rank $k + \ell$ endowed with an indefinite bundle metric of signature (k, ℓ).
- $\mathrm{Sp}(2m)$: E is a real vector bundle of rank $2m$ whose fibers are endowed with nondegenerate alternativing 2-forms; we then call it a **symplectic vector bundle** (cf. Example 25.29 from last semester).

[95] As you might recall from the first semester, this is exactly how one normally defines the vector space structure on the tangent spaces of a smooth manifold: write it down in local coordinates and show that the result is independent of the choice.

Example 42.17. For any Euclidean vector bundle $E \to M$ of rank m, the unit sphere bundle $SE \to M$ and unit disk bundle $\mathbb{D}E \to M$ of Examples 42.1 and 42.2 are both smooth fiber bundles with natural $\mathrm{O}(m)$-structures. Indeed, any local trivialization $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \times \mathcal{U}_\alpha \times \mathbb{R}^m$ in the $\mathrm{O}(m)$-structure of E can be restricted to SE and $\mathbb{D}E$, thus identifying their fibers with S^{m-1} and $\mathbb{D}^m$ respectively. The $\mathrm{O}(m)$-structures inherited by the two fiber bundles thus have exactly the same $\mathrm{O}(m)$-valued transition functions as E, but with $\mathrm{O}(m)$ acting on S^{m-1} and $\mathbb{D}^m$ instead of $\mathbb{R}^m$. It follows for instance that the fibers of SE carry natural Riemannian metrics, defined to match the standard metric on S^{m-1}; this does not depend on a choice of trivializations since the $\mathrm{O}(m)$-valued transition functions act on S^{m-1} by isometries.

Example 42.18. For a vector bundle $E \to M$ of rank m over $\mathbb{F}$, the Grassmann and Stiefel bundles $\mathrm{Gr}_k(E)$ and $\mathrm{St}_k(E)$ of Example 42.3 inherit natural $\mathrm{GL}(m,\mathbb{F})$-structures whose transition functions are the same as for E, but acting smoothly on the manifolds $\mathrm{Gr}_k(\mathbb{F}^m)$ and $\mathrm{St}_k(\mathbb{F}^m)$ respectively instead of $\mathbb{F}^m$. Analogous statements hold for $\mathrm{St}_k^{\mathrm{O}}(E)$ and $\widetilde{\mathrm{Gr}}_k(E)$ if E is endowed with an $\mathrm{O}(m)$-structure or a $\mathrm{GL}_+(m,\mathbb{R})$-structure respectively.

Example 42.19. We defined in Example 42.4 the frame bundle $FE = \mathrm{St}_m(E)$ of a vector bundle $E \to M$: assuming E has rank m over $\mathbb{F}$, this is the special case of Example 42.18 with $k = m$, thus FE inherits from E a $\mathrm{GL}(m,\mathbb{F})$-structure with the same transition functions, which act on $\mathrm{St}_m(\mathbb{F}^m) = \mathrm{GL}(m,\mathbb{F})$ by left multiplication. If E has a positive bundle metric, thus endowing it with an $\mathrm{O}(\mathbb{F}^m)$-structure, then one also has the orthonormal frame bundle $F^{\mathrm{O}}(E) = \mathrm{St}_m^{\mathrm{O}}(E)$, which analogously inherits from E an $\mathrm{O}(\mathbb{F}^m)$-structure with the same transition functions acting on $\mathrm{St}_m^{\mathrm{O}}(\mathbb{F}^m) = \mathrm{O}(\mathbb{F}^m)$ by left multiplication. (Recall that, by definition, $\mathrm{O}(\mathbb{R}^m) = \mathrm{O}(m)$ and $\mathrm{O}(\mathbb{C}^m) = \mathrm{U}(m)$.)

Exercise 42.20. Show that G-structures are well behaved under the pullback operation defined in §42.1. Concretely, if $\pi : M \to E$ is a fiber bundle and $f : N \to M$ a smooth map, then associating to each local trivialization Φ_α of E the pullback trivialization $f^*\Phi_\alpha$ and to each transition function $g_{\beta\alpha}$ on $\mathcal{U}_\alpha \cap \mathcal{U}_\beta \subset M$ the function $g_{\beta\alpha} \circ f$ on $f^{-1}(\mathcal{U}_\alpha) \cap f^{-1}(\mathcal{U}_\beta) \subset N$ turns any G-bundle atlas $\mathcal{A}$ on E into a G-bundle atlas $f^*\mathcal{A}$ on f^*E. Show moreover that any G-bundle isomorphism $\Psi : (E^1, \mathcal{A}^1) \to (E^2, \mathcal{A}^2)$ similarly determines a G-bundle isomorphism $f^*\Psi : (f^*E^1, f^*\mathcal{A}^1) \to (f^*E^2, f^*\mathcal{A}^2)$.

42.3. Transition functions determine the bundle. According to our definitions in the previous section, every G-bundle atlas includes a system of G-valued transition functions (satisfying the cocycle condition) as part of its data, and a G-bundle isomorphism implies the existence of a morphism of systems of transition functions. We would now like to invert this relationship and show that, up to isomorphism, everything important about a G-bundle is determined by its transition functions. This implies for instance that for any given G-bundle, one can find a multitude of other related G-bundles by changing the standard fiber and/or its G-action but keeping the same system of transition functions.

Theorem 42.21. *Fix a Lie group G, manifolds F, M and a smooth action $\rho : G \times F \to F$. Then for any system of G-valued transition functions $\mathcal{T} = (\{\mathcal{U}_\alpha\}_{\alpha \in I}, \{g_{\beta\alpha}\}_{(\alpha,\beta)\in I\times I})$ on M, there exists a smooth fiber bundle $\pi : E \to M$ with a G-bundle atlas of the form $(\mathbf{\Phi}, \mathcal{T}, \rho)$.*

Moreover, if $\pi^j : E^j \to M$ for $j = 1, 2$ are two G-bundles with G-compatible bundle atlases of the form $\mathcal{A}^j = (\mathbf{\Phi}^j, \mathcal{T}^j, \rho)$, then any morphism $\mathcal{T}^1 \to \mathcal{T}^2$ canonically determines a G-bundle isomorphism $\Psi : (E^1, \mathcal{A}^1) \to (E^2, \mathcal{A}^2)$.

Proof. Given the system of transition functions $\mathcal{T}$, we define the total space E of our desired fiber bundle as the quotient

$$E := \left(\coprod_{\alpha \in I} \mathcal{U}_\alpha \times F \right) \Big/ \sim,$$

where the equivalence relation is defined such that for all $\alpha, \beta \in I$,

$$\mathcal{U}_\alpha \times F \ni (p, x) \sim (q, y) \in \mathcal{U}_\beta \times F \qquad \Leftrightarrow \qquad p = q \text{ and } y = g_{\beta\alpha}(p)x.$$

The cocycle condition (42.3) guarantees that this is indeed an equivalence relation. The projection $\pi : E \to M : [(p, x)] \mapsto p$ is then well defined, and E admits a unique topology and smooth structure for which π is smooth and the identity map on $\mathcal{U}_\alpha \times F$ descends to a diffeomorphism $E \supset \pi^{-1}(\mathcal{U}_\alpha) \to \mathcal{U}_\alpha \times F$ for each $\alpha \in I$. Interpreting these maps as trivializations gives a suitable G-bundle atlas for $\pi : E \to M$.

Next, given two fiber bundles $\pi^j : E^j \to M$ for $j = 1, 2$, with G-bundle atlases $(\mathbf{\Phi}^j, \mathcal{T}^j, \rho)$ including trivializations $\mathbf{\Phi}^j = \{\Phi_\alpha : E^j|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times F\}_{\alpha \in I_j}$ and systems of G-valued transition functions $\mathcal{T}^j = (\{\mathcal{U}_\alpha\}_{\alpha \in I_j}, \{g_{\beta\alpha}\}_{(\alpha,\beta) \in I_j \times I_j})$, suppose $\{h_{\beta\alpha} : \mathcal{U}_\beta \cap \mathcal{U}_\alpha \to G\}_{(\alpha,\beta) \in I_1 \times I_2}$ is a morphism from $\mathcal{T}^1$ to $\mathcal{T}^2$. There is then a unique map $\Psi : E^1 \to E^2$ determined by the condition

$$\Phi_\beta \circ \Psi \circ \Phi_\alpha^{-1}(p, x) = (p, h_{\beta\alpha}(p)x)$$

for all $\alpha \in I_1$, $\beta \in I_2$, $p \in \mathcal{U}_\alpha \cap \mathcal{U}_\beta$ and $x \in F$; here it is an easy exercise to check that $\Psi|_{E_p}$ does not depend on the choices of $\alpha \in I_1$ and $\beta \in I_2$ with $p \in \mathcal{U}_\alpha \cap \mathcal{U}_\beta$, due to the condition (42.4). $\square$

Remark 42.22. Theorem 42.21 is the main reason why we explicitly required the cocycle condition in our definition of a G-structure. If we had a non-effective action and a badly chosen system of transition functions for which the cocycle condition is not satisfied, then it would not be possible to alter the fiber and group action arbitrarily while keeping the same transition functions.

43. Principal bundles

One consequence of our discussion of fiber bundles thus far is that a single system of G-valued transition functions $\{g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to G\}_{(\alpha,\beta) \in I \times I}$ can give rise to many different bundles, all realized via different choices of the standard fiber F and a left G-action on F. We can formalize this with the following definition.

Definition 43.1. For a smooth manifold M and Lie group G, an **abstract G-bundle** over M is an equivalence class of systems of G-valued transition functions on M, where two systems are considered equivalent if and only if there exists a morphism (see Definition 42.8) between them.

The equivalence relation in this definition makes sense due to the arguments in Lemma 42.10, just as in our definition of G-structures in the previous lecture. Notice what this definition does not mention: an abstract G-bundle does not have fibers, or a total space. The idea is rather that thanks to Theorem 42.21, one can use an abstract G-bundle as the foundation on which to construct a G-bundle with any desired standard fiber F that has a left G-action, and up to isomorphism, every G-bundle comes from such a construction. The fiber bundles constructed in this way from a single abstract bundle are all in some sense equivalent, even though their fibers and total spaces may be very different manifolds. We will see in particular that the notion of connections and parallel transport compatible with a G-structure can be defined in a way that only makes reference to the underlying abstract G-bundle, thus it determines notions of parallel transport simultaneously on all the associated fiber bundles.

The main message of the present lecture is the observation that every abstract G-bundle has a canonical realization in the form of a so-called *principal* fiber bundle, constructed in terms of the natural left action of G on itself. Of all the fiber bundles associated to a particular abstract bundle, the principal bundle is the one with the nicest properties, and it can be used to construct all the other associated bundles without any explicit reference to triviliazations or transition functions. For this reason, it is conventional to derive most of the theory of connections on fiber bundles from the special case of principal bundles, and this is what we will do in the next few lectures.

43.1. Two definitions. The following is the more abstract of the two equivalent definitions we will give, but it fits neatly into the more general context of the previous lecture.

Definition 43.2 (principal bundles, version 1). For a Lie group G and manifold M, a **principal G-bundle** over M is a G-bundle $\pi : E \to M$ for which the standard fiber is G and the structure group G acts on it via multiplication from the left.

If we were to define a vector bundle in the same language, we would say it is a $\mathrm{GL}(m, \mathbb{F})$-bundle with standard fiber $\mathbb{F}^m$ for some $m \geqslant 0$ and $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, such that the structure group $\mathrm{GL}(m, \mathbb{F})$ acts on $\mathbb{F}^m$ in the canonical way via linear maps. This is of course not the most popular way to phrase the definition of a vector bundle, as for instance, one still has to work a bit to deduce from this definition that each fiber has a natural vector space structure. There is an analogous way to reformulate Definition 43.2 in terms of the intrinsic structure carried by the fibers of a principal bundle, and this reformulation is by far the more popular (and useful) version of the definition. We know already that the fibers of a principal G-bundle are diffeomorphic to the structure group G, though in general they will not have natural group structures, just as the manifold $\mathrm{St}_m(V)$ for an m-dimensional vector space V is not a group in any canonical way, even though it is diffeomorphic to $\mathrm{GL}(m, \mathbb{F})$. What the fibers of a principal bundle do have, however, is a free and transitive action of G *from the right.*

Proposition 43.3. *If $\pi : E \to M$ is a principal G-bundle in the sense of Definition 43.2, then it admits a unique smooth right action*

$$E \times G \to E$$

such that for every G-compatible local trivialization $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times G$,

$$\Phi_\alpha^{-1}(p, x)g = \Phi_\alpha^{-1}(p, xg) \qquad \textit{for all } p \in \mathcal{U}_\alpha,\ x, g \in G. \tag{43.1}$$

This G-action on E preserves each of the fibers $E_p \subset E$ and acts on each one freely and transitively. Conversely, if $\pi : E \to M$ is any smooth fiber bundle with a smooth fiber-preserving right action of a Lie group G that is free and transitive on each fiber, then $\pi : E \to M$ admits a unique G-structure making it into a principal G-bundle such that the right G-action is as described above.

Proof. The G-action on a single fiber E_p defined via the relation in (43.1) is manifestly free and transitive since the same is true for the natural right action of G on itself. We claim that this action is also independent of the choice of G-compatible trivialization $(\mathcal{U}_\alpha, \Phi_\alpha)$ with $p \in \mathcal{U}_\alpha$: indeed, if $(\mathcal{U}_\beta, \Phi_\beta)$ is another one, then we have

$$\Phi_\beta^{-1}(p, x) = \Phi_\alpha^{-1} \circ (\Phi_\alpha \circ \Phi_\beta^{-1})(p, x) = \Phi_\alpha^{-1}(p, g_{\alpha\beta}(p)x)$$

for all $x \in G$, and thus for $g \in G$,

$$\Phi_\beta^{-1}(p, x)g = \Phi_\alpha^{-1}(p, g_{\alpha\beta}(p)xg) = \Phi_\beta^{-1}(p, xg).$$

The crucial property we are using here is the associativity of the group law in G, which can also be expressed as the fact that for any two elements $g, h \in G$, the left and right translation diffeomorphisms $L_g, R_h : G \to G$ always commute with each other. To put it another way, the natural left and right actions of G on itself commute with each other: this is why describing $\pi : E \to M$ in local trivializations via a *left* action of the structure group on the standard fiber allows us to define a global *right* action on E that is independent of any choice of trivialization. We could also have done it the other way around; this is just a convention.

For the converse, suppose $\pi : E \to M$ is a fiber bundle with a smooth right action $E \times G \to E$ that restricts to a free and transitive action on each fiber. Using local trivializations of $E \to M$, one can find an open covering $\{\mathcal{U}_\alpha\}_{\alpha \in I}$ of M such that E admits a smooth local section $s_\alpha \in \Gamma(E|_{\mathcal{U}_\alpha})$ for each $\alpha \in I$. The fact that G acts freely and transitively on each fiber then implies that the map

$$\mathcal{U}_\alpha \times G \to E|_{\mathcal{U}_\alpha} : (p, g) \mapsto s_\alpha(p)g$$

is a diffeomorphism, so we define a local trivialiazation $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times G$ as its inverse. On the overlap $\mathcal{U}_\alpha \cap \mathcal{U}_\beta$ for any $(\alpha, \beta) \in I \times I$, freeness and transitivity also imply that there is a unique smooth function $g_{\alpha\beta} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to G$ such that $s_\beta = s_\alpha g_{\alpha\beta}$, and one easily checks that the two corresponding trivializations are then related by $\Phi_\alpha \circ \Phi_\beta^{-1}(p, x) = (p, g_{\alpha\beta}(p)x)$. Since the left action of G on itself is free (and therefore effective), the collection of functions $\{g_{\alpha\beta}\}_{(\alpha,\beta) \in I \times I}$ automatically satisfies the cocycle condition, so all of this data together forms a G-bundle atlas for $\pi : E \to M$ with standard fiber G. □

Exercise 43.4. Extend the correspondence in Proposition 43.3 as follows: show that if $\pi^j : E^j \to M$ for $j = 1, 2$ are two principal G-bundles in the sense of Definition 43.2, then a smooth map $\Psi : E^1 \to E^2$ that is fiber preserving (i.e. it sends $E^1_p \to E^2_p$ for every $p \in M$) is a G-bundle isomorphism if and only if it is equivariant with respect to the right G-actions on E^1 and E^2, meaning

$$\Psi(xg) = \Psi(x)g \qquad \text{for all } x \in E,\ g \in G.$$

The proposition implies that Definition 43.2 is equivalent to the following:

Definition 43.5 (principal bundles, version 2). For a Lie group G and manifold M, a **principal G-bundle** over M is a smooth fiber bundle $\pi : E \to M$ that is equipped with a smooth and fiber-preserving right action $E \times G \to E$ whose restriction to each fiber is free and transitive. Given two principal G-bundles $E^1, E^2 \to M$ in this sense, a **principal bundle isomorphism** is a smooth map $\Psi : E^1 \to E^2$ that is fiber preserving and G-equivariant.

The following fact buried in the proof of Proposition 43.3 is worth drawing attention to:

Proposition 43.6. *Given a principal G-bundle $\pi : E \to M$ and an open subset $\mathcal{U} \subset M$, there is a natural bijection between the space of local sections $\Gamma(E|_{\mathcal{U}})$ over $\mathcal{U}$ and the set of all G-compatible trivializations $\Phi : E|_{\mathcal{U}} \to \mathcal{U} \times G$ over $\mathcal{U}$: concretely, a trivialization Φ gives rise to the section $s(p) = \Phi^{-1}(p, e)$, and can be recovered from this section via the relation*

$$\Phi^{-1}(p, g) = s(p)g \qquad \textit{for } (p, g) \in \mathcal{U} \times G.$$

In particular, a global section $s \in \Gamma(E)$ exists if and only if the bundle is trivial. □

Most interesting principal bundles one can come up with turn out to be nontrivial, so one of the messages of Proposition 43.6 is that on principal bundles, one should not typically expect global sections to exist. In fact, many important problems involving the existence of geometric structures on manifolds can be reduced to the question of whether a particular principal bundle admits a global section. Algebraic topologists have developed quite powerful methods for solving the latter problem; the subject is known as *obstruction theory* (see e.g. [**Ste51**]).

Example 43.7. For any smooth right action $M \times G \to M$ that is free and proper, the slice theorem (Theorem 40.25) associates to a submanifold $\Sigma \subset M$ satisfying certain properties an embedding

$$\Sigma \times G \to M : (p, g) \mapsto pg$$

that is a diffeomorphism onto the G-orbit of Σ. We can interpret such diffeomorphisms as inverses of local trivializations that make the quotient projection $\pi : M \to M/G$ into a smooth fiber bundle, and since G acts freely and transitively on each of its orbits in M, the G-action makes $\pi : M \to M/G$ into a principal G-bundle. The slice $\Sigma \subset M$ is then equivalent to a local section of this bundle over $\mathcal{U} := \pi(\Sigma) \subset M/G$, i.e. there exists a unique section $s : \mathcal{U} \to M$ of $\pi : M \to M/G$ whose image is Σ.

Exercise 43.8. Show that for any principal G-bundle $\pi : E \to M$, the right G-action on E is free and proper, so E/G has a natural smooth structure and is diffeomorphic to M.

Example 43.9. We observed in the previous lecture that for any vector bundle $E \to M$ of rank m over the field $\mathbb{F}$, the frame bundle $FE \to M$ has a natural $\mathrm{GL}(m, \mathbb{F})$ structure in which the structure group acts on the standard fiber $\mathrm{GL}(m, \mathbb{F})$ by left multiplication, thus FE is a principal $\mathrm{GL}(m, \mathbb{F})$-bundle. The right action of $\mathrm{GL}(m, \mathbb{F})$ on each fiber FE_p is easiest to describe if we regard frames on E_p as vector space isomorphisms $\phi : \mathbb{F}^m \to E_p$: the action of a matrix $\mathbf{A} \in \mathrm{GL}(m, \mathbb{F})$ on ϕ is then given by composition of linear transformations,

$$\phi \cdot \mathbf{A} := \phi \circ \mathbf{A} : \mathbb{F}^m \to E_p.$$

If the vector bundle $E \to M$ carries some additional structure reducing its structure group to a Lie subgroup $G \subset \mathrm{GL}(m, \mathbb{F})$, then $FE \to M$ inherits this G-structure, with the structure group G still acting on the standard fiber $\mathrm{GL}(m, \mathbb{F})$ by left multiplication. The latter action however preserves the submanifold $G \subset \mathrm{GL}(m, \mathbb{F})$, thus defining a submanifold

$$F^G E \subset FE$$

that is naturally a principal G-bundle over M, called the **G-frame bundle** of E. Its fibers $F^G E_p$ consist of all frames $(v_1, \dots, v_m) \in FE_p$ that can be identified via some G-compatible local trivialization with the standard basis of $\mathbb{F}^m$. So for instance, if $G = \mathrm{O}(\mathbb{F}^m)$, then $F^G E$ is again the orthonormal frame bundle $F^{\mathrm{O}}(E)$. If $G = \mathrm{SL}(m, \mathbb{R})$, meaning that E is a real vector bundle equipped with a volume form on its fibers, then $F^G E$ is the space of all frames that span parallelepipeds of signed volume 1. The right G-action on $F^G E$ is just the restriction to $F^G E \subset FE$ of the right action of $G \subset \mathrm{GL}(m, \mathbb{F})$ on FE, thus it can also be described as above in terms of composition of invertible linear maps.

Exercise 43.10. Show that if $G \times M \to M$ is a smooth and transitive left group action, then for any $p \in M$, the map $G \to M : p \mapsto gp$ defines a principal G_p-bundle, where the stabilizer G_p acts on G by multiplication from the right.
Hint: You already know that this map descends to a diffeomorphism $G/G_p \to M$, and G_p acts freely and properly on G from the right.

43.2. Associated bundles. The main power of principal G-bundles is that all other G-bundles with the same transition functions can be derived from them. Here is the general construction.

Suppose $\pi : E \to M$ is a principal G-bundle, F is a smooth manifold and $\rho : G \times F \to F : (g, x) \mapsto gx$ is a smooth group action. We can define a smooth left G-action on $E \times F$ by

$$g(\phi, x) := (\phi g^{-1}, gx) \qquad \text{for } g \in G,\ \phi \in E,\ x \in F,$$

and this action is free and proper due to the fact that G acts freely and properly on E (see Exercise 43.8). The quotient

$$E^\rho := E \times_\rho F := (E \times F)\big/G$$

is thus a smooth manifold, and comes equipped with a smooth map

$$\pi^\rho : E^\rho \to M : [\phi, x] \mapsto \pi(\phi),$$

where $[\phi, x]$ denotes the equivalence class in $(E \times F)/G$ represented by a pair $(\phi, x) \in E \times F$. This object will be called the **associated bundle** determined by E, F and ρ. We claim indeed that it is a smooth fiber bundle, and moreover, that it carries a natural G-structure with standard fiber F. To see this, choose a G-bundle atlas $\{(\mathcal{U}_\alpha, \Phi_\alpha)\}_{\alpha \in I}$ for E, and recall that each local trivialization Φ_α is equivalent to a local section $s_\alpha \in \Gamma(E|_{\mathcal{U}_\alpha})$ given by $s_\alpha(p) = \Phi_\alpha^{-1}(p, e)$. These sections are related to each other by the transition functions determined uniquely by our choice of trivializations, namely by

$$s_\alpha = s_\beta g_{\beta\alpha} \qquad \text{on } \mathcal{U}_\alpha \cap \mathcal{U}_\beta.$$

For $p \in \mathcal{U}_\alpha$, every point in the fiber E^ρ_p is then representable as $[s_\alpha(p), x]$ for a unique $x \in F$, due to the fact that G acts freely and transitively on E_p. We can thus define local trivializations of $\pi^\rho : E^\rho \to M$ by

$$\Phi^\rho_\alpha : E^\rho|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times F \qquad \text{such that} \qquad (\Phi^\rho_\alpha)^{-1}(p, x) = [s_\alpha(p), x].$$

If $p \in \mathcal{U}_\alpha \cap \mathcal{U}_\beta$, we then find

$$(\Phi^\rho_\alpha)^{-1}(p, x) = [s_\alpha(p), x] = [s_\beta(p) g_{\beta\alpha}(p), x] = [s_\beta(p), g_{\beta\alpha}(p) x] = (\Phi^\rho_\beta)^{-1}(p, g_{\beta\alpha}(p) x),$$

and thus

$$\Phi^\rho_\beta \circ (\Phi^\rho_\alpha)^{-1}(p, x) = (p, g_{\beta\alpha}(p) x).$$

We've proved:

THEOREM 43.11. *For any principal G-bundle $\pi : E \to M$ and any smooth manifold F with a smooth left G-action $\rho : G \times F \to F$, the associated bundle $\pi^\rho : E^\rho \to M$ is a smooth fiber bundle, and carries a natural G-structure (with standard fiber F) such that any G-compatible bundle atlas on E determines a G-compatible bundle atlas on E^ρ with the same transition functions.* □

One application of this construction is that it inverts the correspondence sending vector bundles to their frame bundles. Indeed, if $F^G E \to M$ is the G-frame bundle for some vector bundle $E \to M$ with structure group $G \subset \mathrm{GL}(m, \mathbb{F})$, then for the standard action ρ of G on $\mathbb{F}^m$ by linear maps, applying the associated bundle construction to $F^G E$ gives a vector bundle $(F^G E)^\rho \to M$ that is isomorphic (as a G-bundle) to E. We know this abstractly because $(F^G E)^\rho$ and E are two bundles that have the same standard fiber with the same left G-action and the same transition functions. However, an explicit isomorphism can also be written down if we regard frames $\phi \in F^G E_p$ as invertible linear maps $\phi : \mathbb{F}^m \to E_p$: the map

$$(F^G E)^\rho = F^G E \times_\rho \mathbb{F}^m \to E : [\phi, \mathbf{v}] \mapsto \phi(\mathbf{v})$$

is then well defined and does the trick.

43.3. Parallel transport. We can now begin the discussion of connections on fiber bundles. The most natural place to start is with the notion of parallel transport. A connection on a fiber bundle $\pi : E \to M$ should associate to any smooth path $\gamma(t) \in M$ a smooth family of diffeomorphisms

$$P^t_\gamma : E_{\gamma(0)} \to E_{\gamma(t)}, \qquad P^0_\gamma = \mathrm{Id},$$

which we refer to as **parallel transport** on E along γ. If E is endowed with a G-structure, then we also want these diffeomorphisms to respect that structure. The precise meaning of the latter condition is a bit cumbersome to define in general, though it seems clear what it should mean in various concrete examples: on a vector bundle, it means P^t_γ is linear, if that bundle also has a bundle metric, then P^t_γ should be orthogonal, and if E is a principal G-bundle, P^t_γ should be equivariant with respect to the right G-action.

We will see below that any connection compatible with a G-structure on a fiber bundle is determined by a corresponding connection on the associated principal G-bundle. In the concrete situation of a vector bundle $E \to M$ of rank m over $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, this just means that *linear* connections on $E \to M$ can always be derived from so-called *principal* connections on the frame bundle $FE \to M$. If you think in terms of parallel transport, then it is easy to see why this is so: a section of FE along γ is the same thing as a frame for E along γ (i.e. a frame for the pullback bundle γ^*E), so if we have a parallel frame, then it is natural to stipulate that a section $v(t) \in E_{\gamma(t)}$ of the vector bundle along γ is parallel if and only if it has constant components with respect to the parallel frame. In symbols, this relationship can be written down most easily by regarding frames as linear isomorphisms $\phi : \mathbb{F}^m \to E_p$; the relationship between parallel transport on FE and E is then

$$P_\gamma^t(\phi(\mathbf{v})) := P_\gamma^t(\phi)(\mathbf{v}) \qquad \text{for all } \mathbf{v} \in \mathbb{F}^m. \tag{43.2}$$

One needs to check of course that the parallel transport defined on E in this way does not depend on the choice of parallel frame, or equivalently, that (43.2) gives the same result if we replace $\phi \in FE_{\gamma(0)}$ by a different frame $\psi \in FE_{\gamma(0)}$ with $\psi(\mathbf{v}) = \phi(\mathbf{v})$. This is where equivariance comes in: since $\mathrm{GL}(m, \mathbb{F})$ acts on the fibers of FE freely and transitively, we can write $\psi = \phi \circ \mathbf{A}$ for a unique $\mathbf{A} \in \mathrm{GL}(m, \mathbb{F})$, with $\mathbf{A}\mathbf{v} = \mathbf{v}$ under the present assumptions, thus

$$P_\gamma^t(\psi(\mathbf{v})) = P_\gamma^t(\phi(\mathbf{A}\mathbf{v})) = (P_\gamma^t(\phi) \circ \mathbf{A})(\mathbf{v}) = P_\gamma^t(\phi \circ \mathbf{A})(\mathbf{v}) = P_\gamma^t(\psi)(\mathbf{v})$$

as required. Conversely, parallel transport on E determines parallel transport on FE, simply by regarding frames as m-tuples of vectors in E; a precise formula in terms of isomorphisms $\phi : \mathbb{F}^m \to E_p$ is

$$P_\gamma^t(\phi) := P_\gamma^t \circ \phi \in FE_{\gamma(t)} \qquad \text{for } \phi \in FE_{\gamma(0)},$$

and it is manifestly $\mathrm{GL}(m, \mathbb{F})$-equivariant. Moreover, if E has extra structure such as a bundle metric, giving it structure group $\mathrm{O}(\mathbb{F}^m)$, then requiring this metric to be preserved under parallel transport means that the parallel transport maps for FE preserve the orthonormal frame bundle $F^{\mathrm{O}}(E)$, and are of course $\mathrm{O}(\mathbb{F}^m)$-equivariant in that context. For the same reasons, any choice of $\mathrm{O}(m)$-equivariant parallel transport maps on $F^{\mathrm{O}}(E)$ determines parallel transport maps on E that preserve the bundle metric.

Let us now reframe the discussion in more general terms.

Definition 43.12. The parallel transport $P_\gamma^t : E_{\gamma(0)} \to E_{\gamma(t)}$ on a G-bundle $\pi : E \to M$ along a path γ through $\gamma(0) = p \in M$ is said to **respect the G-structure** if for some G-compatible local trivialization $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times M$ with $\mathcal{U}_\alpha$ containing the path γ, there exists a smooth path $\Pi_\alpha(t) \in G$ with $\Pi_\alpha(0) = e$ such that

$$P_\gamma^t \circ \Phi_\alpha^{-1}(p, x) = \Phi_\alpha^{-1}(\gamma(t), \Pi_\alpha(t)x).$$

Observe that if the condition in this definition holds for some particular choice of the trivialization $(\mathcal{U}_\alpha, \Phi_\alpha)$, then it holds for all choices, because if $(\mathcal{U}_\beta, \Phi_\beta)$ is another such trivialization, we have

$$\begin{aligned} P_\gamma^t \circ \Phi_\beta^{-1}(p, x) = P_\gamma^t \circ \Phi_\alpha^{-1}(p, g_{\alpha\beta}(p)x) &= \Phi_\alpha^{-1}(\gamma(t), \Pi_\alpha(t) g_{\alpha\beta}(p)x) \\ &= \Phi_\beta^{-1}(\gamma(t), g_{\beta\alpha}(\gamma(t))\Pi_\alpha(t) g_{\alpha\beta}(p)x) \\ &= \Phi_\beta^{-1}(\gamma(t), \Pi_\beta(t)x), \end{aligned}$$

where in the last line we are defining a new path in G through e by

$$\Pi_\beta(t) := g_{\beta\alpha}(\gamma(t))\Pi_\alpha(t) g_{\alpha\beta}(p). \tag{43.3}$$

Definition 43.12 has the disadvantage that it only makes sense for parallel transport along paths that are short enough to be contained in the domain of a local trivialization—in practice, however, the parallel transport maps we consider will arise from flows of vector fields, and can thus be presented as compositions of many flows defined for short times, to which the definition can be applied.

EXERCISE 43.13. Check that Definition 43.12 gives the notion you would expect in the case of vector bundles (with or without bundle metrics).

Let's check explicitly that Definition 43.12 gives the expected notion of *equivariant* parallel transport on principal bundles. Local trivializations $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times G$ in this case correspond to sections $s_\alpha : \mathcal{U}_\alpha \to E$ via $\Phi_\alpha^{-1}(q, g) = s_\alpha(q)g$, thus Definition 43.12 gives

$$P_\gamma^t(s_\alpha(p)g) = s_\alpha(\gamma(t))\Pi_\alpha(t)g \tag{43.4}$$

for all $g \in G$, implying $P_\gamma^t(xg) = P_\gamma^t(x)g$ for all $x \in E_{\gamma(0)}$ and $g \in G$. Conversely, if an equivariant parallel transport map P_γ^t and a local section s_α are given, then since G acts freely and transitively on each fiber, there is a unique function $\Pi_\alpha(t) \in G$ satisfying

$$P_\gamma^t(s_\alpha(p)) = s_\alpha(\gamma(t))\Pi_\alpha(t),$$

and clearly $\Pi_\alpha(0) = e$. Equivariance then reproduces the relation (43.4), and thus Definition 43.12.

It is worth noting that the notion of G-compatible parallel transport can be defined without any actual knowledge of the fibers of a G-bundle: according to the calculation above, if we are given a system of G-valued transition functions $(\{\mathcal{U}_\alpha\}_{\alpha\in I}, \{g_{\beta\alpha}\}_{(\alpha,\beta)\in I\times I})$ underlying the G-structure on our bundle, then parallel transport along a path $\gamma(t) \in M$ is fully determined by a collection of functions $\Pi_\alpha(t) \in G$, defined for each $\alpha \in I$ such that $\gamma(0) \in \mathcal{U}_\alpha$, that satisfy $\Pi_\alpha(0) = e$ and the transformation formula (43.3). In this sense, the notion of G-compatible parallel transport really belongs to the world of *abstract* G-bundles. For a given G-bundle $E \to M$, the functions Π_α may or may not be uniquely determined by the actual parallel transport maps P_γ^t and trivialization $(\mathcal{U}_\alpha, \Phi_\alpha)$; they can be non-unique if the action of G on the standard fiber is not effective. However, they are certainly unique if $E \to M$ is a principal bundle, since the action of G on itself is free. This is a good reason to focus on the case of principal bundles and derive whatever we need to know about other G-bundles from that special case. Since we now know that every G-bundle over M is isomorphic to the associated bundle E^ρ for some principal G-bundle $E \to M$ and left G-action $\rho : G \times F \to F$, the following result tells us how to do this:

PROPOSITION 43.14. *Suppose $\pi : E \to M$ is a principal G-bundle, $\rho : G \times F \to F$ is a left action and $P_\gamma^t : E^\rho_{\gamma(0)} \to E^\rho_{\gamma(t)}$ is a G-compatible family of parallel transport maps along a path $\gamma(t) \in M$ through $p := \gamma(0)$. Then there exists a (not necessarily unique) family of G-compatible parallel transport maps $P_\gamma^t : E_{\gamma(0)} \to E_{\gamma(t)}$ such that for every $[\phi, x] \in E_p^\rho = (E_p \times F)/G$,*

$$P_\gamma^t([\phi, x]) = [P_\gamma^t(\phi), x].$$

PROOF. Choose a local section $s_\alpha : \mathcal{U}_\alpha \to E$ with $p \in \mathcal{U}_\alpha$ and let Φ_α and Φ_α^ρ denote the corresponding local trivializations of E and E^ρ respectively. By assumption there is a smooth function $\Pi_\alpha(t) \in G$ with $\Pi_\alpha(0) = e$ such that $P_\gamma^t \circ (\Phi_\alpha^\rho)^{-1}(p, x) = (\Phi_\alpha^\rho)^{-1}(\gamma(t), \Pi_\alpha(t)x)$ for all $x \in F$, and in terms of the local section s_α of E, this means

$$P_\gamma^t([s_\alpha(p), x]) = [s_\alpha(\gamma(t)), \Pi_\alpha(t)x] = [s_\alpha(\gamma(t))\Pi_\alpha(t), x].$$

The desired parallel transport maps on E are thus the unique family of G-equivariant diffeomorphisms $E_{\gamma(0)} \to E_{\gamma(t)}$ such that $P_\gamma^t(s_\alpha(p)) := s_\alpha(\gamma(t))\Pi_\alpha(t)$. □

REMARK 43.15. If the G-action on F in Proposition 43.14 is not effective, then it may happen that the parallel transport on E^ρ fails to uniquely determine the parallel transport on the associated principal bundle E, because the function $\Pi_\alpha(t) \in G$ is not unique. There is, however, a canonical way to derive parallel transport on E^ρ from parallel transport on E, and this is a reason to focus on the case of principal bundles and derive whatever results we need for arbitrary G-bundles from that case. Proposition 43.14 guarantees that this is possible.

44. Connections on fiber bundles

Before coming to the main topic, here is a bit of notation that will be useful moving forward. For a smooth vector bundle $E \to M$ and integer $k \geqslant 0$, we denote

$$\Omega^k(M, E) := \Gamma(\Lambda^k T^*M \otimes E)$$

and interpret elements $\omega \in \Omega^k(M, E)$ as **bundle-valued** k-forms, i.e. they associate to each $p \in M$ an antisymmetric k-fold multilinear map $\omega_p : T_pM \times \ldots \times T_pM \to E_p$. The case $k = 0$ is included here and means

$$\Omega^0(M, E) = \Gamma(E).$$

Observe that the natural pullback of an E-valued k-form $\omega \in \Omega^k(M, E)$ via a smooth map $f : N \to M$ is a k-form on N with values in the pullback bundle f^*E,

$$f^*\omega \in \Omega^k(N, f^*E), \qquad (f^*\omega)_p(X_1, \ldots, X_k) := \omega_{f(p)}(f_*X_1, \ldots, f_*X_k) \in E_{f(p)} = (f^*E)_p.$$

If V is a vector space, we can define the space of V-valued k-forms $\Omega^k(M, V)$ similarly by viewing V as the trivial vector bundle $M \times V \to M$.

44.1. Connections and covariant derivatives. In this lecture we dispense with G-structures (they will return in the next lecture), and consider an arbitrary smooth fiber bundle $\pi : E \to M$. The total space E is a smooth manifold, and its tangent bundle has a smooth subbundle

$$VE := \ker \pi_* \subset TE,$$

called the **vertical subbundle**; in other words, the vertical tangent space V_xE for each $x \in E_p$ is $T_x(E_p)$, the tangent space to the fiber that x lives in. If we were assuming $E \to M$ to be a vector bundle, the next step would be to identify V_xE canonically with E_p, but we cannot do this if E_p is not a vector space, so we will now have to deal with vertical tangent spaces more directly than we did before.

A connection on $\pi : E \to M$ will associate to every smooth path $\gamma(t) \in M$ through $p := \gamma(0)$ a smooth family of parallel transport diffeomorphisms $P^t_\gamma : E_{\gamma(0)} \to E_{\gamma(t)}$. As discussed in the previous lecture, we would now impose more conditions on the maps P^t_γ if $\pi : E \to M$ were endowed with a G-structure, but without this, we will merely assume that they are diffeomorphisms. For a section $s \in \Gamma(\gamma^*E)$ along the path γ, parallel transport turns $s(t)$ into a path in the single fiber $E_{\gamma(0)}$ and thus produces a **covariant derivative**

$$\nabla_t s(0) := \left.\frac{d}{dt}(P^t_\gamma)^{-1}(s(t))\right|_{t=0} \in V_{s(0)}E. \tag{44.1}$$

Here we see another difference compared with the case of vector bundles: if $E_{\gamma(0)}$ were a vector space, then the smooth path $t \mapsto (P^t_\gamma)^{-1}(s(t)) \in E_{\gamma(0)}$ could be said to have a derivative in $E_{\gamma(0)}$, but if not, then the best we can do is to define a covariant derivative whose value is a vertical tangent vector, rather than an element of the fiber.

Nonetheless, (44.1) is a reasonable definition for the covariant derivative of a section along a path, and we would like to be able to turn it into a definition of a covariant derivative of sections $s \in \Gamma(E)$ with respect to tangent vectors on M, namely

$$\nabla_X s := \nabla_t s(\gamma(t))|_{t=0} \in V_{s(p)}E \qquad \text{for any path } \gamma(t) \in M \text{ with } \gamma(0) = p,\ \dot\gamma(0) = X \in T_pM.$$

If this object is really to be interpreted as a derivative, then we especially want it to have the property that $\nabla_X s$ depends on the tangent vector $X \in TM$ but not on a choice of path representing it, and moreover, we want the map $T_pM \to V_{s(p)}E : X \mapsto \nabla_X s$ to be linear. It turns out that these two conditions determine uniquely what the right definition of a connection should be and how to derive parallel transport from it. We proved this last semester in the case of vector bundles (see Lecture 19), and that proof requires only minor cosmetic modifications for the general case, thus we will not repeat it here, but shall go ahead and state the definition.

DEFINITION 44.1. A **connection** on a smooth fiber bundle $\pi : E \to M$ is a smooth linear subbundle $HE \subset TE$ whose fibers H_xE for $x \in E$ are everywhere complementary to the vertical tangent spaces, i.e.

$$H_xE \oplus V_xE = T_xE.$$

We call HE the **horizontal subbundle** or **horizontal distribution** for the connection. We will refer to the fiberwise-linear projection

$$K : TE \to VE$$

along HE as the **connection map**, or sometimes also the (global) **connection 1-form**, since it can be interpreted as a bundle-valued 1-form

$$K \in \Omega^1(E, VE).$$

For each $p \in M$ and $x \in E_p$, the unique linear map

$$\mathrm{Hor}_x : T_pM \to H_xE$$

such that $\pi_* \circ \mathrm{Hor}_x = \mathbb{1}_{T_pM}$ is called the **horizontal lift** map.

We observe that the connection map $K : TE \to VE$ in this definition uniquely determines the connection by writing

$$HE = \ker K \subset TE,$$

and vice versa. Moreover, *any* bundle-valued 1-form $K \in \Omega^1(E, VE)$ that satisfies $K|_{VE} = \mathbb{1}_{VE}$ has a kernel that is a smooth subbundle complementary to VE in TE, and can thus be regarded as a connection.

EXERCISE 44.2. Use partitions of unity and connection 1-forms $K \in \Omega^1(E, VE)$ to show that every smooth fiber bundle admits a connection.
Comment: This result is not as useful as one might think, since most fiber bundles arising in nature come with extra structure, and one needs to consider only connections that respect that structure. We will deal with this in the next lecture using principal bundles.

Given a connection and all the associated data as described in Definition 44.1, a section $s(t) \in E_{\gamma(t)}$ is now called **parallel** (or also **horizontal** or **covariantly constant**) if, as a path in the total space E, it satisfies

$$\dot{s}(t) \in H_{s(t)}E$$

for all t. This can be rephrased as the condition that $s(t)$ defines a flow line of a certain vector field on the total space of the pullback bundle γ^*E. Indeed, if $I \subset \mathbb{R}$ denotes the domain of the path γ, recall that we defined γ^*E in §42.1 as the submanifold

$$\gamma^*E = \big\{(t, x) \in I \times E \;\big|\; \pi(x) = \gamma(t)\big\} \subset I \times E$$

and by the implicit function theorem (Exercise 42.7), its tangent space at a point $(t, x) \in \gamma^*E$ is then the subspace

$$T_{(t,x)}(\gamma^*E) = \big\{(s, \xi) \in \mathbb{R} \times T_xE \;\big|\; \pi_*\xi = s\dot{\gamma}\big\} \subset \mathbb{R} \times T_xE = T_{(t,x)}(I \times E),$$

where we are using the canonical identification of T_tI with $\mathbb{R}$. The expression

$$\eta(t, x) := (1, \mathrm{Hor}_x(\dot{\gamma}(t)))$$

thus defines a vector field on γ^*E, and the section $s : I \to \gamma^*E$ is parallel if and only if $s(t) = \varphi^t_\eta(s(0))$.

If the flow φ^t_η is defined globally on $(\gamma^*E)_0 = E_{\gamma(0)} \subset \gamma^*E$ for a given t, then it defines a **parallel transport** diffeomorphism

$$P^t_\gamma : E_{\gamma(0)} \to E_{\gamma(t)},$$

which depends smoothly on t. The caveat here is that at this level of generality, if the fibers are noncompact, the flow might indeed not exist globally, so that P^t_γ for each t can be defined only on some open subset of $E_{\gamma(0)}$. This possibility will be excluded when we require our connections to be compatible with G-structures, thus we shall choose not to worry about it for now, and just continue under the pretense that P^t_γ is defined globally for every t in the domain of the path γ. (No important details will depend on it.)

Having defined parallel transport, (44.1) now gives us a covariant derivative $\nabla_t s(t) \in V_{s(t)}E$ for any smooth section $s(t) \in E_{\gamma(t)}$ along a path $\gamma(t) \in M$. (To define it at points $t \neq 0$, just regard t as a constant and replace γ with the path $s \mapsto \gamma(t+s)$, which passes through $\gamma(t)$ at $s = 0$.) One can also write down a simpler formula for covariant derivatives in terms of the connection map $K : TE \to VE$: rewriting P^t_γ via the flow of the vector field $\eta \in \mathfrak{X}(\gamma^*E)$ described above, we have

$$\begin{aligned}\nabla_t s(0) &= \frac{d}{dt}\varphi^{-t}_\eta(s(t))\Big|_{t=0} = \frac{d}{dt}\varphi^{-t}_\eta(s(0))\Big|_{t=0} + \frac{d}{dt}\varphi^0_\eta(s(t))\Big|_{t=0} = -\eta(s(0)) + \frac{d}{dt}(t, s(t))\Big|_{t=0}\\ &= (-1, -\mathrm{Hor}_{s(0)}(\dot\gamma(0))) + (1, \dot s(0))\\ &= (0, \dot s(0) - \mathrm{Hor}_{s(0)}(\dot\gamma(0))) \in V_{(0,s(0))}(\gamma^*E) = \{0\} \times V_{s(0)}E \subset T_{(0,s(0))}(I \times E).\end{aligned}$$

Since $\pi_* \dot s(0) = \dot\gamma(0)$, the horizontal lift term in this last expression is precisely the horizontal part of $\dot s(0) \in T_{s(0)}E$ with respect to the splitting $TE = HE \oplus VE$, thus

$$\nabla_t s(t) = K(\dot s(t)), \tag{44.2}$$

where we have taken the liberty of writing down the formula for general t in the domain of the path γ, since there is nothing special about the point $t = 0$. As a corollary, the covariant derivative of a section $s : M \to E$ with respect to a tangent vector $X \in T_pM$ at a point $p \in M$ can now be written in terms of the tangent map $Ts : TM \to TE$ and connection map $K : TE \to VE$ as

$$\nabla_X s = K(Ts(X)). \tag{44.3}$$

This shows in particular that the map $X \mapsto \nabla_X s$ has the desired property: it is linear and independent of any choice of path tangent to X.

EXERCISE 44.3. For a fiber bundle $\pi : E \to M$ and a smooth map $f : N \to M$, the pullback bundle f^*E comes with a natural smooth map $\Psi : f^*E \to E$ that sends each fiber $(f^*E)_p$ diffeomorphically to the fiber $E_{f(p)}$. Prove:

(a) There is a canonical isomorphism between the vector bundles Ψ^*VE and $V(f^*E)$ over f^*E.
(b) If a connection on $E \to M$ with connection 1-form $K \in \Omega^1(E, VE)$ is given, then $\Psi^*K \in \Omega^1(f^*E, \Psi^*VE) = \Omega^1(f^*E, V(f^*E))$ is a connection 1-form for $f^*E \to N$.
(c) The connection defined on $f^*E \to N$ via part (b) is the unique one with the property that a section $s(t) \in (f^*E)_{\gamma(t)}$ of f^*E along a path $\gamma(t) \in N$ is parallel if and only if it is also parallel when interpreted as a section $s(t) \in E_{f\circ\gamma(t)}$ of E along the path $f \circ \gamma(t) \in M$.

We will refer to the connection on $f^*E \to N$ defined in this exercise as the **pullback connection** determined by a connection on $E \to M$.

If a connection on $\pi : E \to M$ is given, then a section $s \in \Gamma(E|_{\mathcal{U}})$ defined on an open set $\mathcal{U} \subset M$ is called **parallel** (or **horizontal** or **flat**) if its covariant derivative vanishes identically.

More generally, for a smooth map $f : N \to M$, a section $s \in \Gamma(f^*E)$ of E along f is called parallel if it is parallel as a section of $f^*E \to N$ with respect to the pullback connection.

44.2. Flat connections and curvature. We now consider the most fundamental question about a connection on a fiber bundle: do parallel sections of $\pi : E \to M$ exist over open subset $\mathcal{U} \subset M$? This question is much less trivial than asking for parallel sections along a path; the latter always exist because they are flow lines of a vector field, but by the same token, a parallel section $s \in \Gamma(E|_{\mathcal{U}})$ needs to have the property that for any two paths $\alpha, \beta : [0,1] \to \mathcal{U}$ connecting $\alpha(0) = \beta(0) = p$ to $\alpha(1) = \beta(1) = q$, the parallel transport of $s(p)$ along α and β will end up at the same point $s(q)$. The uniqueness of these parallel transport maps makes this seem unlikely in general, and indeed, it is impossible for most connections on fiber bundles over manifolds of dimension at least 2. The obstruction to the local existence of parallel sections on a fiber bundle is measured by its curvature.

Definition 44.4. A connection on the smooth fiber bundle $\pi : E \to M$ is called **flat** if for every $p \in M$ and $x \in E_p$, there exists a neighborhood $\mathcal{U} \subset M$ of p and a parallel section $s \in \Gamma(E|_{\mathcal{U}})$ with $s(p) = x$.

Exercise 44.5. Assume $\pi : E \to M$ is a smooth fiber bundle whose fibers are compact. Prove that $\pi : E \to M$ admits a flat connection if and only if it admits a G-structure where G is a 0-dimensional Lie group.

Flatness can be reframed in the language of integrable distributions. Recall that on a smooth manifold M, a smooth **k-plane distribution** is by definition a smooth subbundle $\xi \subset TM$ of rank k, and a submanifold $\Sigma \subset M$ is called an **integral submanifold** of the distribution ξ if

$$T_p\Sigma \subset \xi_p \qquad \text{for all } p \in \Sigma.$$

The k-plane distribution $\xi \subset TM$ is called **integrable** if for every point $p \in M$, there exists a smooth k-dimensional integral submanifold containing p. Note that the integral submanifold in this definition is not required to be closed, and it may be contained in an arbitrarily small neighborhood of p; integrability of a distribution is thus an essentially local condition.

In the language of distributions, the horizontal subbundle $HE \subset TE$ for a connection on a fiber bundle $\pi : E \to M$ over an n-manifold is a smooth n-plane distribution on E, and a local section $M \supset \mathcal{U} \xrightarrow{s} E$ defines an n-dimensional submanifold $\Sigma := s(\mathcal{U}) \subset E$, which is an *integral* submanifold for the distribution $HE \subset TE$ if and only if the section s is parallel. The connection is thus flat if and only if the distribution $HE \subset TE$ is integrable.

The problem of deciding whether a given distribution is integrable is solved by the Frobenius integrability theorem. We will state and prove the Frobenius theorem below, as a corollary of the answer to the question on how to decide whether a given connection is flat. We carried out the same argument last semester in the context of vector bundles (see Lecture 26), and the general case is no different—in fact, the argument is somewhat more natural in the fiber bundle context, as it is not actually helpful at all to have a linear structure on the fibers.

Definition 44.6. Given a connection $HE \subset TE$ on the fiber bundle $\pi : E \to M$, with connection 1-form $K \in \Omega^1(E, VE)$, the correesponding **curvature 2-form** $F_K \in \Omega^2(E, VE)$ is defined via the formula

$$F_K(\eta, \xi) := -K\big([H(\eta), H(\xi)]\big) \qquad \text{for } \eta, \xi \in \mathfrak{X}(E),$$

where $H : TE \to HE$ denotes the fiberwise-linear projection along VE, i.e. the projection complementary to $K : TE \to VE$.

It is straightforward to check that $F_K(\eta, \xi)$ by this definition is C^∞-linear in both η and ξ, so it defines a smooth and antisymmetric bilinear bundle map $TE \oplus TE \to VE$, which we therefore interpret as a VE-valued 2-form. It vanishes identically if and only if for every pair of horizontal vector fields $\eta, \xi \in \Gamma(HE) \subset \mathfrak{X}(E)$, the vector field $[\eta, \xi] \in \mathfrak{X}(E)$ is also horizontal, i.e. it is also in $\Gamma(HE)$. The minus sign in the definition is unimportant at this juncture; its purpose is only to make F_K consistent with other definitions that we will encounter later.

Here is the main result.

THEOREM 44.7. *A connection on a fiber bundle is flat if and only if its curvature 2-form vanishes identically.*

The main ingredient we need for the proof is the theorem from last semester that vector fields have commuting flows if and only if their Lie brackets vanish. Given the bundle $\pi : E \to M$ and a connection $HE \subset TE$, we can define a linear map

$$\mathfrak{X}(M) \to \mathfrak{X}(E) : X \mapsto X^h$$

by

$$X^h(x) := \mathrm{Hor}_x(X(p)) \in H_xE \subset T_xE \qquad \text{for } x \in E_p,\ p \in M.$$

As it happens, this map is *nearly* a Lie algebra homomorphism—that is, it would be one if we were allowed to simply ignore the vertical part of $[X^h, Y^h]$:

LEMMA 44.8. *For any $X, Y \in \mathfrak{X}(M)$, $[X, Y]^h = H \circ [X^h, Y^h]$.*

PROOF. Since $[X, Y]^h$ and $H \circ [X^h, Y^h]$ are both purely horizontal vector fields on E, it suffices to check that both define the same derivation when restricted to functions in $C^\infty(E)$ that are constant in the vertical directions, i.e. functions of the form $f \circ \pi$ for $f \in C^\infty(M)$. Moreover, the difference between $[X^h, Y^h]$ and $H \circ [X^h, Y^h]$ is purely vertical and thus vanishes when applied to any function of this form, and it therefore suffices to prove

$$\mathcal{L}_{[X^h,Y^h]}(f \circ \pi) = \mathcal{L}_{[X,Y]^h}(f \circ \pi)$$

for all $f \in C^\infty(M)$. For this, we can use the fact that for any $Z \in \mathfrak{X}(M)$, $p \in M$ and $x \in E_p$,

$$\mathcal{L}_{Z^h}(f \circ \pi)(x) = d(f \circ \pi)(Z^h(x)) = df(\pi_* Z^h(x)) = df(Z(p)) = \mathcal{L}_Z f(\pi(x)),$$

giving the relation

$$\mathcal{L}_{Z^h}(f \circ \pi) = \mathcal{L}_Z f \circ \pi.$$

The rest is a straightforward calculation from the definition of the Lie bracket:

$$\begin{aligned}
\mathcal{L}_{[X^h,Y^h]}(f \circ \pi) &= \mathcal{L}_{X^h}\mathcal{L}_{Y^h}(f \circ \pi) - \mathcal{L}_{Y^h}\mathcal{L}_{X^h}(f \circ \pi) \\
&= \mathcal{L}_{X^h}\left((\mathcal{L}_Y f) \circ \pi\right) - \mathcal{L}_{Y^h}\left((\mathcal{L}_X f) \circ \pi\right) = (\mathcal{L}_X\mathcal{L}_Y f) \circ \pi - (\mathcal{L}_Y\mathcal{L}_X f) \circ \pi \\
&= (\mathcal{L}_{[X,Y]} f) \circ \pi = \mathcal{L}_{[X,Y]^h}(f \circ \pi).
\end{aligned}$$

□

PROOF OF THEOREM 44.7. Assume $F_K \equiv 0$. Since the question of flatness is essentially local, we lose no generality if we replace M with a small neighborhood of some point $p \in M$ on which a chart $(x^1, \dots, x^n)$ can be defined. Denote the resulting coordinate vector fields by $X_j := \partial_j \in \mathfrak{X}(M)$ for $j = 1, \dots, n$; as coordinate vector fields, they satisfy $[X_i, X_j] = 0$ for all i and j. Now since $F_K(X_i^h, X_j^h) = 0$, the vector fields $[X_i^h, X_j^h]$ are horizontal, thus by Lemma 44.8,

$$[X_i^h, X_j^h] = H \circ [X_i^h, X_j^h] = [X_i, X_j]^h = 0.$$

It follows that for any $x \in E_p$, we can construct an integral submanifold through x via the commuting flows of X_i^h: it is parametrized by the map

$$\psi(t^1, \ldots, t^n) = \varphi_{X_1^h}^{t^1} \circ \ldots \circ \varphi_{X_n^h}^{t^n}(x) \tag{44.4}$$

for real numbers $t^1, \ldots, t^n$ sufficiently close to 0.

Conversely, if $HE \subset TE$ is a flat connection and thus an integrable distribution, then for every $p \in M$ and $x \in E_p$ there is a parallel section $s : \mathcal{U} \to E$ on some neighborhood $\mathcal{U} \subset M$ of p with $s(p) = x$, and its image $\Sigma := s(\mathcal{U}) \subset E$ is thus an integral submanifold for the horizontal distribution. Any two sections $\eta, \xi \in \Gamma(HE)$ then define vector fields on Σ, and their Lie bracket therefore restricts to Σ as another vector field on Σ, implying that $[\eta, \xi]$ is horizontal along Σ, and in particular at x. Since the point x was chosen arbitrarily, it follows that $[\eta, \xi]$ is horizontal everywhere, and thus $F_K(\eta, \xi) \equiv 0$. □

As a fringe benefit, Theorem 44.7 provides a good framework to attack the more general question of whether a k-plane distribution $\xi \subset TM$ on an n-manifold M is integrable. The situation is not quite the same as the distribution $HE \subset TE$, because M in this general setting is not the total space of a fiber bundle, so there is no "vertical" subbundle $VM \subset TM$ complementary to ξ in the picture. Locally, however, it is always possible to make some choices such that the two situations look exactly the same: in particular, every point $p \in M$ admits a neighborhood $\mathcal{U} \subset M$ that is the total space of a fiber bundle $\pi : \mathcal{U} \to \Sigma$ on which $\xi|_{\mathcal{U}} \subset T\mathcal{U}$ can be regarded as a connection. (The idea is simply to choose $\mathcal{U}$ inside a small coordinate neighborhood and write down the fibration $\pi : \mathcal{U} \to \Sigma$ in suitable local coordinates.) As a consequence, every point $p \in M$ has a neighborhood on which Theorem 44.7 can be applied to determine whether ξ is integrable. The condition $F_K \equiv 0$ does not make sense globally on M since M is not globally the total space of a fiber bundle, but that condition is equivalent to the condition

$$X, Y \in \Gamma(\xi) \quad \Rightarrow \quad [X, Y] \in \Gamma(\xi),$$

which does make sense globally. This proves:

Theorem 44.9 (Frobenius). *A smooth distribution $\xi \subset TM$ on a manifold M is integrable if and only if the Lie bracket $[X, Y] \in \mathfrak{X}(M)$ of every pair of vector fields $X, Y \in \mathfrak{X}(M)$ everywhere tangent to ξ is also everywhere tangent to ξ.* □

45. Principal connections

For a fiber bundle $\pi : E \to M$ with finite-dimensional structure group G, extra conditions need to be placed on the definition of a connection so that parallel transport will respect the G-structure. As indicated in §43.3, the best way to do this is by defining a connection on the associated *principal* G-bundle, which will then determine a connection on every other bundle that has the same transition functions.

Definition 45.1. A connection on a principal G-bundle $\pi : E \to M$ is called a **principal connection** if it gives rise to G-equivariant parallel transport maps $P_\gamma^t : E_{\gamma(0)} \to E_{\gamma(t)}$ along every path $\gamma(t) \in M$.

This definition is conceptually simple but hard to work with in practice, so we will now derive some other conditions that are equivalent to it.

45.1. Equivariant horizontal subbundles. The first step is to determine the implications of equivariant parallel transport for the horizontal subbundle $HE \subset TE$ of a connection. If $\pi : E \to M$ is a principal G-bundle, then each $g \in G$ determines a fiber preserving diffeomorphism

$$R_g : E \to E : \phi \mapsto \phi g.$$

Choose a path $\gamma(t) \in M$ with $\gamma(0) = p$, $\dot\gamma(0) = X \in T_pM$, and let $\phi \in E_p$. Given a principal connection, the horizontal lift isomorphisms have the property

$$\mathrm{Hor}_{\phi g}(X) = \left.\frac{d}{dt} P^t_\gamma(\phi g)\right|_{t=0} = \left.\frac{d}{dt} R_g \circ P^t_\gamma(\phi)\right|_{t=0} = TR_g\left(\left.\frac{d}{dt}P^t_\gamma(\phi)\right|_{t=0}\right) = TR_g\left(\mathrm{Hor}_\phi(X)\right).$$

We conclude $H_{\phi g}E = TR_g(H_\phi E)$. Conversely, for any connection on $\pi : E \to M$ with this property, a section $s(t) \in E_{\gamma(t)}$ of E along γ will be parallel if and only if the section $t \mapsto R_g(s(t)) = s(t)g$ is also parallel for every $g \in G$, implying that P^t_γ is G-equivariant.[96] We've proved:

PROPOSITION 45.2. *On a principal G-bundle $\pi : E \to M$, a connection $HE \subset TE$ is a principal connection if and only if $TR_g(HE) = HE$ for every $g \in G$.* $\square$

45.2. Lie algebra-valued connection 1-forms. Next, we reformulate $TR_g(HE) = HE$ as a condition on the connection 1-form $K \in \Omega^1(E, VE)$.

Recall from §40.2 that the group action $E \times G \to E$ determines a linear map $\mathfrak{g} \to \mathfrak{X}(E) : X \mapsto X^F$, where the *fundamental vector field* determined by $X \in \mathfrak{g}$ is given by

$$X^F(\phi) = \left.\frac{d}{dt}\phi \exp(tX)\right|_{t=0}.$$

The fundamental vector field X^F vanishes at a point $\phi \in E$ if and only if X belongs to the Lie algebra of the stabilizer subgroup $G_\phi \subset G$; since the action in the present setting is free, that means nontrivial fundamental vector fields on E are *nowhere* vanishing. Moreover, they all point in vertical directions since the G-action is fiber preserving, and it follows that for each $\phi \in E$, the map

$$\mathfrak{g} \to V_\phi E : X \mapsto X^F(\phi)$$

is an isomorphism, thus defining a vector bundle isomorphism between VE and the trivial bundle over E with fiber $\mathfrak{g}$. This makes it possible to reexpress $K \in \Omega^1(E, VE)$ as a $\mathfrak{g}$-valued 1-form

$$A \in \Omega^1(E, \mathfrak{g}) \qquad \text{such that} \qquad K(\xi) = A(\xi)^F(\phi) \text{ for all } \phi \in E,\ \xi \in T_\phi E,$$

and the condition $K|_{VE} = \mathbb{1}_{VE}$ then translates into

$$A(X^F(\phi)) = X \qquad \text{for all } X \in \mathfrak{g} \text{ and } \phi \in E.$$

With this understood, we now ask: what additional condition on a Lie algebra-valued 1-form $A \in \Omega^1(E, \mathfrak{g})$ is necessary and sufficient for the corresponding connection on $E \to M$ to be a principal connection? The answer requires a short lemma as preparation.

LEMMA 45.3. *For $g \in G$ and $X \in \mathfrak{g}$, the pushed-forward vector field $(R_g)_* X^F \in \mathfrak{X}(E)$ matches the fundamental vector field of $\mathrm{Ad}_{g^{-1}}(X) \in \mathfrak{g}$.*

[96] Recall that on a general fiber bundle with noncompact fibers, the diffeomorphisms $P^t_\gamma : E_{\gamma(0)} \to E_{\gamma(t)}$ might not even be defined globally for any given $t \neq 0$, due to vector fields having flows that blow up in finite time. However, G-equivariance prevents this from happening: if $P^t_\gamma(\phi)$ is defined for t in some interval $I \subset \mathbb{R}$, then for every $g \in G$, $P^t_\gamma(\phi g) = P^t_\gamma(\phi)g$ is also defined for $t \in I$.

Proof. For each point $\phi \in E$, we have $\big((R_g)_* X^F\big)(R_g(\phi)) = TR_g(X^F(\phi))$ by definition, thus

$$\begin{aligned}(R_g)_* X^F(\phi g) = TR_g(X^F(\phi)) &= \frac{d}{dt} R_g\left(\phi \exp(tX)\right)\Big|_{t=0} = \frac{d}{dt}\phi \exp(tX) g\Big|_{t=0} \\ &= \frac{d}{dt}\phi g\left(g^{-1}\exp(tX) g\right)\Big|_{t=0} = \frac{d}{dt}\phi g \exp(t\,\mathrm{Ad}_{g^{-1}}(X))\Big|_{t=0} = (\mathrm{Ad}_{g^{-1}}(X))^F(\phi g).\end{aligned}$$

$\square$

Now, at each point $\phi \in E$, the horizontal-vertical splitting allows us to write any tangent vector $\xi \in T_\phi E$ uniquely as $\xi = \xi^h + X^F(\phi)$ for some $\xi^h \in H_\phi E$ and $X \in \mathfrak{g}$. The $\mathfrak{g}$-valued 1-form $A \in \Omega^1(E, \mathfrak{g})$ corresponding to our connection 1-form K then satisfies $A(\xi) = X$, so if $H_{\phi g}E = TR_g(H_\phi E)$, Lemma 45.3 implies

$$R_g^* A(\xi) = A(TR_g(\xi^h) + ((R_g)_* X^F)(\phi g)) = \mathrm{Ad}_{g^{-1}}(X) = \mathrm{Ad}_{g^{-1}} \circ A(\xi)$$

due to the fact that $TR_g(\xi^h) \in H_{\phi g}E \subset \ker A$. Conversely, if $A \in \Omega^1(E, \mathfrak{g})$ satisfies the condition $R_g^* A = \mathrm{Ad}_{g^{-1}} \circ A$, it implies via the calculation above that every $\xi \in H_\phi E$ satisfies $A(TR_g(\xi)) = 0$ and thus $TR_g(HE) = HE$. This proves:

Proposition 45.4. *For a principal G-bundle $\pi : E \to M$, a Lie algebra-valued 1-form $A \in \Omega^1(E, \mathfrak{g})$ defines a principal connection by $HE := \ker A \subset TE$ if and only if it satisfies the following conditions:*

(i) $A(X^F(\phi)) = X$ for all $X \in \mathfrak{g}$ and $\phi \in E$;
(ii) $R_g^* A = \mathrm{Ad}_{g^{-1}} \circ A$ for all $g \in G$.

$\square$

In light of Proposition 45.4, one can (and many authors do) define the term "principal connection" to mean a $\mathfrak{g}$-valued 1-form A satisfying the two conditions listed above, and one then obtains the original definition from this by setting $HE := \ker A$. There are many advantages to viewing principal connections as a special class of $\mathfrak{g}$-valued 1-forms, one of which is that it gives some recognizable structure to the set

$$\mathcal{A}(E) := \{\text{principal connections on } E\},$$

which can now be regarded as an affine space over the vector space of 1-forms $B \in \Omega^1(E, \mathfrak{g})$ that vanish on VE and satisfy the linear condition $R_g^* B = \mathrm{Ad}_{g^{-1}} \circ B$. In particular, the space of connections is naturally a convex set, and local constructions of them can be pieced together via a partition of unity on M, proving:

Theorem 45.5. *Every principal G-bundle admits a principal connection.* $\square$

Exercise 45.6. Work out the details of the proof of Theorem 45.5.

45.3. A digression on exterior algebra. Since we will be encountering vector-valued differential forms more and more often, it's worth pausing to clarify certain algebraic questions, such as: (1) What is the wedge product? (2) What is the exterior derivative? (3) Do they have the properties we think they should? The short answer to the last question is best given in German: *jein*.

For a manifold M and real vector space V, $\Omega^k(M, V)$ is the space of smooth sections of a vector bundle whose fiber over each point $p \in M$ is $\Lambda^k T_p^* M \otimes V$, where for $\omega \in \Lambda^k T_p^* M$ and $v \in V$, we identify

$$\omega v := \omega \otimes v \in \Lambda^k T_p^* M \otimes V$$

with the V-valued alternating k-form $(X_1, \dots, X_k) \mapsto \omega(X_1, \dots, X_k)v$. One cannot define a wedge product of two forms in $\Omega^*(M, V)$ unless V itself is an algebra, i.e. it needs to have a bilinear product structure of its own, so that products of V-valued functions can be defined, and a natural product on $\Omega^*(M, V)$ can then be derived from this. More generally, one can do the following: suppose $V_1, \dots, V_N$ and W are real vector spaces and

$$\mu : V_1 \times \dots \times V_N \to W$$

is a multilinear map, which we could equivalently regard as a linear map $V_1 \otimes \dots \otimes V_N \to W$. For any tuple of integers $k_1, \dots, k_N \geqslant 0$ there is then an obvious linear map

$$\begin{aligned} \Lambda^{k_1} T_p^* M \otimes V_1 \otimes \dots \otimes \Lambda^{k_N} T_p^* M \otimes V_N &\to \Lambda^{k_1 + \dots + k_N} T_P^* M \otimes W \\ \omega_1 \otimes v_1 \otimes \dots \otimes \omega_N \otimes v_N &\mapsto (\omega_1 \wedge \dots \wedge \omega_N) \otimes \mu(v_1, \dots, v_N), \end{aligned}$$

which can also be regarded as a multilinear map

$$(\Lambda^{k_1} T_p^* M \otimes V_1) \times \dots \times (\Lambda^{k_N} T_p^* M \otimes V_N) \to \Lambda^{k_1 + \dots + k_N} T_p^* M \otimes W.$$

Applying this at every point $p \in M$, we obtain a multilinear map between the corresponding spaces of vector-valued forms,

$$\mu : \Omega^{k_1}(M, V_1) \times \dots \times \Omega^{k_N}(M, V_N) \to \Omega^{k_1 + \dots + k_N}(M, W),$$

uniquely determined by the property that for any $\omega_j \in \Omega^{k_j}(M)$ and $v_j \in V$ for $j = 1, \dots, N$,

$$\mu(\omega_1 v_1, \dots, \omega_N v_N) = (\omega_1 \wedge \dots \wedge \omega_N) \; \mu(v_1, \dots, v_N).$$

The case $N = 2$ is the most important for us in the near term, though the general case will also arise when we discuss characteristic classes. For $N = 2$, a bilinear map $\mu : V_1 \times V_2 \to W$ gives rise to a bilinear map

$$\mu : \Omega^k(M, V_1) \times \Omega^\ell(M, V_2) \to \Omega^{k+\ell}(M, W)$$

for each $k, \ell \geqslant 0$, and the characterization above can be combined with the explicit formula (9.6) for the wedge product of real-valued forms to give the formula
(45.1)

$$\mu(\alpha, \beta)(X_1, \dots, X_{k+\ell}) = \frac{1}{k!\ell!} \sum_{\sigma \in S_{k+\ell}} (-1)^{|\sigma|} \mu(\alpha(X_{\sigma(1)}, \dots, X_{\sigma(k)}), \beta(X_{\sigma(k+1)}, \dots, X_{\sigma(k+\ell)})).$$

Depending on the nature of the bilinear map $\mu : V_1 \times V_2 \to W$ in this picture, it may or may not seem appropriate to denote $\alpha \wedge \beta := \mu(\alpha, \beta)$. This is somewhat common in situations where $V_1 = V_2 = W =: V$, so that V is an algebra with μ as its bilinear product, but here the wedge product notation needs to be handled with care, because in general there is no guarantee that it will be either associative or graded commutative. Instead, we have:

EXERCISE 45.7. Assume V is a real vector space with a bilinear product $\mu(v, w) := vw$, and denote $\alpha \wedge \beta := \mu(\alpha, \beta)$ for $\alpha, \beta \in \Omega^*(M, V)$. Prove:

(a) If the product on V is associative, then the wedge product on $\Omega^*(M, V)$ is also associative.
(b) If the product on V is commutative, then the wedge product on $\Omega^*(M, V)$ is graded commutative, i.e. $\alpha \wedge \beta = (-1)^{|\alpha| \cdot |\beta|} \beta \wedge \alpha$.
(c) If V is a commutative and associative algebra with an identity element $\mathbb{1} \in V$, then $1 \otimes \mathbb{1} \in \Omega^0(M, V)$ is an identity element for the wedge product on $\Omega^*(M, V)$.

Assume now that V is an associative algebra, and we are also given a real vector space W with a bilinear map $\nu : V \times W \to W$ making W a left V-module, and denote $\alpha \wedge \beta := \nu(\alpha, \beta) \in \Omega^*(M, W)$ for $\alpha \in \Omega^*(M, V)$ and $\beta \in \Omega^*(M, W)$. Prove:

(d) The pairing $\Omega^*(M, V) \times \Omega^*(M, W) \xrightarrow{\wedge} \Omega^*(M, W)$ makes $\Omega^*(M, W)$ a left $\Omega^*(M, V)$-module, and it is unital if V has an identity element.

Next, suppose $\mathfrak{g}$ is a Lie algebra with bracket $\mu(v,w) := [v,w]$, and denote $[\alpha,\beta] := \mu(\alpha,\beta)$ for $\alpha,\beta \in \Omega^*(M,\mathfrak{g})$. Prove:

(e) The bracket on $\Omega^*(M,\mathfrak{g})$ is **graded anticommutative**

$$[\alpha,\beta] + (-1)^{|\alpha|\cdot|\beta|}[\beta,\alpha] = 0$$

and satisfies a **graded Jacobi identity**[97]

$$[\alpha,[\beta,\gamma]] + (-1)^{|\alpha|\cdot(|\beta|+|\gamma|)}[\beta,[\gamma,\alpha]] + (-1)^{|\gamma|\cdot(|\alpha|+|\beta|)}[\gamma,[\alpha,\beta]] = 0.$$

Finally, suppose $\mathfrak{g}$ is a Lie algebra and V is a vector space with a bilinear map $\nu : \mathfrak{g} \times V \to V$ defined by $\nu(X,v) := \rho(X)v$ for some Lie algebra representation $\rho : \mathfrak{g} \to \mathfrak{gl}(V) = \mathrm{End}(V)$. Writing $\alpha \wedge \beta := \nu(\alpha,\beta) \in \Omega^*(M,V)$ for $\alpha \in \Omega^*(M,\mathfrak{g})$ and $\beta \in \Omega^*(M,V)$, prove:

(f) The bracket on $\Omega^*(M,\mathfrak{g})$ and its wedge product with $\Omega^*(M,V)$ are related by

$$[\alpha,\beta] \wedge \gamma = \alpha \wedge (\beta \wedge \gamma) - (-1)^{|\alpha|\cdot|\beta|}\beta \wedge (\alpha \wedge \gamma)$$

for $\alpha,\beta \in \Omega^*(M,\mathfrak{g})$ and $\gamma \in \Omega^*(M,V)$.

Hint: It suffices in every case to restrict your attention to forms that are products of real-valued forms with vectors in V.

Whenever $\mathfrak{g}$ is a Lie algebra, we will from now on use the prescription in Exercise 45.7 to define a bilinear bracket $[\ ,\ ]$ on $\Omega^*(M,\mathfrak{g})$, and if V is an associative algebra, we similarly define an associative wedge product on $\Omega^*(M,V)$. A commonly occurring special case is the matrix algebra $\mathfrak{g} = V = \mathbb{F}^{m\times m}$ with its commutator bracket, in which both definitions are sensible, but the following caveats should be observed:

- $\alpha \wedge \beta \neq -\beta \wedge \alpha$ in general for $\alpha,\beta \in \Omega^1(M,\mathbb{F}^{m\times m})$ since $\mathbb{F}^{m\times m}$ is not commutative. In particular, $\alpha \wedge \alpha \in \Omega^2(M,\mathbb{F}^{m\times m})$ can be nonzero.
- $[\alpha,\beta] = \alpha \wedge \beta - (-1)^{|\alpha|\cdot|\beta|}\beta \wedge \alpha = \alpha \wedge \beta + \beta \wedge \alpha = [\beta,\alpha]$ for $\alpha,\beta \in \Omega^1(M,\mathfrak{g})$, without a minus sign, thus $[\alpha,\alpha] \in \Omega^2(M,\mathfrak{g})$ can also be nonzero. Indeed, the formula (45.1) in this case gives

$$[\alpha,\alpha](X,Y) = [\alpha(X),\alpha(Y)] - [\alpha(Y),\alpha(X)] = 2[\alpha(X),\alpha(Y)].$$

For any vector space V, the exterior derivative $d : \Omega^*(M,V) \to \Omega^*(M,V)$ has a simple definition as the unique linear map such that

$$d(\omega v) = (d\omega)v \qquad \text{for all } \omega \in \Omega^*(M),\ v \in V.$$

It is straightforward to show that this operator satisfies $d^2 = 0$, as well as the following graded Leibniz rule: for any multilinear map $\mu : V_1 \times \ldots \times V_N \to W$ and forms $\alpha_j \in \Omega^{k_j}(M,V_j)$ for $j = 1,\ldots,N$,

$$\begin{aligned} d\big(\mu(\alpha_1,\ldots,\alpha_N)\big) &= \mu(d\alpha_1,\alpha_2,\ldots,\alpha_N) + (-1)^{k_1}\mu(\alpha_1,d\alpha_2,\ldots,\alpha_N) \\ &\quad + \ldots + (-1)^{k_1+\ldots+k_{N-1}}\mu(\alpha_1,\ldots,\alpha_{N-1},d\alpha_N). \end{aligned} \tag{45.2}$$

So for instance, the exterior derivative on $\Omega^*(M,\mathfrak{g})$ for a Lie algebra $\mathfrak{g}$ satisfies

$$d[\alpha,\beta] = [d\alpha,\beta] + (-1)^{|\alpha|}[\alpha,d\beta].$$

[97] You can remember the signs in the graded Jacobi identity if you observe the following rule: take the usual Jacobi identity, but wherever the order of the elements α,β,γ has been permuted, insert a minus sign for every time two elements of odd degree have been interchanged. The general theory behind such sign rules can be formulated nicely in categorical terms; see e.g. [Var04].

The discussion of products above generalizes easily to bundle-valued forms: if $E_1, \dots, E_N, F \to M$ are smooth vector bundles, a smooth linear bundle map

$$\mu : E_1 \otimes \dots \otimes E_N \to F$$

defines multilinear maps on the corresponding fibers, so that one naturally obtains a multilinear map

$$\mu : \Omega^{k_1}(M, E_1) \times \dots \times \Omega^{k_N}(M, E_N) \to \Omega^{k_1+\dots+k_N}(M, F)$$

for each tuple of integers $k_1, \dots, k_N \geqslant 0$. An easy special case is to take $N = 2$ with the trivial real line bundle as either E_1 or E_2, and the map $\mu : E_1 \otimes E_2 \to F(= E_1 \text{ or } E_2)$ defined by scalar multiplication, which gives rise to more-or-less obvious definitions of wedge products $\Omega^k(M) \times \Omega^\ell(M, E) \to \Omega^{k+\ell}(M, E)$ and $\Omega^k(M, E) \times \Omega^\ell(M) \to \Omega^{k+\ell}(M, E)$, making $\Omega^*(M, E)$ a left and right $\Omega^*(M)$-module.

Defining an exterior derivative on $\Omega^*(M, E)$ is less straightforward: the definition given above on $\Omega^*(M, V)$ relies tacitly on our understanding that each vector $v \in V$ can be interpreted as a *constant* vector-valued function, whereas vector bundles do not come with an intrinsic notion of constant sections. What's needed, therefore, is a choice of connection on E, which defines the covariant derivative operator $\nabla : \Omega^0(M, E) = \Gamma(E) \to \Gamma(\mathrm{Hom}(TM, E)) = \Omega^1(M, E)$. Given this, the **covariant exterior derivative**

$$d_\nabla : \Omega^k(M, E) \to \Omega^{k+1}(M, E)$$

is defined for each $k \geqslant 0$ as the unique linear map that matches ∇ for $k = 0$ and satisfies the graded Leibniz rule

$$d_\nabla(\alpha \wedge \beta) = d\alpha \wedge \beta + (-1)^{|\alpha|}\alpha \wedge d_\nabla\beta \qquad \text{for all } \alpha \in \Omega^k(M),\ \beta \in \Omega^\ell(M, E).$$

The uniqueness of d_∇ satisfying these properties is clear since every bundle-valued form is locally a linear combination of real-valued forms multiplied by sections. To show that such an operator exists, one can write down an explicit formula locally using coordinates and a frame for E, then show that this formula satisfies the required Leibniz rule.

EXERCISE 45.8. On a vector bundle $E \to M$ with connection ∇, prove that the operator $d_\nabla : \Omega^k(M, E) \to \Omega^{k+1}(M, E)$ satisfies

$$\begin{aligned} (45.3) \quad d_\nabla\omega(X_0, \dots, X_k) = &\sum_{i=0}^{k} (-1)^i \nabla_{X_i}\left(\omega(X_0, \dots, \hat{X}_i, \dots, X_k)\right) \\ &+ \sum_{0 \leqslant i < j \leqslant k} (-1)^{i+j}\omega\big([X_i, X_j], X_0, \dots, \hat{X}_i, \dots, \hat{X}_j, \dots, X_k\big) \end{aligned}$$

for $X_0, \dots, X_k \in \mathfrak{X}(M)$, where the hats in sequences mean that those terms do not appear. *Hint: We proved the corresponding formula for real-valued forms in §8.2 of last semester's notes. You can make use of that if you focus on the special case $\omega = \alpha\eta$ for some $\alpha \in \Omega^k(M)$ and $\eta \in \Gamma(E)$, which implies the rest via linearity.*

Notice what is missing from this discussion: we are not claiming $d_\nabla^2 = 0$, and in general, it is not true. In fact, applying (45.3) to $\nabla\eta \in \Omega^1(M, E)$ for a section $\eta \in \Gamma(E)$ reproduces the Riemann tensor of the connection:

$$d_\nabla^2\eta(X, Y) = d_\nabla(\nabla\eta)(X, Y) = \nabla_X(\nabla\eta(Y)) - \nabla_Y(\nabla\eta(X)) - \nabla\eta([X, Y]) = R(X, Y)\eta.$$

We'll come back to this observation in the next lecture, where it will be used to show that the Riemann tensor is essentially equivalent to the curvature 2-form on the associated principal bundle, and is thus an obstruction to flatness.

If $E_1, \ldots, E_N, F \to M$ are all equipped with connections and $\mu : E_1 \otimes \ldots \otimes E_N \to F$ is a smooth linear bundle map, one would hope to see a generalization of the Leibniz rule (45.2) in the form

$$d_\nabla\big(\mu(\alpha_1, \ldots, \alpha_N)\big) = \mu(d_\nabla\alpha_1, \alpha_2, \ldots, \alpha_N) + (-1)^{k_1}\mu(\alpha_1, d_\nabla\alpha_2, \ldots, \alpha_N) + \ldots + (-1)^{k_1+\ldots+k_{N-1}}\mu(\alpha_1, \ldots, \alpha_{N-1}, d_\nabla\alpha_N) \tag{45.4}$$

for $\alpha_j \in \Omega^{k_j}(M, E_j)$, $j = 1, \ldots, N$. In general this will not hold for arbitrary choices of connections and bundle map μ, e.g. for $k_1 = \ldots = k_N = 0$, (45.4) says

$$\nabla_X\big(\mu(\eta_1 \otimes \ldots \otimes \eta_N)\big) = \mu(\nabla_X\eta_1 \otimes \eta_2 \ldots \otimes \eta_N) + \ldots + \mu(\eta_1 \otimes \ldots \otimes \eta_{N-1} \otimes \nabla_X\eta_N)$$

for all $X \in \mathfrak{X}(M)$ and $\eta_j \in \Gamma(E_j)$, $j = 1, \ldots, N$, and the latter is true if and only if $\nabla\mu \equiv 0$ for the connection induced on $\mathrm{Hom}(E_1 \otimes \ldots \otimes E_N, F)$ by the connections on $E_1, \ldots, E_N$ and F (see §33.2). This does hold in many situations that naturally arise: for instance, if $E, F \to M$ are two bundles with connections, then the induced connection on $\mathrm{Hom}(E, F)$ is defined so that the canonical bundle map

$$\mathrm{Hom}(E, F) \otimes E \to F : A \otimes \eta \mapsto A\eta$$

is parallel.

Exercise 45.9. Show that the graded Leibniz rule (45.4) holds if and only if the bundle map $\mu \in \Gamma(\mathrm{Hom}(E_1 \otimes \ldots \otimes E_N, F))$ is parallel with respect to the natural connection induced by the connections on $E_1, \ldots, E_N$ and F.

Exercise 45.10. On any Lie group G, the **Maurer-Cartan form** is defined as the unique $\mathfrak{g}$-valued left-invariant 1-form $\theta \in \Omega^1(G, \mathfrak{g})$ such that $\theta_e = \mathbb{1}_{\mathfrak{g}}$.

(a) Prove that θ satisfies the so-called **Maurer-Cartan equation**:

$$d\theta + \frac{1}{2}[\theta, \theta] = 0.$$

Hint: The expression on the left is a $\mathfrak{g}$-valued 2-form on G, and it suffices to evaluate it on an arbitrary pair of left-invariant vector fields.

(b) Prove that θ transforms under right translations by

$$R_g^*\theta = \mathrm{Ad}_{g^{-1}} \circ \theta \qquad \text{for } g \in G.$$

45.4. The curvature 2-form. When $\pi : E \to M$ is a principal G-bundle, the vector bundle isomorphism $VE \cong E \times \mathfrak{g}$ defined via fundamental vector fields allows us to rewrite the curvature 2-form $F_K \in \Omega^2(E, VE)$ from §44.2 as a Lie algebra-valued 2-form

$$F_A \in \Omega^2(E, \mathfrak{g}), \qquad F_A(\eta, \xi) := -A([H(\eta), H(\xi)]),$$

where $H : TE \to HE$ again denotes the fiberwise-linear projection along VE. The words "curvature 2-form" will from now on refer to $F_A \in \Omega^2(E, \mathfrak{g})$ whenever the context is a principal connection.

Since $F_K \in \Omega^2(E, VE)$ and $F_A \in \Omega^2(E, \mathfrak{g})$ are completely equivalent objects, Theorem 44.7 tells us that a principal connection is flat if and only if $F_A \equiv 0$. The next theorem,[98] tells us that F_A is *nearly* a closed 2-form and is also *nearly* the exterior derivative of A, except for correction terms that vanish if the structure group G is abelian. These relations underlie many important results in differential geometry, including the Gauss-Bonnet theorem, the characterization of flatness via the Riemann tensor, and the Chern-Weil theory of characteristic classes.

[98] The first relation in Theorem 45.11 is taken as a *definition* of the curvature 2-form in some books. The calculations in the proof then need to be carried out in order to show that F_A also has something to do with an integrability condition for the horizontal subbundle. We have done things the other way around.

THEOREM 45.11. *A principal connection* 1*-form* $A \in \Omega^1(E, \mathfrak{g})$ *and its curvature* 2*-form* $F_A \in \Omega^2(E, \mathfrak{g})$ *satisfy the following relations:*

(i) $F_A = dA + \frac{1}{2}[A, A]$ *(second structural equation)*[99]
(ii) $dF_A = [F_A, A]$ *(second Bianchi identity)*

PROOF. In verbose form, the first equation says that for every pair of vector fields $\eta, \xi \in \mathfrak{X}(E)$,

$$dA(\eta, \xi) = F_A(\eta, \xi) - [A(\eta), A(\xi)],$$

where it should be stressed that the bracket appearing on the right hand side is the Lie bracket on $\mathfrak{g}$, not a bracket of vector fields. We will prove this via the formula (45.3), or rather the special case of it in which the bundle and connection are trivial (which was essentially the definition of the exterior derivative given in Lecture 8 last semester). In light of the splitting $TE = HE \oplus VE$, any vector field $\eta \in \mathfrak{X}(E)$ has the same value at a given point $\phi \in E$ as $X^h + Y^F$ for some $X \in \mathfrak{X}(M)$ and $Y \in \mathfrak{g}$, where $X^h \in \Gamma(HE)$ is the horizontal lift of X defined by $X^h(\phi) := \mathrm{Hor}_\phi(X(\pi(\phi)))$, and $Y^F \in \Gamma(VE)$ is the fundamental vector field for Y. It suffices now to consider three cases: either η and ξ are both horizontal, or both vertical, or one of each.

Case 1: Both horizontal. Given $X, Y \in \mathfrak{X}(M)$, A vanishes on both X^h and Y^h, thus the term involving the Lie bracket on $\mathfrak{g}$ vanishes, and

$$\begin{aligned} dA(X^h, Y^h) &= \mathcal{L}_{X^h}\left(A(Y^h)\right) - \mathcal{L}_{Y^h}\left(A(X^h)\right) - A([X^h, Y^h]) = -A([H(X^h), H(Y^h)]) \\ &= F_A(X^h, Y^h). \end{aligned}$$

Case 2: Both vertical. Given $X, Y \in \mathfrak{g}$, the $\mathfrak{g}$-valued functions $A(X^F) = X$ and $A(Y^F) = Y$ are constant, and $F_A(X^F, Y^F)$ vanishes, so we find

$$\begin{aligned} dA(X^F, Y^F) &= \mathcal{L}_{X^F}\left(A(Y^F)\right) - \mathcal{L}_{Y^F}\left(A(X^F)\right) - A([X^F, Y^F]) = -A([X, Y]^F) = -[X, Y] \\ &= -[A(X^F), A(Y^F)], \end{aligned}$$

where we are using the fact from Theorem 40.8 and Exercise 40.10 that $\mathfrak{g} \to \mathfrak{X}(E) : X \mapsto X^F$ is a Lie algebra homomorphism.

Case 3: One horizontal and one vertical. Given $X \in \mathfrak{X}(M)$ and $Y \in \mathfrak{g}$, we claim first that the flows of $X^h \in \mathfrak{X}(E)$ and $Y^F \in \mathfrak{X}(E)$ commute, implying $[X^h, Y^F] = 0$. To see this, pick $p \in M$ and $\phi \in E_p$, and let $\gamma(t) := \varphi^t_X(p)$, so $\sigma(t) := \varphi^t_{X^h}(\phi)$ is the unique horizontal lift of γ with $\sigma(0) = \phi$. Since parallel transport is G-equivariant, the section $t \mapsto \sigma(t)g$ is also a horizontal lift of γ for each $g \in G$, and thus satisfies $\sigma(t)g = \varphi^t_{X^h}(\phi g)$. Now using the obvious analogue of Proposition 40.7 for right group actions to write down the flow of Y^F, we find

$$\varphi^s_{Y^F} \circ \varphi^t_{X^h}(\phi) = \sigma(t)\exp(sY) = \varphi^t_{X^h}(\phi\exp(sY)) = \varphi^t_{X^h} \circ \varphi^s_{Y^F}(\phi),$$

thus proving the claim. Plugging X^h and Y^F into F_A gives 0 since Y^F is vertical, and the term with the Lie bracket on $\mathfrak{g}$ also vanishes since $A(X^h) = 0$, so we are left with

$$dA(X^h, Y^F) = \mathcal{L}_{X^h}\left(A(Y^F)\right) - \mathcal{L}_{Y^F}\left(A(X^h)\right) - A([X^h, Y^F]) = \mathcal{L}_{X^h}(Y) - \mathcal{L}_{Y^F}(0) - A(0) = 0.$$

This completes the proof of the first identity.

[99]Like the second fundamental form, the term "second structural equation" is a remnant of the way that concepts in differential geometry were first presented by the authors who developed the subject; in this case, Henri Cartan. In case you'd wondered, the *first* structural equation (see Exercise 46.16) is defined in the context of the frame bundle of a tangent bundle, and is essentially a translation of the definition of the torsion tensor into the language of connection 1-forms. We will see in the next lecture that the second structural equation encodes the relationship between the curvature 2-form and the Riemann tensor on any associated vector bundle.

Now that the structural equation is established, the Bianchi identity will follow if we can prove

$$\frac{1}{2}d[A, A] = [F_A, A],$$

since $d^2A = 0$. We could prove this by another direct computation as carried out above,[100] but it is much easier to apply the graded Leibniz rule for d with respect to the bracket $[\ ,\ ]$ and then replace dA with $F_A - \frac{1}{2}[A, A]$: we obtain

$$\frac{1}{2}d[A, A] = \frac{1}{2}[dA, A] - \frac{1}{2}[A, dA] = [dA, A] = [F_A, A] - \frac{1}{2}[[A, A], A] = [F_A, A],$$

where the term $[[A, A], A]$ vanishes due to the graded Jacobi identity. □

46. Curvature on associated vector bundles

In this lecture we shall finally wind the discussion back around to vector bundles, but at a greater level of generality than we have considered before: our vector bundles will now have arbitrary structure groups G, acting smoothly and linearly but not necessarily effectively on the model fiber $\mathbb{F}^m$. Following the philosophy outlined in §43.3, every such bundle is isomorphic to an associated bundle $E^\rho = E \times_\rho V \to M$ for some principal G-bundle $\pi : E \to M$ and a linear group representation $\rho : G \to \mathrm{GL}(V)$ on some vector space V, and the connection we consider on $E^\rho \to M$ will always be determined by a choice of principal connection on $E \to M$.

46.1. Connections on associated bundles. Let us first complete the discussion of §43.3 by clarifying how a principal connection $HE \subset TE$ on a principal G-bundle $\pi : E \to M$ determines G-compatible connections on each of its associated bundles. Assume $\rho : G \times F \to F$ is a smooth group action and write $E^\rho = E \times_\rho F$ as in §43.2. If the parallel transport $P^t_\gamma : E_{\gamma(0)} \to E_{\gamma(t)}$ along a path $\gamma(t) \in M$ through $p := \gamma(0)$ is given, the associated diffeomorphisms $P^t_\gamma : E^\rho_{\gamma(0)} \to E^\rho_{\gamma(t)}$ need to satisfy

$$P^t_\gamma([\phi, x]) = [P^t_\gamma(\phi), x]$$

for all $\phi \in E_p$ and $x \in F$, which is well defined because $P^t_\gamma : E_{\gamma(0)} \to E_{\gamma(t)}$ is G-equivariant. This dictates defining the horizontal lift maps $\mathrm{Hor}_{[\phi,x]} : T_pM \to T_{[\phi,x]}E^\rho$ in terms of the corresponding maps $\mathrm{Hor}_\phi : T_pM \to T_\phi E$ by

$$\mathrm{Hor}_{[\phi,x]}(X) := [\mathrm{Hor}_\phi(X), 0] \in T_{[\phi,x]}E^\rho = T_{(\phi,x)}(E \times F)\Big/T_{(\phi,x)}\left(G \cdot (\phi, x)\right),$$

where we are using Exercise 40.27 to identify tangent spaces of the quotient $E^\rho = (E \times F)/G$ with quotient vector spaces. The tangent space at (ϕ, x) to its G-orbit is spanned by pairs of fundamental vector fields $(-X^F(\phi), X^F(x)) \in T_\phi E \times T_x F$ for $X \in \mathfrak{g}$, and since $X^F(\phi)$ is a nontrivial vertical vector for every $X \neq 0$, it follows that the image of the map $T_pM \to T_{(\phi,x)}(E \times F) : X \mapsto (\mathrm{Hor}_\phi(X), 0)$ intersects this subspace trivially, and therefore has an injective projection to the quotient. Moreover, the vertical subspace $V_{[\phi,x]}E^\rho$ is represented by all tangent vectors in $T_{(\phi,x)}(E \times F)$ that are equivalent in the quotient to vectors of the form $(0, Y) \in T_\phi E \times T_x F$, and these will always be vertical in the first component, implying that the image of $\mathrm{Hor}_{[\phi,x]} : T_pM \to T_{[\phi,x]}E^\rho$ intersects the vertical subspace trivially. In other words, there is a connection on $E^\rho \to M$ defined by

$$H_{[\phi,x]}E^\rho := \mathrm{im}\,\mathrm{Hor}_{[\phi,x]} \subset T_{[\phi,x]}E^\rho,$$

for which the induced parallel transport maps are as described above and are thus compatible with the G-structure. By construction, the G-compatible connection on E^ρ is then flat whenever the

[100] An earlier version of these notes actually contained that computation, which occupied about a page, but I later realized that it wasn't necessary.

principal connection on E is flat, since parallel local sections of E^ρ can be constructed explicitly out of parallel local sections of E. The converse will also hold if the G-action on F is effective, since in this case the parallel transport on E^ρ also uniquely determines the parallel transport on E.

We will not need it, but just for completeness, here is an explicit formula for the connection map $K : TE^\rho \to VE^\rho$ induced by a principal connection on $E \to M$ with connection 1-form $A \in \Omega^1(E, \mathfrak{g})$: identifying $T_{[\phi,x]}E^\rho$ with a quotient vector space as described above and $V_{[\phi,x]}E^\rho$ with T_xF,

$$K_{[\phi,x]}([\eta, \xi]) = \xi + A(\eta)^F(x) \in T_xF = V_{[\phi,x]}E^\rho.$$

We leave it as an exercise to verify that this map is well defined and is the fiberwise-linear projection of TE^ρ to VE^ρ along the horizontal subbundle HE^ρ defined above.

46.2. The linear case. For the remainder of this lecture, we shall assume the standard fiber F of our associated bundle is a finite-dimensional (real or complex) vector space V, and G acts on it linearly via a representation

$$\rho : G \to \mathrm{GL}(V),$$

so that $E^\rho = E \times_\rho V$ is a vector bundle. Recall that the derivative of ρ at $e \in G$ also gives us a Lie algebra representation

$$\rho_* : \mathfrak{g} \to \mathfrak{gl}(V) = \mathrm{End}(V),$$

so elements $X \in \mathfrak{g}$ can also be said to act on V via (not necessarily invertible) linear maps $v \mapsto \rho_*(X)v$, such that $\rho_*([X,Y]) = \rho_*(X)\rho_*(Y) - \rho_*(Y)\rho_*(X)$. The parallel transport maps defined on E^ρ via a principal connection on E are now linear, and additionally preserve whatever structure on the fibers is dictated by the structure group G. In light of the canonical isomorphisms $V_vE^\rho = T_v(E^\rho_p) = E^\rho_p$ for each $v \in E^\rho_p$, it is appropriate in this setting to regard the covariant derivative $\nabla_t s(t)$ of a section $s(t) \in E^\rho_{\gamma(t)}$ along a path $\gamma(t) \in M$ as taking values in the fibers $E^\rho_{\gamma(t)}$, and one then checks that for any smooth function $f(t) \in \mathbb{R}$, there is a Leibniz rule:

$$\nabla_t(fs)(0) = \frac{d}{dt}P_\gamma^{-1}(f(t)s(t))\bigg|_{t=0} = \frac{d}{dt}f(t)P_\gamma^{-1}(s(t))\bigg|_{t=0} = \dot{f}(0)s(0) + f(0)\nabla_t s(0).$$

The covariant derivatives of sections thus define a linear operator $\nabla : \Gamma(E^\rho) \to \Gamma(\mathrm{Hom}(TM, E^\rho))$ satisfying the Leibniz rule $\nabla(f\eta) = df(\cdot)\eta + f\,\nabla\eta$, and so we are back to the original definition of connections on a vector bundle given in Lecture 32.

We can relate the covariant derivative operator on E^ρ to the principal connection on E in the following manner. For every section $\eta \in \Gamma(E^\rho)$, there is an associated function $\widehat{\eta} : E \to V$ defined via the condition

$$\eta(p) = [\phi, \widehat{\eta}(\phi)] \in E^\rho_p \qquad \text{for all } p \in M \text{ and } \phi \in E_p. \tag{46.1}$$

Not every smooth function $\widehat{\eta} : E \to V$ corresponds to a section of E^ρ in this way; the right hand side is independent of the choice of $\phi \in E_p$ if and only if $\widehat{\eta}$ satisfies the condition

$$\rho(g) \circ \widehat{\eta}(\phi g) = \widehat{\eta}(\phi) \qquad \text{or equivalently} \qquad \widehat{\eta} \circ R_g = \rho(g^{-1}) \circ \widehat{\eta}$$

for all $g \in G$. Interpreting R_g and $\rho(g^{-1})$ as right G-actions on E and V respectively, we will call a function $\widehat{\eta} : E \to V$ **ρ-equivariant** whenever it satisfies the condition $\widehat{\eta} \circ R_g = \rho(g^{-1}) \circ \widehat{\eta}$ for all $g \in G$; if preferred, one could instead frame it in terms of left G-actions by rewriting the condition as $\widehat{\eta} \circ R_{g^{-1}} = \rho(g) \circ \widehat{\eta}$. In any case, we see that there is a natural bijective correspondence between sections of E^ρ and ρ-equivariant functions $E \to V$. Given $p \in M$, $X \in T_pM$ and a section $\eta \in \Gamma(E^\rho)$, let us now choose a path $\gamma(t) \in M$ through $\gamma(0) = p$ with $\dot{\gamma}(0) = X$ and a parallel section $\phi(t) \in E_{\gamma(t)}$ of E along γ. By the definition of parallel transport on E^ρ, we then have

$$P_\gamma^t([\phi(0), v]) = [\phi(t), v]$$

for all $v \in V$, thus

$$\nabla_X \eta = \nabla_t \eta(\gamma(t))|_{t=0} = \frac{d}{dt}(P_\gamma^t)^{-1}\big(\eta(\gamma(t))\big)\Big|_{t=0} = \frac{d}{dt}[\phi(0), \hat{\eta}(\phi(t))]\Big|_{t=0}.$$

The derivative of the path $\hat{\eta}(\phi(t)) \in V$ at $t = 0$ is $d\hat{\eta}(\mathrm{Hor}_{\phi(0)}(X))$ since $\phi(t)$ is a horizontal lift of $\gamma(t)$, thus we obtain the succinct formula

$$\nabla_X \eta = [\phi, d\hat{\eta}(\mathrm{Hor}_\phi(X))] \in E_p^\rho \qquad \text{for any } p \in M,\ X \in T_pM,\ \phi \in E_p. \tag{46.2}$$

In particular, a section $\eta \in \Gamma(E^\rho)$ is parallel if and only if the corresponding function $\hat{\eta} : E \to V$ has vanishing derivatives in all horizontal directions.

Recall from Exercise 45.9 that covariant exterior derivatives satisfy Leibniz rules with respect to smooth multilinear bundle maps that are parallel. The following result serves as a useful source of parallel bundle maps.

Proposition 46.1. *Assume $\pi : E \to M$ is a principal G-bundle, $\rho : G \to \mathrm{GL}(V)$ and $\tau : G \to \mathrm{GL}(W)$ are two linear group representations on finite-dimensional vector spaces V and W respectively, and $\psi : V \to W$ is a G-equivariant linear map, i.e. it satisfies $\psi \circ \rho(g) = \tau(g) \circ \psi$ for all $g \in G$. Then there is a smooth linear bundle map $\Psi : E^\rho \to E^\tau$ defined by*

$$\Psi([\phi, v]) = [\phi, \psi(v)],$$

and for any choice of principal connection on $E \to M$, $\Psi \in \Gamma(\mathrm{Hom}(E^\rho, E^\tau))$ is parallel with respect to the associated connections on E^ρ and E^τ.

Proof. The map $\Psi : E^\rho \to E^\tau$ is well defined due to the assumption that $\psi : V \to W$ is equivariant, and it is linear on each fiber because ψ is linear. Given a principal connection on $E \to M$ and the associated linear connections on E^ρ and E^τ, it is also clear from the definitions that Ψ has the following property: if $s(t) \in E^\rho_{\gamma(t)}$ is a parallel section of E^ρ along a path $\gamma(t) \in M$, then $\Psi \circ s(t) \in E^\tau_{\gamma(t)}$ is a parallel section of E^τ along γ. This follows because both can be written in terms of a parallel section of E along γ, paired with constant vectors $v \in V$ and $\psi(v) \in W$ respectively. Since any section along a path can be expressed pointwise as a linear combination of parallel sections along that path, this is enough information to deduce $\nabla\Psi \equiv 0$ from the Leibniz rule; we leave the details as an exercise. □

Exercise 46.2. Given two representations $\rho : G \to \mathrm{GL}(V)$ and $\tau : G \to \mathrm{GL}(W)$, the natural representation $\mathrm{Hom}(\rho, \tau)$ induced on $\mathrm{Hom}(V, W)$ is defined via the linear G-action

$$gA := \tau(g) \circ A \circ \rho(g)^{-1} \qquad \text{for } A \in \mathrm{Hom}(V, W).$$

Show that for any principal G-bundle $E \to M$, there is a natural vector bundle isomorphism between the associated bundle $E^{\mathrm{Hom}(\rho,\tau)}$ and $\mathrm{Hom}(E^\rho, E^\tau)$, and it is parallel for any choice of principal connection on E. Under this isomorphism, what kind of equivariant function $E \to \mathrm{Hom}(V, W)$ does the parallel bundle map $E^\rho \to E^\tau$ in Proposition 46.1 correspond to?

46.3. Equivariant exterior algebra. Recall that for a principal connection, the connection 1-form $A \in \Omega^1(E, \mathfrak{g})$ is G-equivariant in the sense that

$$R_g^* A = \mathrm{Ad}_{g^{-1}} \circ A \qquad \text{for all } g \in G.$$

For the curvature 2-form $F_A \in \Omega^2(E, \mathfrak{g})$, it is an easy exercise using its definition $F_A(\xi, \eta) = -A([H\xi, H\eta])$ to show that F_A satisfies the same equivariance property, and additionally, that $F_A(\xi, \eta) = 0$ whenever either of ξ or η is vertical. These properties suggest that instead of viewing F_A as a 2-form defined on the total space E, we really ought to let it descend to the quotient $E/G \cong M$, analogously to the way that ρ-equivariant vector-valued functions $\hat{\eta} : E \to V$ are equivalent to sections $\eta : M \to E^\rho$ of an associated vector bundle over M. The resulting formalism for equivariant differential forms turns out to provide an elegant language for proving results about connections and curvature without the need for local trivializations.

Assume as usual that $\rho : G \to \mathrm{GL}(V)$ is a linear representation of G on some finite-dimensional vector space V. If a principal connection on $E \to M$ has been chosen, then we can use a mild generalization of the relation in §46.2 to associate an equivariant V-valued k-form on E to any bundle-valued form in $\Omega^k(M, E^\rho)$, i.e. given $\omega \in \Omega^k(M, E^\rho)$, we claim there is a canonical choice of $\hat{\omega} \in \Omega^k(E, V)$ satisfying the relation

$$\omega_p(X_1, \ldots, X_k) = [\phi, \hat{\omega}_\phi(\mathrm{Hor}_\phi(X_1), \ldots, \mathrm{Hor}_\phi(X_k))] \qquad p \in M,\ \phi \in E_p,\ X_1, \ldots, X_k \in T_pM. \tag{46.3}$$

The right hand side determines $\hat{\omega}$ only on k-tuples of horizontal vectors, so we are not claiming that there is a *unique* $\hat{\omega} \in \Omega^k(E, V)$ satisfying this relation, but it will indeed become unique if we impose the additional condition that $\hat{\omega}$ should be a **horizontal** k-form, meaning

$$H^*\hat{\omega} = \hat{\omega}$$

for the fiberwise-linear projection $H : TE \to HE$ along VE. This is equivalent to the condition that $\hat{\omega}(X_1, \ldots, X_k)$ must vanish whenever any of the vectors $X_1, \ldots, X_k$ is vertical, and if we require this in (46.3), then $\hat{\omega}$ is determined uniquely. Moreover, the fact that $\phi \in E_p$ does not appear on the left hand side forces $\hat{\omega}$ to satisfy an equivariance condition, namely

$$\rho(g) \circ R_g^*\hat{\omega} = \hat{\omega} \qquad \text{or equivalently} \qquad R_g^*\hat{\omega} = \rho(g^{-1}) \circ \hat{\omega}$$

for all $g \in G$. We shall refer to the k-forms satisfying this condition as **ρ-equivariant**. Whenever $\hat{\omega} \in \Omega^k(E, V)$ is ρ-equivariant, there is a unique $\omega \in \Omega^k(M, E^\rho)$ for which (46.3) is satisfied. We've proved:

THEOREM 46.3. *For any principal G-bundle $\pi : E \to M$ with a principal connection and any linear representation $\rho : G \to \mathrm{GL}(V)$, the relation (46.3) defines a natural isomorphism for each $k \geqslant 0$ between the space of $\Omega^k(M, E^\rho)$ of smooth E^ρ-valued k-forms on M and the space*

$$\Omega^k_\rho(E, V) \subset \Omega^k(E, V)$$

consisting of all smooth V-valued k-forms on E that are horizontal and ρ-equivariant. □

EXAMPLE 46.4. The curvature 2-form $F_A \in \Omega^2(E, \mathfrak{g})$ for a principal connection is horizontal and Ad-equivariant. The connection 1-form $A \in \Omega^1(E, \mathfrak{g})$ is also Ad-invariant, though not horizontal; however, the difference between any two connection 1-forms on a principal G-bundle is both Ad-invariant and horizontal.

The isomorphism $\Omega^*(M, E^\rho) \cong \Omega^*_\rho(E, V)$ identifies the covariant derivative operator $\nabla : \Gamma(E^\rho) = \Omega^0(M, E^\rho) \to \Omega^1(M, E^\rho)$ with a linear map

$$\Omega^0_\rho(E, V) \xrightarrow{d_A} \Omega^1_\rho(E, V).$$

It turns out that d_A can be expressed via a simple formula in terms of the differential $d : \Omega^0(E, V) \to \Omega^1(E, V)$ and the connection 1-form $A \in \Omega^1(E, \mathfrak{g})$. For each $p \in M$, $X \in T_pM$, $\phi \in E_p$ and $\eta \in \Gamma(E^\rho)$, combining (46.3) with (46.2) gives

$$\nabla_X \eta = [\phi, d_A\hat{\eta}(\mathrm{Hor}_\phi(X))] = [\phi, d\hat{\eta}(\mathrm{Hor}_\phi(X))],$$

so $d_A\hat{\eta} \in \Omega^1_\rho(E, V)$ is the unique horizontal 1-form that matches $d\hat{\eta}$ in horizontal directions. In other words, d_A is the restriction to $\Omega^0_\rho(E, V) \subset \Omega^0(E, V) = C^\infty(E, V)$ of the operator

$$d_A : \Omega^0(E, V) \to \Omega^1(E, V), \qquad d_A f := df \circ H, \tag{46.4}$$

which is also sometimes called the **covariant derivative** for functions on E. In order to write d_A more explicitly, let us assume again that $f : E \to V$ is ρ-equivariant and use the connection 1-form $A \in \Omega^1(E, \mathfrak{g})$ as a vertical projection: then for $\phi \in E_p$ and $\xi \in T_\phi E$,

$$\begin{aligned} d_A f(\xi) &= df(H\xi) = df(\xi - A(\xi)^F(\phi)) = df(\xi) - df(A(\xi)^F(\phi)) \\ &= df(\xi) - \frac{d}{dt} f\left(\phi \exp(tA(\xi))\right)\Big|_{t=0} = df(\xi) - \frac{d}{dt}\rho(\exp(-tA(\xi)))f(\phi)\Big|_{t=0} \\ &= df(\xi) + \rho_*(A(\xi))f(\phi). \end{aligned}$$

The last term in this expression can be interpreted as a wedge product of vector-valued forms in the spirit of Exercise 45.7(f): the Lie algebra representation $\rho_* : \mathfrak{g} \to \mathfrak{gl}(V)$ determines a bilinear map $\mathfrak{g} \times V \to V : (X, v) \mapsto \rho_*(X)v$ and thus a wedge product

$$\Omega^*(E, \mathfrak{g}) \times \Omega^*(E, V) \xrightarrow{\wedge} \Omega^*(E, V),$$

with which the formula above can be written as

$$d_A f = df + A \wedge f \qquad \text{for } f \in \Omega^0_\rho(E, V). \tag{46.5}$$

It should be stressed that this is not a valid formula for $d_A f$ on arbitrary functions $f : E \to V$, but only on those that are ρ-equivariant—which is the case we care about.

Like the covariant derivative on E^ρ, the operator d_A extends naturally to a **covariant exterior derivative**

$$d_A : \Omega^k_\rho(E, V) \to \Omega^{k+1}_\rho(E, V)$$

for each $k \geqslant 0$ that is equivalent via Theorem 46.3 to the operator $d_\nabla : \Omega^k(M, E^\rho) \to \Omega^{k+1}(M, E^\rho)$. We can extrapolate from the case $k = 0$ above to guess two formulas for d_A: the first is

$$d_A\omega(\xi_0, \ldots, \xi_k) := d\omega(H\xi_0, \ldots, H\xi_k),$$

which defines more generally a linear operator $d_A : \Omega^k(E, V) \to \Omega^{k+1}(E, V)$ on all (not just equivariant) V-valued forms. Using the fact that $TR_g(HE) = HE$, one verifies easily that $d_A\omega$ is ρ-equivariant whenever ω is, and since it is manifestly also horizontal, d_A preserves the subspace $\Omega^*_\rho(E, V)$. To see that $d_A : \Omega^k_\rho(E, V) \to \Omega^{k+1}_\rho(E, V)$ really is equivalent to $d_\nabla : \Omega^k(M, E^\rho) \to \Omega^{k+1}(M, E^\rho)$ beyond the case $k = 0$, it suffices to establish that d_A satisfies a corresponding Leibniz rule. Recall that the Leibniz rule for d_∇ is based on the wedge product of bundle-valued forms on M with real-valued forms. One can interpret $\Omega^*(M)$ as $\Omega^*(M, E^{\text{triv}})$ for the trivial representation $\text{triv} : G \to \text{GL}(\mathbb{R}) : g \mapsto \mathbb{1}$ acting on $\mathbb{R}$, as the associated vector bundle for this representation is just the trivial real line bundle over M. Under the correspondence of Theorem 46.3, $\Omega^*(M)$ is thus identified with $\Omega^*_{\text{triv}}(E, \mathbb{R})$, the space of real-valued horizontal forms $\omega \in \Omega^*(E)$ that are *G-invariant*, meaning they satisfy

$$R_g^*\omega = \omega \qquad \text{for all } g \in G.$$

Exercise 46.5. For any $\alpha \in \Omega^k(M)$ and $\beta \in \Omega^\ell(M, E^\rho)$, using Theorem 46.3 to write $\hat{\alpha} \in \Omega^k_{\text{triv}}(E, \mathbb{R})$ and $\hat{\beta} \in \Omega^\ell_\rho(E, V)$, prove that $\widehat{\alpha \wedge \beta} = \hat{\alpha} \wedge \hat{\beta} \in \Omega^{k+\ell}(E, V)$.

Exercise 46.6. Prove that for $\omega \in \Omega^k_{\text{triv}}(E, \mathbb{R})$, $d_A\omega = d\omega$.
Hint: One only really needs to check this for $k = 0$, in which case the horizontality of α is a vacuous condition, though G-invariance is important!

Exercise 46.7. Prove that if $\alpha \in \Omega^k_{\text{triv}}(E, \mathbb{R})$ and $\beta \in \Omega^\ell(E, V)$ is horizontal, then

$$d_A(\alpha \wedge \beta) = d\alpha \wedge \beta + (-1)^{|\alpha|}\alpha \wedge d_A\beta.$$

This applies in particular whenever $\beta \in \Omega^\ell_\rho(E, V)$.

The second formula we can guess for d_A generalizes (46.5): define

$$D_A : \Omega^k(E, V) \to \Omega^{k+1}(E, V) : \omega \mapsto d\omega + A \wedge \omega,$$

where again the bilinear map $\mathfrak{g} \times V \to V$ arising from the Lie algebra representation $\rho_* : \mathfrak{g} \to \mathfrak{gl}(V)$ is used for defining the product $A \wedge \omega$. The appropriate Leibniz rule for this operator is easy to prove: it takes the form

$$D_A(\alpha \wedge \beta) = d\alpha \wedge \beta + (-1)^{|\alpha|}\alpha \wedge D_A\beta \qquad \text{for } \alpha \in \Omega^*(E),\ \beta \in \Omega^*(E, V),$$

and applies in particular whenever $\alpha \in \Omega^*_{\text{triv}}(E, \mathbb{R})$ and $\beta \in \Omega^*_\rho(E, V)$. Since $D_A = d_A$ on $\Omega^0_\rho(E, V)$, it follows that the two operators also match on $\Omega^k_\rho(E, V)$ for all $k \geqslant 0$. Here is a summary:

THEOREM 46.8. *The operator $d_A = H^*d : \Omega^*(E, V) \to \Omega^*(E, V)$ preserves the space $\Omega^*_\rho(E, V) \subset \Omega^*(E, V)$ of horizontal ρ-equivariant forms, and satisfies*

$$d_A\omega = d\omega + A \wedge \omega$$

on this subspace. Moreover, it is equivalent under the isomorphism of Theorem 46.3 to the covariant exterior derivative $d_\nabla : \Omega^*(M, E^\rho) \to \Omega^*(M, E^\rho)$. □

46.4. The Riemann tensor revisited. It is time to get some real mileage out of the second structural equation $F_A = dA + \frac{1}{2}[A, A]$. We saw in §45.3 that on the associated vector bundle $E^\rho \to M$, the Riemann tensor

$$R : TM \oplus TM \oplus E^\rho \to E^\rho : (X, Y, \eta) \mapsto R(X, Y)\eta$$

appears if one iterates the covariant exterior derivative on $\Omega^0(M, E^\rho)$, i.e.

$$d^2_\nabla\eta(X, Y) = R(X, Y)\eta \qquad \text{for } \eta \in \Gamma(E^\rho).$$

By Theorem 46.8, this operator is equivalent to

$$d^2_A : \Omega^0_\rho(E, V) \to \Omega^2_\rho(E, V).$$

More generally, the formula $d_A\omega = d\omega + A \wedge \omega$ enables us to compute $d^2_A : \Omega^k_\rho(E, V) \to \Omega^{k+2}_\rho(E, V)$ in terms of the curvature 2-form: using the second structural equation and Exercise 45.7(f), we find

$$\begin{aligned} d^2_A\omega &= d(d\omega + A \wedge \omega) + A \wedge (d\omega + A \wedge \omega) = d(A \wedge \omega) + A \wedge d\omega + A \wedge (A \wedge \omega) \\ &= dA \wedge \omega - A \wedge d\omega + A \wedge d\omega + \frac{1}{2}(A \wedge (A \wedge \omega) - (-1)^{|A|\cdot|A|}A \wedge (A \wedge \omega)) \\ &= dA \wedge \omega + \frac{1}{2}[A, A] \wedge \omega = \left(dA + \frac{1}{2}[A, A]\right) \wedge \omega. \end{aligned}$$

In light of Theorem 45.11, this proves:

THEOREM 46.9. *For $\omega \in \Omega^*_\rho(E, V)$, $d^2_A\omega = F_A \wedge \omega$.* □

To understand what this means about the Riemann tensor, we can translate it into a statement about bundle-valued forms on M. Since $F_A \in \Omega^2_{\text{Ad}}(E, \mathfrak{g})$, the curvature form is equivalent via Theorem 46.3 to a 2-form

$$\Omega_A \in \Omega^2(M, \text{Ad}(E))$$

taking values in the so-called **adjoint bundle**, defined as the associated vector bundle $\text{Ad}(E) = E \times_{\text{Ad}} \mathfrak{g}$ whose standard fiber is the Lie algebra $\mathfrak{g}$, with G-valued transition functions acting on it via the adjoint representation. The next exercise shows that the Lie algebra homomorphism $\rho_* : \mathfrak{g} \to \mathfrak{gl}(V) = \text{End}(V)$ determines a natural smooth (and parallel) linear bundle map $\text{Ad}(E) \to \text{End}(E^\rho)$ whose kernel is a linear subbundle of $\text{Ad}(E)$; in particular, each element of the fiber $\text{Ad}(E)_p$ over a point $p \in M$ defines a linear map $E^\rho_p \to E^\rho_p$.

EXERCISE 46.10. Given the representation $\rho : G \to \mathrm{GL}(V)$, the induced representation $\mathrm{End}(\rho)$ of G on $\mathrm{End}(V) = \mathrm{Hom}(V, V)$ is defined via the G-action

$$g \cdot A := \rho(g) \circ A \circ \rho(g^{-1}) \in \mathrm{End}(V) \qquad \text{for } g \in G,\ A \in \mathrm{End}(V).$$

(a) Show that the linear map $\rho_* : \mathfrak{g} \to \mathfrak{gl}(V) = \mathrm{End}(V)$ is G-equivariant with respect to the representations Ad on $\mathfrak{g}$ and $\mathrm{End}(\rho)$ on $\mathrm{End}(V)$.

(b) Find a natural parallel bundle isomorphism from the associated vector bundle $E^{\mathrm{End}(\rho)}$ to $\mathrm{End}(E^\rho) = \mathrm{Hom}(E^\rho, E^\rho)$.

It follows now from Proposition 46.1 that ρ_* induces a natural parallel bundle map $\mathrm{Ad}(E) \to E^{\mathrm{End}(\rho)}$, which we can regard as a parallel bundle map $\mathrm{Ad}(E) \to \mathrm{End}(E^\rho)$ due to the isomorphism in part (b).

(c) Show that the kernel of the natural bundle map $\mathrm{Ad}(E) \to \mathrm{End}(E^\rho)$ is a smooth subbundle of $\mathrm{Ad}(E)$ whose fibers are all isomorphic to the kernel of $\rho_* : \mathfrak{g} \to \mathrm{End}(V)$.

(d) Show that every fiber of $\mathrm{Ad}(E)$ has a natural Lie algebra structure such that the bundle map $\mathrm{Ad}(E) \to \mathrm{End}(E^\rho)$ defines a Lie algebra homomorphism on each fiber, where the bracket on fibers of $\mathrm{End}(E^\rho)$ is the commutator.

In light of the natural bundle map $\mathrm{Ad}(E) \to \mathrm{End}(E^\rho)$ and the obvious fiberwise-bilinear pairing $\mathrm{End}(E^\rho) \oplus E^\rho \to E^\rho$ defined via evaluation, there is now a natural wedge product

$$\Omega^k(M, \mathrm{Ad}(E)) \times \Omega^\ell(M, E^\rho) \xrightarrow{\wedge} \Omega^{k+\ell}(M, E^\rho),$$

and Theorem 46.9 translates into the relation

$$d_\nabla^2 \omega = \Omega_A \wedge \omega \qquad \text{for } \omega \in \Omega^*(M, E^\rho). \tag{46.6}$$

Applied to a section $\eta \in \Gamma(E^\rho)$, this expresses the Riemann tensor as

$$R(\cdot, \cdot)\eta = \Omega_A \wedge \eta \in \Omega^2(M, E^\rho).$$

Clearly, R must vanish if the connection on E is flat. It is possible for the converse to be false, depending on the Lie algebra representation $\rho_* : \mathfrak{g} \to \mathfrak{gl}(V)$; indeed, $\Omega_A \wedge \eta$ can vanish for all η without Ω_A itself being 0, but this happens if and only if the values taken by Ω_A all lie in the subbundle of $\mathrm{Ad}(E)$ defined as the kernel of the natural bundle map $\mathrm{Ad}(E) \to \mathrm{End}(E^\rho)$ described in Exercise 46.10. The rank of this subbundle is the dimension of the kernel of the Lie algebra representation $\rho_* : \mathfrak{g} \to \mathfrak{gl}(V)$, so we obtain an especially strong condition if the latter is injective, which is true in many of the situations we care about, e.g. it is automatic if the original representation $\rho : G \to \mathrm{GL}(V)$ is faithful, meaning that G acts effectively on V. We conclude:

THEOREM 46.11. *For the connection ∇ on the associated vector bundle $E^\rho \to M$ determined by a principal connection $A \in \Omega^1(E, \mathfrak{g})$ on $\pi : E \to M$, the following conditions are equivalent:*

(i) The Riemann tensor of ∇ vanishes;

(ii) The equivariant curvature 2-form $F_A \in \Omega^2(E, \mathfrak{g})$ takes all its values in the kernel of $\rho_* : \mathfrak{g} \to \mathfrak{gl}(V)$;

(iii) The bundle-valued curvature 2-form $\Omega_A \in \Omega^2(M, \mathrm{Ad}(E))$ takes all its values in the kernel of the natural bundle map $\mathrm{Ad}(E) \to \mathrm{End}(E^\rho)$ described in Exercise 46.10.

□

COROLLARY 46.12. *If the principal connection on $E \to M$ is flat, then the Riemann tensor on $E^\rho \to M$ vanishes, and the converse is also true if the Lie algebra representation $\rho_* : \mathfrak{g} \to \mathfrak{gl}(V)$ is injective; it holds in particular if the representation $\rho : G \to \mathrm{GL}(V)$ is faithful.* □

COROLLARY 46.13. *A connection on a vector bundle is flat if and only if its Riemann tensor vanishes.*

PROOF. The definition of the Riemann tensor does not depend on the structure group of the vector bundle, so given an m-plane bundle over $\mathbb{F}$, we are free to take $\mathrm{GL}(m,\mathbb{F})$ as the structure group and thus view our bundle as $E^{\mathrm{Id}} \to M$ for a principal $\mathrm{GL}(m,\mathbb{F})$-bundle $E \to M$ and the canonical representation $\mathrm{Id} : \mathrm{GL}(m,\mathbb{F}) \to \mathrm{GL}(m,\mathbb{F})$. Since the latter is faithful, vanishing of the Riemann tensor implies via Corollary 46.12 that the connection is flat. □

In the setting of pseudo-Riemannian geometry, the vector bundle in question is a tangent bundle TM of rank n and the structure group is $\mathrm{O}(k,\ell) \subset \mathrm{GL}(n,\mathbb{R})$ for some $k,\ell \geqslant 0$ with $k+\ell = n$, acting on the standard fiber $\mathbb{R}^n$ via its canonical representation. By Corollary 46.12, the Levi-Cività connection is then flat if and only if the Riemann tensor vanishes. In this situation, one can go further and deduce from the symmetry of the connection that any parallel local orthonormal frame $X_1,\dots,X_n$ also satisfies $[X_i,X_j] = 0$ for all i and j, so that it generates a local coordinate chart in which the metric has constant components $g_{ij} = \langle X_i, X_j\rangle$, proving:

COROLLARY 46.14. *A pseudo-Riemannian manifold is locally flat if and only if its Riemann tensor vanishes.* □

EXERCISE 46.15. In what situation can two distinct principal connections on $\pi : E \to M$ determine the same connection on the associated vector bundle E^ρ? Show in particular that this is possible if and only if the Lie algebra representation $\rho_* : \mathfrak{g} \to \mathfrak{gl}(V)$ is not injective.

EXERCISE 46.16. Assume $\pi : E := F^G(TM) \to M$ is the G-frame bundle of the tangent bundle of an n-manifold M, where $TM \to M$ has been equipped with a G-structure for some matrix group $G \subset \mathrm{GL}(n,\mathbb{R})$. Let $\rho : G \to \mathrm{GL}(n,\mathbb{R})$ denote the inclusion, which defines a linear left G-action on $\mathbb{R}^n$ for which TM is isomorphic to the associated vector bundle $E^\rho := (E \times \mathbb{R}^n)/G$. There is a **tautological 1-form**

$$\theta \in \Omega^1(E,\mathbb{R}^n)$$

defined by $\theta_\phi(\xi) := \phi^{-1}(\pi_*\xi)$ for $\xi \in T_\phi E$, where we regard frames $\phi \in E_p$ at points $p \in M$ as vector space isomorphisms $\phi : \mathbb{R}^n \to T_pM$. Given a connection ∇ on TM induced by a choice of principal connection $A \in \Omega^1(E,\mathfrak{g})$ on E, the torsion tensor $T \in \Gamma(T^1_2 M)$ can be interpreted as a bundle-valued 2-form $T \in \Omega^2(M,TM) = \Omega^2(M,E^\rho)$, thus it is naturally equivalent to some ρ-equivariant horizontal 2-form $\tau \in \Omega^2_\rho(E,\mathbb{R}^n)$. The **first structural equation** of Cartan is the relation

$$\tau = d\theta + A \wedge \theta,$$

where the wedge product of $A \in \Omega^1(E,\mathfrak{g})$ with $\theta \in \Omega^1(E,\mathbb{R}^n)$ is defined in terms of the bilinear map $\mathfrak{g} \times \mathbb{R}^n \to \mathbb{R}^n : (X,v) \mapsto \rho_*(X)v$. Prove the equation.
Hint: You can use the same approach that we used to prove the second structural equation in Theorem 45.11, but there is also a much quicker way. Notice that θ is horizontal and ρ-equivariant. What bundle-valued 1-form on M is it equivalent to?

EXERCISE 46.17. In many older or more elementary treatments (including the first semester of this course), connections and curvature on vector bundles are described mainly in terms of locally-defined objects that depend on choices of trivializations, without ever mentioning a principal bundle. This exercise is meant to help you translate between the local picture and the more global perspective that we've adopted in the last few lectures.

Assume $\pi : E \to M$ is a principal G-bundle, with a connection 1-form $A \in \Omega^1(E,\mathfrak{g})$ and curvature 2-form $F \in \Omega^2(E,\mathfrak{g})$, and $\{s_\alpha \in \Gamma(E|_{\mathcal{U}_\alpha})\}_{\alpha\in I}$ is a collection of local sections on open sets $\mathcal{U}_\alpha$ that cover M. For any vector space V and $\omega \in \Omega^k(E,V)$ with $k \geqslant 0$, we can pull back ω via the maps $s_\alpha : \mathcal{U}_\alpha \to E$ to define local V-valued k-forms on M,

$$\omega_\alpha := s_\alpha^*\omega \in \Omega^k(\mathcal{U}_\alpha, V), \qquad \alpha \in I$$

The 1-forms $\{A_\alpha \in \Omega^1(\mathcal{U}_\alpha, \mathfrak{g})\}_{\alpha\in I}$ and 2-forms $\{F_\alpha \in \Omega^2(\mathcal{U}_\alpha, \mathfrak{g})\}_{\alpha\in I}$ are called the **local connection** and **local curvature forms** respectively. Prove:

(a) The connection on $\pi : E \to M$ is uniquely determined by the collection of local connection forms $\{A_\alpha \in \Omega^1(\mathcal{U}_\alpha, \mathfrak{g})\}_{\alpha\in I}$, and its curvature 2-form is similarly determined by the local curvature forms $\{F_\alpha \in \Omega^2(\mathcal{U}_\alpha, \mathfrak{g})\}_{\alpha\in I}$. (Are analogous statements true for *all* forms in $\Omega^*(E, \mathfrak{g})$?)

(b) $F_\alpha = dA_\alpha + \frac{1}{2}[A_\alpha, A_\alpha]$ and $dF_\alpha = [F_\alpha, A_\alpha]$ for each $\alpha \in I$.

Now suppose $\rho : G \to \mathrm{GL}(V)$ is a representation of G on some finite-dimensional vector space V, with induced Lie algebra representation $\rho_* : \mathfrak{g} \to \mathfrak{gl}(V)$, and let $E^\rho = (E \times V)/G \to M$ denote the associated vector bundle, which carries a connection ∇ determined by $A \in \Omega^1(E, \mathfrak{g})$. As shown in §43.2, the local sections $\{s_\alpha \in \Gamma(E|_{\mathcal{U}_\alpha})\}_{\alpha\in I}$ determine a G-bundle atlas $\{\Phi_\alpha : E^\rho|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times V\}_{\alpha\in I}$ for E^ρ, where $\Phi_\alpha^{-1}(p, v) = [s_\alpha(p), v] \in E^\rho_p$ for $p \in \mathcal{U}_\alpha$ and $v \in V$, and the corresponding system of transition functions $g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to G$ is determined by

$$s_\alpha = s_\beta g_{\beta\alpha} \qquad \text{on } \mathcal{U}_\alpha \cap \mathcal{U}_\beta.$$

For each $\omega \in \Omega^k(M, E^\rho)$, $k \geqslant 0$, let $\hat{\omega} \in \Omega^k_\rho(E, V)$ denote the ρ-equivariant horizontal form that corresponds to it under the natural isomorphism $\Omega^k(M, E^\rho) \cong \Omega^k_\rho(E, V)$, and denote $\omega_\alpha := \hat{\omega}_\alpha = s_\alpha^*\hat{\omega} \in \Omega^k(\mathcal{U}_\alpha, V)$ for each $\alpha \in I$. Given $\omega \in \Omega^k(M, E^\rho)$ and $\alpha, \beta \in I$, prove:

(c) $\omega_\alpha \in \Omega^k(\mathcal{U}_\alpha, V)$ is the local representation of ω with respect to the trivialization Φ_α, meaning

$$\Phi_\alpha(\omega(X_1, \ldots, X_k)) = (p, \omega_\alpha(X_1, \ldots, X_k)) \qquad \text{for } X_1, \ldots, X_k \in T_pM,\ p \in \mathcal{U}_\alpha.$$

(d) $\omega_\beta = \rho(g_{\beta\alpha}) \circ \omega_\alpha$ on $\mathcal{U}_\alpha \cap \mathcal{U}_\beta$.

(e) $(d_\nabla\omega)_\alpha = d\omega_\alpha + A_\alpha \wedge \omega_\alpha$, where the wedge product of $A_\alpha \in \Omega^1(\mathcal{U}_\alpha, \mathfrak{g})$ with $\omega_\alpha \in \Omega^k(\mathcal{U}_\alpha, V)$ is defined in terms of the bilinear map $\mathfrak{g} \times V \to V : (X, v) \mapsto \rho_*(X)v$. In particular, for a section $\eta \in \Gamma(E^\rho) = \Omega^0(M, E^\rho)$ and $X \in \mathfrak{X}(\mathcal{U}_\alpha)$, one obtains

$$(\nabla_X\eta)_\alpha = d\eta_\alpha(X) + \rho_*(A_\alpha(X))\eta_\alpha.$$

Finally, prove the following transformation formulas for the local connection and curvature forms: given $p \in \mathcal{U}_\alpha \cap \mathcal{U}_\beta$ and $X, Y \in T_pM$,

(f) $F_\beta(X, Y) = \mathrm{Ad}_{g_{\beta\alpha}(p)} \circ F_\alpha(X, Y)$

(g) $A_\beta(X) = \mathrm{Ad}_{g_{\beta\alpha}(p)} \circ A_\alpha(X) + TL_{g_{\beta\alpha}(p)} \circ Tg_{\alpha\beta}(X)$, where $L_g : G \to G$ denotes left translation $h \mapsto gh$.

In the special case where $G \subset \mathrm{GL}(m, \mathbb{F})$ is a matrix group acting in the obvious way on $V = \mathbb{F}^m$, the transformation formulas of parts (f) and (g) can be written in the simplified form

$$F_\beta = gF_\alpha g^{-1}, \qquad A_\beta = gA_\alpha g^{-1} + g\, dg^{-1},$$

where we abbreviate $g := g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to G$. The second formula is known to physicists as a *gauge transformation*.

47. Chern-Weil theory

The formalism we've developed for connections and curvature on principal bundles has an interesting application that has nothing directly to do with covariant differentiation or local flatness: it can be used to define topological invariants of vector bundles. These invariants are called characteristic classes, and they have been essential tools in topology since at least the middle of the 20th century. One reason for this is that, in a crude sense, characteristic classes make the distinction between topological manifolds and smooth manifolds detectable via the standard methods of algebraic topology; a smooth manifold has a natural vector bundle associated to it, namely its tangent bundle, while a topological manifold does not. Milnor's famous discovery of exotic smooth structures on the 7-sphere, for example, made crucial use of the Pontryagin classes.

The most powerful invariants in mathematics are typically those which admit at least two completely different constructions based on different choices of auxiliary data, sometimes even living in different categories, but which then turn out to be equivalent. The standard characteristic classes are good examples of this phenomenon, as it is possible to construct them entirely within the topological category, where notions of smoothness, connections and curvature cannot be defined, but it is also possible to give a purely smooth construction that depends crucially on those notions. The topological version of the construction belongs in a course on algebraic topology, but in this lecture we will outline the smooth version, which is known as *Chern-Weil theory.*

47.1. A brief review of c_1 for line bundles. The simplest case of Chern-Weil theory appeared in Lecture 30 last semester: it relies on the observation that if $E \to M$ is a smooth complex line bundle, then choosing a bundle metric endows it with the structure group $\mathrm{U}(1)$, which is abelian. We framed the discussion last semester in terms of the *local* connection forms $\{A_\alpha \in \Omega^1(\mathcal{U}_\alpha, \mathfrak{g})\}_{\alpha \in I}$ and curvature forms $\{F_\alpha \in \Omega^2(\mathcal{U}_\alpha, \mathfrak{g})\}_{\alpha \in I}$ associated to a G-bundle atlas, as outlined in Exercise 46.17. Having an abelian structure group simplifies several things: notably, the bracket and adjoint representation on $\mathfrak{g}$ are trivial, so the local version of the second structural equation $F_\alpha = dA_\alpha + \frac{1}{2}[A_\alpha, A_\alpha]$ becomes

$$F_\alpha = dA_\alpha \qquad \text{on } \mathcal{U}_\alpha,$$

and the transformation formula $F_\beta = \mathrm{Ad}_{g_{\beta\alpha}} \circ F_\alpha$ just says

$$F_\beta = F_\alpha \qquad \text{on } \mathcal{U}_\alpha \cap \mathcal{U}_\beta,$$

so that the local curvature forms are restrictions to open sets $\mathcal{U}_\alpha \subset M$ of a single globally defined 2-form

$$F \in \Omega^2(M, \mathfrak{g}).$$

This 2-form is closed since it locally matches dA_α, but it might not be exact, because it can happen that none of the individual local connection 1-forms $A_\alpha \in \Omega^1(\mathcal{U}_\alpha, \mathfrak{g})$ can be extended to a global primitive of F. The issue here is precisely that E might be a nontrivial bundle: if it were trivial, then a global trivialization would provide a global connection 1-form that is a primitive of F. We can thus view the failure of F to be exact as an algebraic measurement of the nontriviality of the bundle.

The distinction between closed and exact k-forms is measured by the de Rham cohomology

$$H^k_{\mathrm{dR}}(M) = \ker\left(\Omega^k(M) \xrightarrow{d} \Omega^{k+1}(M)\right) \Big/ \operatorname{im}\left(\Omega^{k-1}(M) \xrightarrow{d} \Omega^k(M)\right), \qquad k \geqslant 0,$$

whose basic properties were covered in Lecture 13 of last semester's course. The end result of this thought process is a definition of the *first Chern class*

$$c_1(E) := \left[-\frac{1}{2\pi i}F\right] \in H^2_{\mathrm{dR}}(M),$$

where the factor of i has been inserted so that the 2-form $F \in \Omega^2(M, \mathfrak{u}(1))$ with values in $\mathfrak{u}(1) = i\mathbb{R} \subset \mathbb{C}^{1\times 1}$ becomes real-valued, and the usefulness of $\frac{1}{2\pi}$ becomes clear as soon as one does some nontrivial computations (cf. §30.2). From the definition, one might expect $c_1(E)$ to depend on the choice of connection, which was completely arbitrary. But on a $\mathrm{U}(1)$-bundle, the difference $\nabla' - \nabla$ between two connections can be understood as a global $\mathfrak{u}(1)$-valued 1-form $\lambda \in \Omega^1(M, \mathfrak{u}(1))$, and the difference between their curvature 2-forms is $d\lambda$, so while the 2-form $-\frac{1}{2\pi i}F \in \Omega^2(M)$ can be changed drastically by changing the connection, its cohomology class remains the same. A more robust way to make this argument is as follows: since the set of connections forms an affine space, we can linearly interpolate between any two connections ∇, ∇' and use this interpolation to define a connection on the pullback bundle $\Pi^*E \to [0,1] \times M$ via the obvious projection $\Pi : [0,1] \times M \to M$, such that the connection matches ∇ on $\{0\} \times M$ and ∇' on $\{1\} \times M$. The curvature of this connection then defines a closed 2-form on $[0,1] \times M$ whose restriction to $\{0\} \times M$ gives one version of our definition of $c_1(E)$, while its restriction to $\{1\} \times M$ gives the other—by the homotopy-invariance of de Rham cohomology, these two restrictions must give the same cohomology class, because the inclusions $M \hookrightarrow \{0\} \times M$ and $M \hookrightarrow \{1\} \times M$ are smoothly homotopic in $[0,1] \times M$. Once one knows that $c_1(E)$ is independent of the choice of connection, it is not hard to show that one also has $c_1(E) = c_1(E')$ whenever the two bundles $E, E' \to M$ are isomorphic.

As any topologist will tell you, there are good philosophical reasons to seek out a cohomology class as a topological invariant of bundles. This has to do with the notion of *functoriality*: in the language of category theory, cohomology theories $H^*(\cdot)$ define *contravariant functors*, meaning that they associate to suitable spaces X an algebraic object $H^*(X)$ and to suitable maps $f : X \to Y$ between spaces a homomorphism $f^* : H^*(Y) \to H^*(X)$ which goes the *other direction*, and is compatible with compositions in the sense that $(f \circ g)^* = g^* f^*$. Conveniently, there is similarly a pullback operation defined for bundles, so that for each choice of structure group G and space X, one can reasonably ask for a correspondence

$$\{G\text{-bundles over } X\} \big/ \text{isomorphism} \xrightarrow{c} H^*(X)$$

that is compatible with pullbacks in the sense that

$$c(f^*E) = f^*c(E)$$

for G-bundles $E \to X$ and suitable maps $f : Y \to X$. This is called the **naturality** property, and any map c on a particular class of bundles that satisfies it is called a **characteristic class** of such bundles. It is not difficult to see that our definition of $c_1(E)$ above satisfies it, because pullback bundles inherit pullback connections whose connection 1-forms and curvature 2-forms are likewise pullbacks (see Exercise 48.3 in the next lecture).

47.2. The main idea. While the definition of $c_1(E)$ sketched above is elegant, its scope is quite limited, e.g. it does not work for complex vector bundles of rank $m \geqslant 2$, since $\mathfrak{u}(m)$ is then a nonabelian Lie algebra, more complicated than the space of pure imaginary numbers. The transformation formula $F_\beta = \mathrm{Ad}_{g_{\beta\alpha}} \circ F_\alpha$ for local curvature 2-forms with respect to different trivializations (see Exercise 46.17(f)) gives a hint of one possible way to proceed. In the concrete case $G = \mathrm{U}(m)$, we can abbreviate $g := g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{U}(m)$ and rewrite this formula as

$$F_\beta(X, Y) = g(p) F_\alpha(X, Y) g(p)^{-1} \in \mathfrak{u}(m), \qquad \text{for } p \in \mathcal{U}_\alpha \cap \mathcal{U}_\beta,\ X, Y \in T_pM,$$

and it follows for instance that

$$F(X, Y) := i \operatorname{tr}(F_\alpha(X, Y)) \in \mathbb{R}$$

defines a real-valued 2-form on M that can be computed at any point $p \in M$ by choosing a local trivialization $(\mathcal{U}_\alpha, \Phi_\alpha)$ with $p \in \mathcal{U}_\alpha$, but does not depend on that choice. It is real-valued because matrices in $\mathfrak{u}(m)$ are anti-Hermitian and thus have only imaginary numbers on the diagonal, and it is independent of the choice because the trace map $\operatorname{tr} : \mathbb{C}^{m \times m} \to \mathbb{C}$ is invariant under conjugation. It is also closed, even if the local curvature forms F_α are not: these satisfy $F_\alpha = dA_\alpha + \frac{1}{2}[A_\alpha, A_\alpha]$ by the second structural equation, but values of $\frac{1}{2}[A_\alpha, A_\alpha] \in \Omega^2(\mathcal{U}_\alpha, \mathfrak{u}(m))$ take the form $[A_\alpha(X), A_\alpha(Y)] = A_\alpha(X)A_\alpha(Y) - A_\alpha(Y)A_\alpha(X)$, so they are traceless since $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA})$. Up to a scaling factor to be specified later, the cohomology class represented by this closed 2-form will serve as a definition of $c_1(E) \in H^2_{\mathrm{dR}}(M)$ for complex vector bundles $E \to M$ of arbitrary rank.

REMARK 47.1. In case the proof above that $dF = 0$ struck you as somewhat *ad hoc*, don't worry, we'll come up with a better one later (cf. Exercise 47.9).

The crucial property of the trace that makes the trick above work is conjugation invariance: in fancier terms, the trace defines an Ad-*invariant* linear function $\mathfrak{u}(m) \to \mathbb{C}$. It is natural to wonder whether any other Ad-invariant functions $\mathfrak{u}(m) \to \mathbb{C}$ we can think of might be put to similar use in defining characteristic classes. One such function is the determinant,

$$\det : \mathfrak{u}(m) \to \mathbb{C},$$

which is conjugation-invariant but not linear; strictly speaking, it is a homogenous polynomial of degree m in the entries of the matrices in $\mathfrak{u}(m)$. As we will review in §47.3 below, being a polynomial of degree m means that it can be written in the form $\det(\mathbf{A}) = Q(\mathbf{A}, \ldots, \mathbf{A})$ for a unique symmetric m-fold multilinear map $Q : \mathfrak{u}(m) \times \ldots \times \mathfrak{u}(m) \to \mathbb{C}$, and we recall from §45.3 that such a multilinear map can be used to define product forms $Q(F_\alpha, \ldots, F_\alpha) \in \Omega^{2m}(\mathcal{U}_\alpha, \mathbb{C})$. This suggests the idea of turning any Ad-invariant k-fold multilinear map $\mu : \mathfrak{g} \times \ldots \times \mathfrak{g} \to \mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ for $k \in \mathbb{N}$ into a $2k$-form $\mu(F_\alpha, \ldots, F_\alpha) \in \Omega^{2k}(\mathcal{U}_\alpha, \mathbb{F})$, which we would then expect to be closed and independent of the choice of trivialization, defining a global closed form whose cohomology class in $H^{2k}_{\mathrm{dR}}(M; \mathbb{F})$ should be a characteristic class.

In order to see why this idea works, let's first translate our presentation of $c_1(E)$ into the more global language of the previous lecture. We assume $E \to M$ is a Hermitian vector bundle of rank $m \in \mathbb{N}$, so its structure group is $\mathrm{U}(m)$ and it can be viewed as an associated bundle for the principal $\mathrm{U}(m)$-bundle

$$F^{\mathrm{O}}(E) := P \to M,$$

its orthonormal frame bundle. The curvature forms $\{F_\alpha \in \Omega^2(\mathcal{U}_\alpha, \mathfrak{g})\}_{\alpha \in I}$ used above are local manifestations of a global object, namely the bundle-valued curvature 2-form $\Omega_A \in \Omega^2(M, \mathrm{Ad}(P))$, or equivalently, the horizontal and Ad-equivariant curvature 2-form $F_A \in \Omega^2_{\mathrm{Ad}}(P, \mathfrak{g})$ on the frame bundle. In the general case, $\mathrm{Ad}(P) \to M$ might be a nontrivial bundle with a nontrivial connection, but we can ask whether Ω_A is *covariantly* closed, meaning $d_\nabla \Omega_A = 0$. By Theorem 46.8, this will be true if and only if $d_A F_A = 0$, and we can use the formula $d_A = d + A \wedge (\cdot)$ to compute this, where in the present case, the wedge product of $A \in \Omega^1(P, \mathfrak{g})$ with $F_A \in \Omega^2(P, \mathfrak{g})$ will be defined in terms of the bilinear map $\mathfrak{g} \times \mathfrak{g} \to \mathfrak{g} : (X, Y) \mapsto \mathrm{ad}_X Y = [X, Y]$, and is thus actually the same thing as the bracket $[F_A, A]$. From the second Bianchi identity (Theorem 45.11), we conclude

$$d_A F_A = dF_A + A \wedge F_A = [F_A, A] + [A, F_A] = 0, \tag{47.1}$$

as the bracket is still antisymmetric whenever the forms we plug into it do not both have odd degree. This implies

$$d_\nabla \Omega_A = 0,$$

thus giving us a new interpretation of the second Bianchi identity: it says that the bundle-valued curvature 2-form on M is covariantly closed. If G is abelian, then several details now become simpler: the triviality of the adjoint representation makes the bundle $\mathrm{Ad}(P) \to M$ canonically isomorphic to the trivial bundle $M \times \mathfrak{g} \to M$ with its trivial connection, so Ω_A gets interpreted as a $\mathfrak{g}$-valued 2-form and the equation $d_\nabla \Omega_A = 0$ becomes $d\Omega_A = 0$. In the case $G = \mathrm{U}(1)$, one can now use the fact that $\mathfrak{u}(1) = i\mathbb{R}$ and define $c_1(E) := [-(1/2\pi i)\Omega_A] \in H^2_{\mathrm{dR}}(M)$ as before. Exercise 46.17(c) implies that this really is the same definition, because the local curvature forms $F_\alpha \in \Omega^2(\mathcal{U}_\alpha, \mathfrak{u}(1))$ are nothing other than the local representatives of $\Omega_A \in \Omega^2(M, \mathrm{Ad}(P))$ with respect to trivializations, implying $\Omega_A = F$.

In the general nonabelian case, proceeding further requires having a multilinear map

$$\psi : \underbrace{\mathfrak{g} \times \ldots \times \mathfrak{g}}_{k} \to \mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$$

for some $k \geqslant 1$ that is also Ad-invariant in the sense that $\psi(\mathrm{Ad}_g(X_1), \ldots, \mathrm{Ad}_g(X_k)) = \psi(X_1, \ldots, X_k)$ for all $g \in G$ and $X_1, \ldots, X_k \in \mathfrak{g}$. This can be interpreted as a G-equivariant linear map $\mathfrak{g} \otimes \ldots \otimes \mathfrak{g} \to \mathbb{F}$, where the tensor product inherits the natural tensor product representation $\mathrm{Ad}^{\otimes k} : G \to \mathrm{GL}(\mathfrak{g}^{\otimes k})$, and $\mathbb{F}$ carries the trivial representation. Proposition 46.1 then gives a parallel bundle map

$$\Psi : \mathrm{Ad}(P)^{\otimes k} \cong P^{(\mathrm{Ad}^{\otimes k})} \to P^{\mathrm{triv}} = M \times \mathbb{F},$$

and we can feed Ω_A into the resulting multilinear map $\Psi : \Omega^{j_1}(M, \mathrm{Ad}(P)) \times \ldots \times \Omega^{j_k}(M, \mathrm{Ad}(P)) \to \Omega^{j_1+\ldots+j_k}(M, \mathbb{F})$, producing a real or complex-valued $2k$-form

$$\omega_k := \Psi(\Omega_A, \ldots, \Omega_A) \in \Omega^{2k}(M, \mathbb{F}).$$

By Exercise 45.9, we can apply the covariant Leibniz rule to conclude in this case that ω_k is also closed, as a consequence of the second Bianchi identity $d_\nabla \Omega_A = 0$:

$$\begin{aligned} d\omega_k = d\left(\Psi(\Omega_A, \ldots, \Omega_A)\right) &= \Psi(d_\nabla \Omega_A, \Omega_A, \ldots, \Omega_A) + \Psi(\Omega_A, d_\nabla \Omega_A, \ldots, \Omega_A) + \ldots \\ &\quad + \Psi(\Omega_A, \ldots, \Omega_A, d_\nabla \Omega_A) = 0. \end{aligned}$$

It will be straightforward to verify that the cohomology class

$$[\omega_k] \in H^{2k}_{\mathrm{dR}}(M; \mathbb{F})$$

is then a characteristic class, and we will work out the details of this in §47.4 below. Let's pause briefly to clarify some notation: we are writing

$$H^*_{\mathrm{dR}}(M; \mathbb{R}) := H^*_{\mathrm{dR}}(M)$$

for the usual de Rham cohomology as a real vector space, and $H^*_{\mathrm{dR}}(M; \mathbb{C})$ for its analogue based on complex-valued forms, which is algebraically just the complexification of $H^*_{\mathrm{dR}}(M; \mathbb{R})$. It will turn out that the most important characteristic classes we construct can be viewed as elements of $H^*_{\mathrm{dR}}(M; \mathbb{R})$, but it is convenient to have the freedom of defining some of them first in $H^*_{\mathrm{dR}}(M; \mathbb{C})$.

REMARK 47.2. Most people think of characteristic classes as objects associated to *vector* bundles in particular, but the construction of Chern-Weil theory does not actually use the fibers in any way—what it depends on rather is the underlying system of transition functions, i.e. the "abstract" G-bundles (see Definition 43.1), on top of which fiber bundles with structure group G can be built. Since every abstract G-bundle corresponds canonically to a principal G-bundle, the correct theoretical perspective is therefore to define characteristic classes $c(E) \in H^*(M)$ for *principal* G-bundles $E \to M$ and then define

$$c(E^\rho) := c(E)$$

for all the associated bundles, which includes all vector bundles. We will adopt this perspective in the following.

47.3. Polynomial functions of differential forms. As outlined above, our goal is to represent characteristic classes by plugging curvature 2-forms ω into k-fold multilinear functions $\mu : \mathfrak{g} \times \ldots \times \mathfrak{g} \to \mathbb{F}$ to extract closed $2k$-forms $\mu(\omega, \ldots, \omega)$. Algebraically, functions of the form

$$\mathbb{F}^m \to \mathbb{F} : \mathbf{v} \mapsto \mu(\underbrace{\mathbf{v}, \ldots, \mathbf{v}}_{k})$$

for a multilinear function $\mu : \mathbb{F}^m \times \ldots \times \mathbb{F}^m \to \mathbb{F}$ are homogeneous polynomials of degree k. What follows is a brief digression to clarify a few algebraic facts about such functions.

Assume throughout that V is a vector space of finite dimension $m \in \mathbb{N}$ over $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$. A function $f : V \to \mathbb{F}$ is called a **homogeneous polynomial of degree** $k \geqslant 0$ on V if composing it with an isomorphism $\phi : \mathbb{F}^m \to V$ makes $f \circ \phi : \mathbb{F}^m \to \mathbb{F}$ a homogeneous polynomial of degree k in the standard coordinates $(z^1, \ldots, z^m) \in \mathbb{F}^m$, i.e.

$$f \circ \phi(z^1, \ldots, z^m) = a_{i_1 \ldots i_k} z^{i_1} \ldots z^{i_k}$$

for some set of coefficients $a_{i_1 \ldots i_k} \in \mathbb{F}$. It is easy to check that if f satisfies this condition for some choice of isomorphism $\phi : \mathbb{F}^m \to V$, then it satisfies it for every other choice.

Proposition 47.3. *A function $f : V \to \mathbb{F}$ is a homogeneous polynomial of degree k if and only if it takes the form $f(v) = Q(v, \ldots, v)$ for some k-fold multilinear function $Q : V \times \ldots \times V \to \mathbb{F}$. Moreover, for a given f, there is a unique* symmetric *multilinear function Q satisfying this condition.*

Proof. We can recover the symmetric form Q from the polynomial f by differentiation: according to Taylor's formula,

$$f(v) = \frac{1}{k!} D^k f(0)(v, \ldots, v),$$

so we define $Q(v_1, \ldots, v_k) := \frac{1}{k!} D^k f(0)(v_1, \ldots, v_k)$. □

We will call a function $f : V \to \mathbb{F}$ a (not necessarily homogeneous) **polynomial** on V if it is a finite sum of homogeneous polynomials. More generally, it is sometimes useful to consider **formal power series** on V, by which we mean arbitrary infinite sums

$$f := \sum_{k=0}^{\infty} f_k,$$

such that for each $k \geqslant 0$, the term f_k in the sum is a homogeneous polynomial of degree k. The word "formal" refers to the fact that we do not require these sums to converge, thus a formal power series cannot generally be regarded as a function $V \to \mathbb{F}$, but is instead a purely algebraic object. The set of polynomials and the set of formal power series on V both have natural product structures and thus form algebras over $\mathbb{F}$,

$$\mathbb{F}[V] := \{\text{polynomials on } V\}, \qquad \mathbb{F}[[V]] := \{\text{formal power series on } V\},$$

so e.g. for each pair of formal power series $f = \sum_k f_k$ and $g = \sum_k g_k$, the homogeneous degree k part of $fg \in \mathbb{F}[[V]]$ is $\sum_{j=0}^k f_j g_{k-j}$.

We next consider the operation defined by polynomials on V-valued differential forms on a manifold M. If $f : V \to \mathbb{F}$ is homogeneous of degree k and $Q : V \times \ldots \times V \to \mathbb{F}$ is the symmetric k-fold multilinear map such that $Q(v, \ldots, v) = f(v)$, then Q determines an operation

$$Q : \Omega^{j_1}(M, V) \times \ldots \times \Omega^{j_k}(M, V) \to \Omega^{j_1 + \ldots + j_k}(M, \mathbb{F}),$$

and for each $\omega \in \Omega^j(M, V)$ we define

$$f(\omega) := Q(\omega, \ldots, \omega) \in \Omega^{kj}(M, \mathbb{F}).$$

We can write down an explicit formula for $f(\omega)$ after choosing a basis $e_1, \ldots, e_m$ of V and writing $\omega = \omega^i e_i$ for $\omega^1, \ldots, \omega^m \in \Omega^j(M, \mathbb{F})$: we then have

$$f(\omega) = Q(\omega^{i_1} e_{i_1}, \ldots, \omega^{i_k} e_{i_k}) = Q(e_{i_1}, \ldots, e_{i_k})\, \omega^{i_1} \wedge \ldots \wedge \omega^{i_k}.$$

Notice that if j is odd and $k \geqslant 2$, then $f(\omega) = 0$ since swapping any two of the indices in this last sum changes the sign of the wedge product without changing Q, thus proving $f(\omega) = -f(\omega)$. On the other hand if j is even, then the same trick shows that we could freely have dropped the assumption that Q is symmetric in using it to define $f(\omega)$, as Q can be modified by any permutation of its k variables without changing $Q(\omega, \ldots, \omega)$.

Having defined $f(\omega)$ when f is a homogeneous polynomial, we can now also define

$$f(\omega) := f_0(\omega) + \ldots + f_N(\omega) \in \Omega^*(M, \mathbb{F})$$

for an arbitrary polynomial $f = f_0 + \ldots + f_N$; if f is not homogeneous and ω has positive degree, then $f(\omega)$ will be a finite sum of forms of various degrees. When ω has positive degree, we can even define

$$f(\omega) := f_0(\omega) + f_1(\omega) + f_2(\omega) + \ldots \in \Omega^*(M, \mathbb{F})$$

for an arbitrary formal power series $f = f_0 + f_1 + f_2 + \ldots$ on V, because only the finitely many terms $f_k(\omega) \in \Omega^*(M, \mathbb{F})$ with degree $|f_k(\omega)| = k|\omega| \leqslant \dim M$ are actually nonzero.

Exercise 47.4. Show that for any fixed $\omega \in \Omega^j(M, V)$ of degree $j > 0$, the map $\mathbb{F}[[V]] \to \Omega^*(M, \mathbb{F})$ satisfies

$$(fg)(\omega) = f(\omega) \wedge g(\omega) \qquad \text{for all } f, g \in \mathbb{F}[[V]],$$

and the same is true for $\omega \in \Omega^0(M, V)$ if we restrict to polynomials $f, g \in \mathbb{F}[V]$.

Suppose now that G is a Lie group with a representation $\rho : G \to \mathrm{GL}(V)$ on V. A polynomial $f \in \mathbb{F}[V]$ will be called **ρ-invariant** if it satisfies

$$f(\rho(g)v) = f(v) \qquad \text{for all } g \in G,\ v \in V,$$

and a multilinear map $Q : V \times \ldots \times V \to \mathbb{F}$ is called **ρ-invariant** if it satisfies

$$Q(\rho(g)v_1, \ldots, \rho(g)v_k) = Q(v_1, \ldots, v_k) \qquad \text{for all } g \in G,\ v_1, \ldots, v_k \in V.$$

A ρ-invariant k-fold multilinear map is thus equivalent to a G-equivariant linear map $V^{\otimes k} \to \mathbb{F}$ if we define the G-action on $V^{\otimes k}$ via the k-fold tensor product of the representation ρ, along with the trivial G-action on $\mathbb{F}$.

Proposition 47.5. *A homogeneous polynomial $f : V \to \mathbb{F}$ of degree $k \in \mathbb{N}$ is ρ-invariant if and only if the corresponding symmetric multilinear map $Q : V \times \ldots \times V \to \mathbb{F}$ is ρ-invariant. Moreover, a non-homogeneous polynomial is ρ-invariant if and only if its degree k homogeneous term is ρ-invariant for every $k \geqslant 0$.*

Proof. For any $f \in \mathbb{F}[V]$, Taylor's formula allows us to write

$$f(v) = \sum_{k=0}^{N} \frac{1}{k!} D^k f(0)(v, \ldots, v)$$

for some finite N, and we see that $f : V \to \mathbb{F}$ is ρ-invariant if and only if all of the derivatives $D^k f(0) : V \times \ldots \times V \to \mathbb{F}$ are ρ-invariant. □

The following definition seems reasonable in light of Proposition 47.5: a formal power series $f \in \mathbb{F}[[V]]$ is **ρ-invariant** if for every $k \geqslant 0$, its degree k homogeneous term is ρ-invariant. Notice that the product of two ρ-invariant polynomials or formal power series is also ρ-invariant, so the subspaces

$$\mathbb{F}[V]^\rho := \left\{ f \in \mathbb{F}[V] \;\middle|\; f \text{ is } \rho\text{-invariant} \right\}, \qquad \mathbb{F}[[V]]^\rho := \left\{ f \in \mathbb{F}[[V]] \;\middle|\; f \text{ is } \rho\text{-invariant} \right\}$$

are also algebras.

Exercise 47.6. Suppose $\rho_j : G \to \mathrm{GL}(V_j)$ for $j = 1, \ldots, k$ and $\tau : G \to \mathrm{GL}(W)$ are representations, and $\mu : V_1 \times \ldots \times V_k \to W$ is a multilinear map that is G-equivariant in the sense that $\mu(\rho_1(g)v_1, \ldots, \rho_k(g)v_k) = \tau(g)\mu(v_1, \ldots, v_k)$ for all $g \in G$ and $v_1, \ldots, v_k \in V$. Show that the induced multilinear map $\mu : \Omega^*(M, V_1) \times \ldots \times \Omega^*(M, V_k) \to \Omega^*(M, W)$ is also G-equivariant, where G acts on vector-valued forms $\omega \in \Omega^*(M, V_1)$ by $g \cdot \omega := \rho_1(g) \circ \omega$ and so forth.

Exercise 47.7. Under the assumptions of Exercise 47.6, if $\pi : E \to M$ is a principal G-bundle, show that

$$\omega_j \in \Omega^*_{\rho_j}(E, V_j) \text{ for } j = 1, \dots, k \quad \Rightarrow \quad \mu(\omega_1, \dots, \omega_k) \in \Omega^*_\tau(E, W).$$

Corollary 47.8 (of Exercise 47.7). *Given a principal G-bundle $\pi : E \to M$ and a representation $\rho : G \to \mathrm{GL}(V)$, if $\omega \in \Omega^j_\rho(E, V)$ and $f \in \mathbb{F}[V]^\rho$, then $f(\omega) \in \Omega^*_{\mathrm{triv}}(E, \mathbb{F})$, meaning $f(\omega)$ is horizontal and satisfies $R^*_g\big(f(\omega)\big) = f(\omega)$ for all $g \in G$. The same also holds for formal power series $f \in \mathbb{F}[[V]]^\rho$ if $j > 0$.* □

47.4. The Chern-Weil homomorphism. We now complete the general construction of characteristic classes for principal G-bundles $\pi : E \to M$ via Chern-Weil theory. The idea sketched in §47.2 was based on the bundle-valued curvature 2-form $\Omega_A \in \Omega^2(M, \mathrm{Ad}(E))$, but it is slightly easier (though equivalent) to work with equivariant vector-valued forms on E, so in this section we shall do that instead.

Choose a connection 1-form $A \in \Omega^1(E, \mathfrak{g})$ on E and let $F_A \in \Omega^2_{\mathrm{Ad}}(E, \mathfrak{g})$ denote its curvature 2-form. For each Ad-invariant formal power series $f \in \mathbb{F}[[\mathfrak{g}]]^{\mathrm{Ad}}$ on the Lie algebra, Corollary 47.8 implies that

$$f(F_A) \in \Omega^*(E, \mathbb{F})$$

is horizontal and G-invariant, so under the natural isomorphism $\Omega^*_{\mathrm{triv}}(E, \mathbb{F}) \cong \Omega^*(M, E^{\mathrm{triv}})$ of Theorem 46.3, it corresponds to a form on M with values in the vector bundle E^{triv} associated to the trivial representation $\mathrm{triv} : G \to \mathrm{GL}(1, \mathbb{F})$, which means an $\mathbb{F}$-valued form

$$\omega^f_A \in \Omega^*(M, \mathbb{F}).$$

In this situation, the relation (46.3) between ω^f_A and $f(F_A)$ translates to

$$f(F_A) = \pi^* \omega^f_A.$$

We claim that ω^f_A is closed. The claim follows from two observations: first, we saw in (47.1) that $d_A F_A = 0$, due to the second Bianchi identity. Second, the operator $d_A := H^* d$ satisfies an obvious Leibniz rule under wedge products of horizontal forms on E, and it follows that it also satisfies such a rule under products induced by any multilinear map $\mu : V_1 \times \ldots \times V_k \to W$, i.e.

$$d_A\big(\mu(\omega_1, \dots, \omega_k)\big) = \mu(d_A\omega_1, \dots, \omega_k) + \ldots + (-1)^{|\omega_1| + \ldots + |\omega_{k-1}|} \mu(\omega_1, \dots, d_A\omega_k),$$

so long as the forms $\omega_j \in \Omega^*(E, V_j)$ for $j = 1, \dots, k$ are all horizontal. Since the homogeneous term in $f \in \mathbb{F}[[\mathfrak{g}]]^{\mathrm{Ad}}$ of each degree $k \in \mathbb{N}$ can be expressed as a multilinear map $\mathfrak{g}^{\otimes k} \to \mathbb{F}$, it follows that $d_A\big(f(F_A)\big) = 0$. But since $f(F_A) = \pi^*\omega^f_A$, $d\big(f(F_A)\big) = \pi^* d\omega^f_A$ is itself horizontal, implying $0 = d_A\big(f(F_A)\big) = d\big(f(F_A)\big) = \pi^* d\omega^f_A$, and thus

$$d\omega^f_A = 0.$$

Exercise 47.9. For a homogeneous Ad-invariant polynomial $f(X) = Q(X, \dots, X)$ of degree k on $\mathfrak{g}$, the symmetric k-fold multilinear map Q defines a G-equivariant linear map $\mathfrak{g}^{\otimes k} \to \mathbb{F}$, so by Proposition 46.1, it also determines a parallel bundle map $Q : \mathrm{Ad}(E)^{\otimes k} \to E^{\mathrm{triv}} = M \times \mathbb{F}$, which can be used to define a differential form $Q(\Omega_A, \dots, \Omega_A) \in \Omega^{2k}(M, \mathbb{F})$. Show that $Q(\Omega_A, \dots, \Omega_A) = \omega^f_A$, and deduce from this a second proof that $d\omega^f_A = 0$ based on $d_\nabla \Omega_A = 0$.

Lemma 47.10. *The cohomology class $[\omega^f_A] \in H^*_{\mathrm{dR}}(M; \mathbb{F})$ is independent of the choice of connection $A \in \Omega^1(E, \mathfrak{g})$.*

PROOF. Given two connection 1-forms $A^0, A^1 \in \Omega^1(E, \mathfrak{g})$, the linear interpolation $A^t := tA^1 + (1-t)A^0 \in \Omega^1(E, \mathfrak{g})$ defines a smooth family of connection 1-forms. Let $\Pi : [0,1] \times M \to M$ denote the projection $(t,p) \mapsto p$. The pullback bundle $\Pi^* E$ is then a principal G-bundle over $[0,1] \times M$ whose total space is naturally diffeomorphic to $[0,1] \times E$, so the 1-form $\widehat{A} \in \Omega^1([0,1] \times E, \mathfrak{g})$ defined by

$$\widehat{A}_{(t,\phi)}(s,\xi) := A^t_\phi(\xi)$$

can be regarded as a 1-form on $\Pi^* E$, which is then easily seen to be a connection 1-form for the principal bundle $\Pi^* E \to [0,1] \times M$. Writing

$$i_t : E \hookrightarrow [0,1] \times E = \Pi^* E : \phi \mapsto (t,\phi)$$

for each $t \in [0,1]$, we have $i_0^* \widehat{A} = A^0$ and $i_1^* \widehat{A} = A^1$, and the corresponding curvature 2-forms satisfy

$$i_0^* F_{\widehat{A}} = F_{A^0}, \qquad i_1^* F_{\widehat{A}} = F_{A^1}.$$

Plugging all three of these curvature forms into $f \in \mathbb{F}[[\mathfrak{g}]]^{\mathrm{Ad}}$ as described above then produces closed forms $\omega^f_{A^0}, \omega^f_{A^1} \in \Omega^*(M, \mathbb{F})$ and $\omega^f_{\widehat{A}} \in \Omega^*([0,1] \times M, \mathbb{F})$ such that for $j = 0, 1$, $\omega^f_{A^j}$ is the pullback of $\omega^f_{\widehat{A}}$ via the inclusion $M \cong \{j\} \times M \hookrightarrow [0,1] \times M$. Since these two inclusions are smoothly homotopic, they induce the same map $H^*_{\mathrm{dR}}([0,1] \times M; \mathbb{F}) \to H^*_{\mathrm{dR}}(M; \mathbb{F})$. □

DEFINITION 47.11. For any principal G-bundle $\pi : E \to M$ and $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, the **Chern-Weil homomorphism** is the map

$$\mathbb{F}[[\mathfrak{g}]]^{\mathrm{Ad}} \to H^*_{\mathrm{dR}}(M, \mathbb{F}) : f \mapsto c_f(E) := [\omega^f_A]$$

defined by choosing any connection 1-form $A \in \Omega^1(E, \mathfrak{g})$ and finding the unique $\omega^f_A \in \Omega^*(M, \mathbb{F})$ such that $\pi^* \omega^f_A = f(F_A) \in \Omega^*(E, \mathbb{F})$.

Thanks to the graded Leibniz rule, the wedge product of differential forms descends to an associative and graded commutative product on de Rham cohomology, known as the **cup product**

$$H^k_{\mathrm{dR}}(M; \mathbb{F}) \times H^\ell_{\mathrm{dR}}(M; \mathbb{F}) \to H^{k+\ell}_{\mathrm{dR}}(M, \mathbb{F}) : ([\alpha], [\beta]) \mapsto [\alpha] \cup [\beta] := [\alpha \wedge \beta].$$

By Exercise 47.4, we then have:

THEOREM 47.12. *For every principal G-bundle $\pi : E \to M$, the Chern-Weil homomorphism $\mathbb{F}[[\mathfrak{g}]]^{\mathrm{Ad}} \to H^*(M; \mathbb{F}) : f \mapsto c_f(E)$ satisfies*

$$c_{fg}(E) = c_f(E) \cup c_g(E)$$

for all $f, g \in \mathbb{F}[[\mathfrak{g}]]^{\mathrm{Ad}}$, i.e. it is an algebra homomorphism. □

We will prove in the next lecture that for every fixed $f \in \mathbb{F}[[\mathfrak{g}]]^{\mathrm{Ad}}$, the association of the class $c_f(E) \in H^*_{\mathrm{dR}}(M; \mathbb{F})$ to each principal G-bundle $\pi : E \to M$ satisfies the required naturality property and thus defines a characteristic class. Beyond this, the properties of such classes generally depend on the choice of Ad-invariant polynomial or formal power series f, and we will see in the next lecture that for the groups which arise in most applications, there is a canonical set of choices of f to use, giving rise to the standard Chern and Pontryagin classes.

48. Chern, Pontryagin, and Euler

48.1. The local perspective on the Chern-Weil homomorphism. In the previous lecture we saw two closely related ways to define the Chern-Weil homomorphism

$$\mathbb{F}[[\mathfrak{g}]]^{\mathrm{Ad}} \to H^*_{\mathrm{dR}}(M;\mathbb{F}) : f \mapsto c_f(E)$$

for a principal G-bundle $\pi : E \to M$. Both require an initial choice of connection 1-form $A \in \Omega^1(E,\mathfrak{g})$, though by Lemma 47.10, the eventual definition of $c_f(E) \in H^*_{\mathrm{dR}}(M;\mathbb{F})$ is independent of this choice. In the first approach, one plugs the horizontal and Ad-equivariant curvature 2-form $F_A \in \Omega^2_{\mathrm{Ad}}(E,\mathfrak{g})$ into $f \in \mathbb{F}[[\mathfrak{g}]]^{\mathrm{Ad}}$, giving rise to a scalar-valued horizontal form

$$f(F_A) \in \Omega^*(E,\mathbb{F})$$

that is G-invariant due to the Ad-invariance of f, and is therefore the pullback via $\pi : E \to M$ of some form $\omega^f_A \in \Omega^*(M,\mathbb{F})$. From this perspective, ω^f_A is closed due to the second Bianchi identity $d_A F_A = 0$: the latter implies via a Leibniz rule that $f(F_A)$ is also annihilated by d_A, which is equivalent to being closed since $d\big(f(F_A)\big) = \pi^* d\omega^f_A$ is horizontal.

The second perspective was outlined in Exercise 47.9, and is easiest to explain if we assume $f \in \mathbb{F}[[\mathfrak{g}]]^{\mathrm{Ad}}$ is a homogeneous polynomial of degree k (from which the general case follows just by summing homogeneous terms). The idea in this case is to write $f(X) = Q(X,\dots,X)$ for a symmetric and Ad-invariant k-fold multilinear map $Q : \mathfrak{g}\times\dots\times\mathfrak{g} \to \mathbb{F}$, which by Proposition 46.1 gives rise to a parallel bundle map $Q : \mathrm{Ad}(E)^{\otimes k} \to E^{\mathrm{triv}} = M \times \mathbb{F}$. One then uses the bundle-valued curvature 2-form $\Omega_A \in \Omega^2(M,\mathrm{Ad}(E))$, which is equivalent to $F_A \in \Omega^2_{\mathrm{Ad}}(E,\mathfrak{g})$ under the isomorphism of Theorem 46.3, to define

$$\omega^f_A := f(\Omega_A) := Q(\Omega_A,\dots,\Omega_A) \in \Omega^*(M,E^{\mathrm{triv}}) = \Omega^*(M,\mathbb{F}),$$

and $d\omega^f_A = 0$ follows from another Leibniz rule and the bundle-valued version of the second Bianchi identity, which is the equation $d_\nabla \Omega_A = 0$. In either case, $c_f(E) \in H^*_{\mathrm{dR}}(M;\mathbb{F})$ is defined as the cohomology class represented by the closed $\mathbb{F}$-valued form ω^f_A.

Here is a third perspective that is sometimes useful, and generalizes our original presentation of the first Chern class for line bundles. As explained in Exercise 46.17, we can associate to every local section $s_\alpha \in \Gamma(E|_{\mathcal{U}_\alpha})$ the local connection and curvature forms

$$A_\alpha := s^*_\alpha A \in \Omega^1(\mathcal{U}_\alpha,\mathfrak{g}), \qquad F_\alpha := s^*_\alpha F_A \in \Omega^2(\mathcal{U}_\alpha,\mathfrak{g}),$$

and interpret F_α as the local representation of the bundle-valued curvature form $\Omega_A \in \Omega^2(M,\mathrm{Ad}(E))$ with respect to the trivialization $\mathrm{Ad}(E)|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathfrak{g}$ determined by the section s_α. If $s_\beta \in \Gamma(E|_{\mathcal{U}_\beta})$ is related to s_α by $s_\alpha = s_\beta g_{\beta\alpha}$ for a transition function $g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to G$, then the resulting local curvature 2-forms are related on $\mathcal{U}_\alpha \cap \mathcal{U}_\beta$ by

$$F_\beta = \mathrm{Ad}_{g_{\beta\alpha}} \circ F_\alpha.$$

It follows that for any $f \in \mathbb{F}[[\mathfrak{g}]]^{\mathrm{Ad}}$, the scalar-valued forms $f(F_\alpha) \in \Omega^*(\mathcal{U}_\alpha,\mathbb{F})$ and $f(F_\beta) \in \Omega^*(\mathcal{U}_\beta,\mathbb{F})$ match where they overlap,

$$f(F_\alpha) = f(F_\beta) \qquad \text{on } \mathcal{U}_\alpha \cap \mathcal{U}_\beta,$$

implying that both are restrictions of a global scalar-valued form on M, and we claim that that form is again $\omega^f_A \in \Omega^*(M,\mathbb{F})$, i.e.

$$f(F_\alpha) = \omega^f_A|_{\mathcal{U}_\alpha}.$$

To see this quickly, we can use the first definition of ω^f_A via the relation $f(F_A) = \pi^*\omega^f_A$: after checking that the operation of pulling back forms commutes with the operation of feeding them into polynomials, this gives

$$f(F_\alpha) = f(s_\alpha^* F_A) = s_\alpha^* f(F_A) = s_\alpha^* \pi^* \omega_A^f = (\pi \circ s_\alpha)^* \omega_A^f = \omega_A^f|_{\mathcal{U}_\alpha}$$

since $\pi \circ s_\alpha$ is the identity map on $\mathcal{U}_\alpha$. It is less easy to see via local curvature forms why ω_A^f is closed and its cohomology class independent of choices, but this is one of the big advantages of having the freedom to switch to the global perspective.

THEOREM 48.1. *For any $f \in \mathbb{F}[[\mathfrak{g}]]^{\mathrm{Ad}}$, $c_f(E) = 0$ if the bundle $\pi : E \to M$ is trivial.*

PROOF. On a trivial bundle, we can choose the trivial connection, which is flat, thus $F_A = 0$ and it follows that $f(F_A)$ and ω_A^f vanish. □

THEOREM 48.2. *For each $f \in \mathbb{F}[[\mathfrak{g}]]^{\mathrm{Ad}}$, the class c_f has the following naturality property: for every principal G-bundle $\pi : E \to M$ and smooth map $\varphi : N \to M$, $c_f(\varphi^* E) = \varphi^* c_f(E)$.*

PROOF. Choose a connection $A \in \Omega^1(E, \mathfrak{g})$ on $E \to M$ and fix the induced pullback connection on $\varphi^* E \to N$. By Exercise 48.3 below, every local trivialization of E over a set $\mathcal{U}_\alpha \subset M$ with local curvature form $F_\alpha \in \Omega^2(\mathcal{U}_\alpha, \mathfrak{g})$ induces a local trivialization of $\varphi^* E$ over $\varphi^{-1}(\mathcal{U}_\alpha) \subset N$ for which the local curvature form of the pullback connection is $\varphi^* F_\alpha \in \Omega^2(\varphi^{-1}(\mathcal{U}_\alpha), \mathfrak{g})$. Plugging both into the polynomial f, it follows that $\varphi^* \omega_A^f \in \Omega^*(N, \mathbb{F})$ is a representative of $c_f(\varphi^* E)$. □

EXERCISE 48.3. Suppose $E \to M$ is a principal G-bundle with connection 1-form $A \in \Omega^1(E, \mathfrak{g})$ and curvature 2-form $F_A \in \Omega^2(E, \mathfrak{g})$, $\rho : G \to \mathrm{GL}(V)$ is a representation, ∇ is the associated connection on the associated vector bundle $E^\rho \to M$, and $f : N \to M$ is a smooth map. Let

$$\Psi : f^* E \to E$$

denote the canonical fiber-preserving map that sends $(f^* E)_p \subset f^* E$ to $E_{f(p)} \subset E$ for each $p \in N$ as the identity map. Prove:

(a) The connection 1-form for the pullback connection (cf. Exercise 44.3) on the principal G-bundle $f^* E \to N$ is $\Psi^* A \in \Omega^1(f^* E, \mathfrak{g})$, and its curvature 2-form is $\Psi^* F_A \in \Omega^2(f^* E, \mathfrak{g})$.
(b) The associated bundle $(f^* E)^\rho \to N$ is naturally isomorphic to the pullback $f^* E^\rho \to N$, and under this identification, the connection on $(f^* E)^\rho$ determined by $\Psi^* A$ is the pullback of ∇.
(c) Under the isomorphism $\Omega^k(M, E^\rho) \to \Omega^k_\rho(E, V) : \omega \mapsto \hat{\omega}$ of Theorem 46.3, $f^* \omega \in \Omega^k(N, f^*(E^\rho)) = \Omega^k(N, (f^* E)^\rho)$ satisfies $\widehat{f^* \omega} = \Psi^* \hat{\omega} \in \Omega^k_\rho(f^* E, V)$. In particular, this identifies the bundle-valued connection 2-form $\Omega_{\Psi^* A} \in \Omega^2(N, (f^* E)^\rho)$ for the pullback connection with $f^* \Omega_A \in \Omega^2(N, f^* E^\rho)$.
(d) For a G-compatible local trivialization $\Phi_\alpha : E^\rho|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times V$ and associated local connection and curvature forms $A_\alpha \in \Omega^1(\mathcal{U}_\alpha, \mathfrak{g})$ and $F_\alpha \in \Omega^2(\mathcal{U}_\alpha, \mathfrak{g})$ as in Exercise 46.17, the corresponding local connection and curvature forms for the pullback connection on $f^* E^\rho$ with respect to the pullback trivialization $f^* \Phi_\alpha : (f^* E^\rho)|_{f^{-1}(\mathcal{U}_\alpha)} \to f^{-1}(\mathcal{U}_\alpha) \times V$ are $f^* A_\alpha \in \Omega^1(f^{-1}(\mathcal{U}_\alpha), \mathfrak{g})$ and $f^* F_\alpha \in \Omega^2(f^{-1}(\mathcal{U}_\alpha), \mathfrak{g})$ respectively.

48.2. Chern classes. Concrete examples of characteristic classes arise from examples of Ad-invariant polynomials on various Lie algebras. For $G := \mathrm{GL}(m, \mathbb{C})$, the function

$$f : \mathfrak{gl}(m, \mathbb{C}) \to \mathbb{C} : \mathbf{A} \mapsto \det\left(\mathbb{1} - \frac{1}{2\pi i}\mathbf{A}\right)$$

is Ad-invariant and takes the form

$$f(\mathbf{A}) = f_0(\mathbf{A}) + f_1(\mathbf{A}) + \ldots + f_m(\mathbf{A}) = 1 - \frac{1}{2\pi i}\operatorname{tr}(\mathbf{A}) + \ldots + \left(-\frac{1}{2\pi i}\right)^m \det(\mathbf{A}).$$

The characteristic class of a principal $\mathrm{GL}(m,\mathbb{C})$-bundle $\pi : E \to M$ determined by this polynomial is called the **total Chern class**

$$c(E) := c_f(E) = 1 + c_1(E) + \ldots + c_m(E) \in H^*_{\mathrm{dR}}(M;\mathbb{C}),$$

and for $k \in \mathbb{N}$, the part arising from the degree k homogeneous term in f gives rise to the **kth Chern class**

$$c_k(E) \in H^{2k}_{\mathrm{dR}}(M;\mathbb{C}).$$

For a complex vector bundle $E \to M$ of rank m, the kth Chern class $c_k(E)$ for $k \in \mathbb{N}$ is then defined as the kth Chern class of its frame bundle, a principal $\mathrm{GL}(m,\mathbb{F})$-bundle $FE \to M$. Note that by this definition, $c_k(E)$ is automatically 0 whenever k is bigger than the rank m, and $c_m(E)$ is thus sometimes called the *top* Chern class of E. You should take a moment to convince yourself that our new definition of $c_1(E)$ matches the one we already had for Hermitian line bundles. (The new definition also makes it obvious that $c_1(E)$ does not depend on any choice of bundle metric, even though such a choice was required for our original definition.)

Since every principal $\mathrm{GL}(m,\mathbb{C})$-bundle is the frame bundle of a complex vector bundle with rank m, all important results about the Chern classes for principal bundles can be stated and proved as results about vector bundles, and this is sometimes more convenient. A useful observation in this context is the following:

LEMMA 48.4. *For a vector bundle $E \to M$ of rank m over $\mathbb{F} \in \{\mathbb{R},\mathbb{C}\}$, any linear connection ∇ on E uniquely determines a correpsonding principal connection on the frame bundle $FE \to M$, and thus uniquely determines the local connection 1-forms $A_\alpha \in \Omega^1(\mathcal{U}_\alpha,\mathfrak{gl}(m,\mathbb{F}))$ and curvature 2-forms $F_\alpha \in \Omega^2(\mathcal{U}_\alpha,\mathfrak{gl}(m,\mathbb{F}))$ associated to that principal connection and choices of local frames $s_\alpha \in \Gamma(FE|_{\mathcal{U}_\alpha})$ over regions $\mathcal{U}_\alpha \subset M$.* □

This statement is obvious if one thinks of the way that parallel transport of frames is uniquely determined by parallel transport of vectors, but it's worth drawing specific attention to it anyway because certain natural generalizations of Lemma 48.4 are false: in general, a principal connection on a principal G-bundle $\pi : E \to M$ is not uniquely determined by the connection it induces on any given associated vector bundle $E^\rho \to M$. We observed this in Remark 43.15 in the more general context of associated fiber bundles, but the danger only actually arises if G does not act effectively on the standard fiber, i.e. if the representation $\rho : G \to \mathrm{GL}(V)$ giving rise to the associated vector bundle $E^\rho \to M$ is not faithful. Lemma 48.4 is not susceptible to this danger, because in order to produce a vector bundle $E \to M$ as an associated bundle of its own frame bundle, the representation required is the identity map $\mathrm{Id} : \mathrm{GL}(m,\mathbb{F}) \to \mathrm{GL}(m,\mathbb{F})$, and that is indeed a faithful representation.

PROPOSITION 48.5. *The Chern classes take values in real cohomology $H^*_{\mathrm{dR}}(M;\mathbb{R})$, i.e. for every complex vector bundle $E \to M$ and $k \in \mathbb{N}$, $c_k(E) \in H^{2k}_{\mathrm{dR}}(M;\mathbb{C})$ can be represented by a real-valued closed $2k$-form on M.*

PROOF. A quick summary of the proof is as follows: since every vector bundle admits a positive bundle metric, the structure group can always be reduced from $\mathrm{GL}(m,\mathbb{C})$ to $\mathrm{U}(m)$, whose Lie algebra has the property that $\det(\mathbb{1} - (1/2\pi i)\mathbf{A}) \in \mathbb{R}$ for all $\mathbf{A} \in \mathfrak{u}(m)$.

Here it is with some more details. Using a partition of unity, we can always construct a Hermitian bundle metric on the vector bundle $E \to M$, and then choose a connection ∇ that is compatible with it, i.e. a connection that is induced by a principal connection on the corresponding orthonormal frame bundle. This connection determines a principal connection on the general frame bundle $FE \to M$ with the special property that for any local section $s_\alpha \in \Gamma(FE|_{\mathcal{U}_\alpha})$ that is an orthonormal frame, the corresponding local connection and curvature forms $A_\alpha \in \Omega^1(\mathcal{U}_\alpha,\mathfrak{gl}(m,\mathbb{C}))$ and $F_\alpha \in \Omega^2(\mathcal{U}_\alpha,\mathfrak{gl}(m,\mathbb{C}))$ respectively both take values in the Lie subalgebra $\mathfrak{u}(m) \subset \mathfrak{gl}(m,\mathbb{C})$.

Plugging F_α into the polynomial $f(\mathbf{A}) = \det(\mathbb{1} - (1/2\pi i)\mathbf{A})$ then produces a real-valued form $\omega_A^f \in \Omega^*(M, \mathbb{R})$ since $\mathbb{1} - (1/2\pi i)\mathbf{A}$ is Hermitian whenever $\mathbf{A} \in \mathfrak{u}(m)$, implying that its determinant is real. □

The best method for computing $c_1(E)$ when $E \to \Sigma$ is a complex line bundle over a closed oriented surface Σ was discussed in Lecture 30 last semester: it uses a theorem that expresses $\int_\Sigma \omega$ for any 2-form ω representing $c_1(E)$ as the signed count of zeroes of a generic section of E. Most interesting computations of Chern classes for higher-rank bundles are based on a combination of that result with the following theorem.

THEOREM 48.6 (Whitney product formula). *For any pair of complex vector bundles $E, F \to M$, the total Chern class satisfies $c(E \oplus F) = c(E) \cup c(F)$.*

In the setting of this theorem, the portion of $c(E) \cup c(F) = (1 + c_1(E) + \ldots) \cup (1 + c_1(F) + \ldots) = 1 + c_1(E) + c_1(F) + \ldots$ living in $H^2_{\mathrm{dR}}(M)$ is $c_1(E) + c_1(F)$, thus:

COROLLARY 48.7. *The first Chern class of complex vector bundles is additive with respect to direct sums.* □

PROOF OF THEOREM 48.6. As in the proof of Proposition 48.5, the main idea is to reduce the structure group of $E \oplus F$ and choose a connection that is compatible with this reduction. Assuming E and F have rank m and n respectively, constructing local frames for $E \oplus F$ out of overlapping local frames for E and F reduces its structure group from $\mathrm{GL}(m+n, \mathbb{C})$ to the subgroup

$$G := \mathrm{GL}(m, \mathbb{C}) \times \mathrm{GL}(n, \mathbb{C}) \subset \mathrm{GL}(m+n, \mathbb{C})$$

consisting of matrices in block form $\begin{pmatrix} \mathbf{A} & 0 \\ 0 & \mathbf{B} \end{pmatrix}$ for $\mathbf{A} \in \mathrm{GL}(m, \mathbb{C})$ and $\mathbf{B} \in \mathrm{GL}(n, \mathbb{C})$. Choosing connections separately on E and F, there is also a natural direct sum connection on $E \oplus F$ that is compatible with its G-structure, so the resulting local curvature 2-forms F_α take values in $\mathfrak{g} = \mathfrak{gl}(m, \mathbb{C}) \times \mathfrak{gl}(n, \mathbb{C}) \subset \mathfrak{gl}(m+n, \mathbb{C})$. For $\mathbf{M} = \begin{pmatrix} \mathbf{A} & 0 \\ 0 & \mathbf{B} \end{pmatrix} \in \mathfrak{g}$, we have

$$f(\mathbf{M}) = \det \begin{pmatrix} \mathbb{1} - \frac{1}{2\pi i}\mathbf{A} & 0 \\ 0 & \mathbb{1} - \frac{1}{2\pi i}\mathbf{B} \end{pmatrix} = \det\left(\mathbb{1} - \frac{1}{2\pi i}\mathbf{A}\right) \cdot \det\left(\mathbb{1} - \frac{1}{2\pi i}\mathbf{B}\right) = f(\mathbf{A}) f(\mathbf{B}).$$

It thus follows from Exercise 47.4 that the resulting representative of $c(E \oplus F)$ is a wedge product of two elements of $\Omega^*(M, \mathbb{C})$ representing $c(E)$ and $c(F)$ respectively. □

EXERCISE 48.8. For a principal $\mathrm{U}(1)$-bundle $\pi : E \to M$ with connection $A \in \Omega^1(E, \mathfrak{g})$, we've seen that the bundle-valued curvature 2-form $\Omega_A \in \Omega^2(M, \mathrm{Ad}(E))$ can be regarded as a closed imaginary-valued 2-form $\Omega_A \in \Omega^2(M, \mathfrak{u}(1)) = \Omega^2(M, i\mathbb{R})$ for which $c_1(E) = [-(1/2\pi i)\Omega_A] \in H^2_{\mathrm{dR}}(M)$. Prove the following converse: For any closed 2-form $\omega \in \Omega^2(M)$, if $[\omega] = c_1(E)$, then the bundle $\pi : E \to M$ admits a principal connection whose curvature 2-form (regarded as an imaginary-valued 2-form on M) is $-2\pi i \omega$.

EXERCISE 48.9. Regarding S^{2n+1} as the unit sphere in $\mathbb{C}^{n+1}$, the group $\mathrm{U}(1) \subset \mathrm{GL}(1, \mathbb{C}) = \mathbb{C}^*$ acts on S^{2n+1} via scalar multiplication, thus defining a principal $\mathrm{U}(1)$-bundle

$$\pi : S^{2n+1} \to \mathbb{CP}^n : (z_0, \ldots, z_n) \mapsto [z_0 : \ldots : z_n]$$

that can also be viewed as the orthonormal frame bundle of the tautological line bundle $E \to \mathbb{CP}^n$ with its canonical bundle metric (cf. Exercise 41.13). Writing $\langle z, w \rangle := \sum_{j=0}^n \bar{z}^j w^j$ for the standard Hermitian inner product on $\mathbb{C}^{n+1}$, we can define a 1-form $\lambda \in \Omega^1(\mathbb{C}^{n+1})$ and 2-form $\omega \in \Omega^2(\mathbb{C}^{n+1})$ by

$$\lambda_z(X) := \mathrm{Re}\langle iz, X \rangle, \qquad \omega_z(X, Y) := \mathrm{Re}\langle iX, Y \rangle$$

for $z \in \mathbb{C}^{n+1}$ and $X, Y \in T_z\mathbb{C}^{n+1} = \mathbb{C}^{n+1}$. Notice that ω satisfies $\omega(X, iX) > 0$ whenever $X \neq 0$.

(a) Show that $A := i\lambda|_{TS^{2n+1}} \in \Omega^1(S^{2n+1}, \mathfrak{u}(1))$ is a connection 1-form for the principal bundle $\pi : S^{2n+1} \to \mathbb{CP}^n$, and the resulting horizontal subspace $H_z S^{2n+1} \subset T_z S^{2n+1}$ for each $z \in S^{2n+1}$ is a complex subspace of $\mathbb{C}^{n+1}$.

(b) Show that for the connection in part (a), the curvature 2-form is $F_A = 2i\omega|_{TS^{2n+1}}$.

(c) Show that the first Chern class of the bundle $\pi : S^{2n+1} \to \mathbb{CP}^n$ can be represented by a closed 2-form $\alpha \in \Omega^2(\mathbb{CP}^n)$ that satisfies $\int_\Sigma \alpha \neq 0$ for every closed complex 1-dimensional submanifold $\Sigma \subset \mathbb{CP}^n$ (which is also a canonically oriented real 2-dimensional submanifold). Conclude that the first Chern class is nonzero.
Hint: The quotient projection $\Pi : \mathbb{C}^{n+1}\backslash\{0\} \to (\mathbb{C}^{n+1}\backslash\{0\})/\mathbb{C}^* = \mathbb{CP}^n$ is a holomorphic map between complex manifolds, implying in particular that $T_z\Pi : \mathbb{C}^{n+1} = T_z(\mathbb{C}^{n+1}\backslash\{0\}) \to T_{[z]}\mathbb{CP}^n$ is complex linear for every $z \in \mathbb{C}^{n+1}\backslash\{0\}$.

Example 48.10. For integers $m > k > 0$, let $E_k^m \to \mathrm{Gr}_k(\mathbb{C}^m)$ denote the tautological complex vector bundle of rank k over the Grassmannian of complex k-planes in $\mathbb{C}^m$ (see Exercise 41.13). The example $E_1^m \to \mathrm{Gr}_1(\mathbb{C}^m) = \mathbb{CP}^{m-1}$ was mentioned in Exercise 48.9 above, which shows that $c_1(E_1^m) \neq 0 \in H^2_{\mathrm{dR}}(\mathbb{CP}^{m-1})$. The following argument extends this to the conclusion that $c_1(E_k^m) \neq 0 \in H^2_{\mathrm{dR}}(\mathrm{Gr}_k(\mathbb{C}^m))$ also for every $m > k \geqslant 2$. One can define a natural embedding $\varphi : \mathbb{CP}^1 = \mathrm{Gr}_1(\mathbb{C}^2) \hookrightarrow \mathrm{Gr}_k(\mathbb{C}^m)$ by

$$\varphi(\ell) := \ell \times \mathbb{C}^{k-1} \times \{0\} \subset \mathbb{C}^2 \times \mathbb{C}^{k-1} \times \mathbb{C}^{m-k-1} = \mathbb{C}^m,$$

which has the property that $\varphi^* E_k^m \to \mathbb{CP}^1$ is the direct sum of $E_1^2 \to \mathbb{CP}^1$ with a trivial bundle whose fiber over every point is the subspace $\mathbb{C}^{k-1} \times \{0\} \subset \mathbb{C}^{m-2}$. Since the trivial bundle has vanishing first Chern class, it follows from the Whitney product formula that

$$c_1(\varphi^* E_k^m) = c_1(E_1^2) \neq 0 \in H^2_{\mathrm{dR}}(\mathbb{CP}^2),$$

and by naturality, this is the same thing as $\varphi^* c_1(E_k^m)$, implying $c_1(E_k^m) \neq 0$.

Remark 48.11. The computation of $c(E_k^m) \in H^*_{\mathrm{dR}}(\mathrm{Gr}_k(\mathbb{C}^m))$ for the tautological bundles $E_k^m \to \mathrm{Gr}_k(\mathbb{C}^m)$ is important for the following reason. According to fundamental results in homotopy theory, the tautological bundles serve as *universal* vector bundles, meaning for instance that every complex vector bundle $E \to M$ of rank k over a compact manifold M is isomorphic to a pullback $\varphi^* E_k^m$ for some smooth map $\varphi : M \to \mathrm{Gr}_k(\mathbb{C}^m)$ with $m > k$ sufficiently large, and similarly, two bundles $\varphi_0^* E_k^m$ and $\varphi_1^* E_k^m$ presented in this way are—assuming again that m is sufficiently large—isomorphic if and only if the maps $\varphi_0, \varphi_1 : M \to \mathrm{Gr}_k(\mathbb{C}^m)$ are smoothly homotopic. Thanks to naturality, the Chern classes of all smooth complex k-plane bundles are therefore completely determined by the Chern classes of $E_k^m \to \mathrm{Gr}_k(\mathbb{C}^m)$ for $m \gg k$, and the main step in showing that the construction of Chern classes via curvature matches the corresponding construction in topology is to show that both give the same result for tautological bundles.

A cleaner version of the statement above about pullbacks of $E_k^m \to \mathrm{Gr}_k(\mathbb{C}^m)$ is obtained if one lets $m \to \infty$: there is a limiting space $\mathrm{Gr}_k(\mathbb{C}^\infty)$, which is unfortunately not a smooth manifold, but is instead an infinite-dimensional cell complex and also has a natural tautological bundle $E_k^\infty \to \mathrm{Gr}_k(\mathbb{C}^\infty)$, understood in this case as a topological vector bundle. The general statement—which is even valid when the base M is noncompact—is that the correspondence $\varphi \mapsto \varphi^* E_k^\infty$ defines a bijection from the set $[M, \mathrm{Gr}_k(\mathbb{C}^\infty)]$ of homotopy classes of continuous maps $M \to \mathrm{Gr}_k(\mathbb{C}^\infty)$ to the set of isomorphism classes of k-plane bundles over M. Equivalently, one can pass to the frame bundle $FE_k^\infty \to \mathrm{Gr}_k(\mathbb{C}^\infty)$ and find a similar bijection between $[M, \mathrm{Gr}_k(\mathbb{C}^\infty)]$ and the set of isomorphism classes of principal $\mathrm{GL}(k,\mathbb{C})$-bundles over M. For this reason, $E\,\mathrm{GL}(k,\mathbb{C}) := FE_k^\infty \to \mathrm{Gr}_k(\mathbb{C}^\infty)$ is called a *universal* principal $\mathrm{GL}(k,\mathbb{C})$-bundle, and its base

$$B\,\mathrm{GL}(k,\mathbb{C}) := \mathrm{Gr}_k(\mathbb{C}^\infty)$$

is called a *classifying space* for the group $\mathrm{GL}(k, \mathbb{C})$. By a fundamental result of Milnor [**Mil56**], every topological group G admits a classifying space BG, a topological space which comes with a universal principal G-bundle $EG \to BG$, which puts the set of all isomorphism classes of principal G-bundles over any given space M into bijective correspondence with the set $[M, BG]$ of homotopy classes of maps $M \to BG$. The existence of classifying spaces makes it possible to place the entire theory of characteristic classes onto an axiomatic footing: by naturality, every characteristic class for principal G-bundles corresponds to a choice of cohomology class on the classifying space BG. The classifying spaces themselves are typically not smooth manifolds, but they usually can be *approximated* in some sense by smooth manifolds such as $\mathrm{Gr}_k(\mathbb{C}^m)$ for $m \gg k$, which makes the methods of Chern-Weil theory applicable even for universal bundles.

48.3. Pontryagin classes. For real vector bundles or principal $\mathrm{GL}(m, \mathbb{R})$-bundles, a natural choice of Ad-invariant polynomial is obtained by removing the i from the one we used for defining the total Chern class:[101]

$$f : \mathfrak{gl}(m, \mathbb{R}) \to \mathbb{R} : \mathbf{A} \mapsto \det\left(\mathbb{1} - \frac{1}{2\pi}\mathbf{A}\right).$$

The resulting characteristic class for principal $\mathrm{GL}(m, \mathbb{R})$-bundles $\pi : E \to M$ is called the **total Pontryagin class**

$$p(E) := c_f(E) \in H^*_{\mathrm{dR}}(M),$$

and whenever $E \to M$ is a real vector bundle of rank m, we define $p(E)$ to be the total Pontryagin class of its frame bundle $FE \to M$, which is a principal $\mathrm{GL}(m, \mathbb{R})$-bundle. Before breaking $p(E)$ down into individual Pontryagin classes of specific degrees, it is worth proving the analogue of Proposition 48.5 in this setting, which says something a bit surprising:

Proposition 48.12. *Only the homogeneous terms in $f : \mathfrak{gl}(m, \mathbb{R}) \to \mathbb{R}$ with even degree can make nontrivial contributions to the total Pontryagin class, hence $p(E) \in H^*_{\mathrm{dR}}(M)$ is represented by a finite sum of forms whose degrees are all divisible by* 4.

Proof. As in Proposition 48.5, the main reason for this is that every vector bundle $E \to M$ admits a bundle metric, which in this case reduces the structure group from $\mathrm{GL}(m, \mathbb{R})$ to $\mathrm{O}(m)$. After choosing a connection compatible with this $\mathrm{O}(m)$-structure, the result follows from a simple algebraic observation: the restriction of the polynomial $f : \mathfrak{gl}(m, \mathbb{R}) \to \mathbb{R}$ to $\mathfrak{o}(m) \subset \mathfrak{gl}(m, \mathbb{R})$ contains no nontrivial homogeneous terms of odd degree. This is true because every $\mathbf{A} \in \mathfrak{o}(m)$ is antisymmetric, but $f(\mathbf{A}) = \det(\mathbb{1} - (1/2\pi i)\mathbf{A})$ does not change when $\mathbf{A}$ is replaced by its transpose, implying $f(\mathbf{A}) = f(-\mathbf{A})$ for $\mathbf{A} \in \mathfrak{o}(m)$. The latter implies that the derivatives of f at 0 with odd order all vanish, so f is a sum of homogeneous terms with even degree. □

Instead of proceeding directly in analogy with the Chern classes and defining $p_k(E) \in H^{2k}_{\mathrm{dR}}(M)$ as the characteristic class arising from the degree k part of $f : \mathfrak{gl}(m, \mathbb{R}) \to \mathbb{R}$, Proposition 48.12 reveals that it is more sensible to define the **kth Pontryagin class**

$$p_k(E) \in H^{4k}_{\mathrm{dR}}(M)$$

for each $k \in \mathbb{N}$ as the class arising from the degree $2k$ part of f.

The similarity of the polynomials used for the Pontryagin and Chern classes suggests that there should be a close relationship between them. To see this, recall that every real vector bundle $E \to M$ of rank m has a **complexification**

$$E^{\mathbb{C}} \to M,$$

[101] Actually, when I presented this in lecture I kept the i in the formula, but I have since decided to remove it. The reason why is explained in Remark 48.14.

which is most easily defined as the real tensor product of E with the trivial complex line bundle $M \times \mathbb{C} \to M$, and $E^{\mathbb{C}}$ can then be regarded as a complex vector bundle, also of rank m, where complex scalar multiplication is defined by $\lambda(v \otimes z) := v \otimes \lambda z$ for $\lambda \in \mathbb{C}$. It is straightforward to check that any bundle atlas for E naturally determines a bundle atlas for $E^{\mathbb{C}}$ that has the same $\mathrm{GL}(m, \mathbb{R})$-valued transition functions, but acting on standard fiber $\mathbb{C}^m$ instead of $\mathbb{R}^m$. In more abstract terms, if we view E as the associated vector bundle $(FE)^\rho$ for the canonical representation $\rho = \mathrm{Id} : \mathrm{GL}(m, \mathbb{R}) \to \mathrm{GL}(m, \mathbb{R})$, then $E^{\mathbb{C}}$ is also an associated bundle, obtained by keeping the same representation of $\mathrm{GL}(m, \mathbb{R})$ but viewing it as a complex representation on $\mathbb{C}^m$.

Exercise 48.13. Show that for any real vector bundle $E \to M$ and $k \in \mathbb{N}$, $c_k(E^{\mathbb{C}}) = 0$ if k is odd and $c_{2k}(E^{\mathbb{C}}) = (-1)^k p_k(E)$.

Remark 48.14. There is a lack of universal agreement on sign conventions in the definitions of certain characteristic classes. I am attempting in these notes to follow what I perceive as the majority view, but in lecture I inadvertently followed a different one (consistent with [**BT82**, **Tu17**]) before I had become completely aware of the problem. In this alternative convention, one takes the polynomial in the definition of the total Pontryagin class to be exactly the same (complex-valued) polynomial as for the total Chern class, just restricted to $\mathfrak{gl}(m, \mathbb{R}) \subset \mathfrak{gl}(m, \mathbb{C})$. Under this definition, Proposition 48.12 still holds, so only the terms of even degree contribute to $p(E)$, which is therefore a real cohomology class, but $p_k(E)$ gets replaced by $-p_k(E)$ when k is odd. This change has the benefit of turning Exercise 48.13 into the slightly prettier formula $c_{2k}(E^{\mathbb{C}}) = p_k(E)$. But most authors do not follow that convention, and thus keep the extra sign in Exercise 48.13.

While we're talking about signs, it should be noted that in the original classic reference on characteristic classes [**MS74**], Milnor and Stasheff appear to use the polynomial $f(\mathbf{A}) := \det(\mathbb{1} + (1/2\pi i)\mathbf{A})$ for their definition of the total Chern class, which differs from ours by the insertion of a sign in front of $\mathbf{A}$. If all other things were equal, this would mean that their definition of $c_k(E)$ for k odd differs from ours by a sign, but I suspect there must be another sign discrepancy in [**MS74**] that cancels this one out, and I haven't found it yet. It should be a nonnegotiable principle that for a complex bundle of rank m, the top Chern class $c_m(E)$ matches the Euler class of E as a canonically oriented real bundle of rank $2m$ (cf. Exercise 48.20 below). For a line bundle over a surface, that means in particular that $c_1(E)$ computes the signed count of zeroes of a generic section as explained in §30.2, and this is precisely what the factor of $-\frac{1}{2\pi i}$ in our definition achieves.

48.4. The Euler class. In topology, there are two further families of standard characteristic classes for a real vector bundle $E \to M$ of rank m: the *Stiefel-Whitney classes*

$$w_k(E) \in H^k(M; \mathbb{Z}_2), \qquad k = 1, \ldots, m,$$

and the *Euler class*

$$e(E) \in H^m(M),$$

which is only defined if the bundle $E \to M$ is orientable and depends on a choice of orientation. We briefly encountered the first Stiefel-Whitney class $w_1(E) \in H^1(M; \mathbb{Z}_2)$ in Remark 32.6, as it serves as the obstruction to orientability for a real vector bundle, and we will see later that the second Stiefel-Whitney class plays a similar role for spin structures. Unfortunately, the Stiefel-Whitney classes can only be defined in cohomology theories with $\mathbb{Z}_2$ coefficients, which makes them inaccessible to de Rham cohomology. Chern-Weil theory does however have a natural construction for the Euler class, at least for oriented real vector bundles of even rank.[102] (We will see in Remark 48.17 that the case of odd rank is not interesting.)

[102]The contents of this section were only fleetingly mentioned in lecture, i.e. the definition of the Euler class in terms of the Pfaffian was stated, but there was no time to discuss any properties of the Euler class, nor the actual

The construction of $e(E) \in H^{2m}_{\mathrm{dR}}(M)$ for an oriented vector bundle $E \to M$ of rank $2m$ is based on an Ad-invariant homogeneous polynomial of degree m on $\mathfrak{so}(2m)$, called the **Pfaffian**,

$$\mathrm{Pf} : \mathfrak{so}(2m) \to \mathbb{R}.$$

To write it down, suppose V is a real vector space of dimension $2m$, endowed with an inner product $\langle\ ,\ \rangle$ and an orientation, and let

$$d\mathrm{vol} \in \Lambda^{2m} V^*$$

denote the canonical volume form, i.e. the one which evaluates to 1 on any positively-oriented orthonormal basis. Let $\mathrm{SO}(V) \subset \mathrm{O}(V)$ denote the identity component of the group $\mathrm{O}(V)$ of orthogonal transformations on V, so the Lie algebra $\mathfrak{so}(V) = \mathfrak{o}(V)$ is the space of linear maps $A \in \mathrm{End}(V)$ that are skew-symmetric with respect to $\langle\ ,\ \rangle$. It follows that we can associate to any $A \in \mathfrak{so}(V)$ an alternating 2-form

$$\omega_A \in \Lambda^2 V^*, \qquad \omega_A(v, w) := \langle v, Aw \rangle.$$

The m-fold wedge product of ω_A with itself is then a scalar multiple of $d\mathrm{vol}$ since $\dim \Lambda^{2m} V^* = 1$, so we can define $\mathrm{Pf}(A) \in \mathbb{R}$ via the relation

$$\frac{1}{m!} \omega_A^m = \mathrm{Pf}(A) \cdot d\mathrm{vol}.$$

The 2-form ω_A depends linearly on A, thus the left hand side of this relation is a degree m homogeneous polynomial function of A, and therefore so is $\mathrm{Pf}(A)$. The combinatorial factor on the left has the following justification. Let us choose a positive orthonormal basis $e_1, \ldots, e_{2m}$ of V and write the resulting matrix entries of A as $A_{ij} := \langle e_i, Ae_j \rangle$. We now have $d\mathrm{vol}(e_1, \ldots, e_{2m}) = 1$ by definition, and writing $e^1_*, \ldots, e^{2m}_* \in V^*$ for the dual basis, we also have

$$\omega_A = \frac{1}{2} A_{ij} e^i_* \wedge e^j_*.$$

Applying the usual combinatorial formula (9.3) for wedge products of 1-forms then gives

$$\begin{aligned} \frac{1}{m!} \omega_A^m(e_1, \ldots, e_{2m}) &= \frac{1}{2^m \cdot m!} A_{i_1 j_1} \cdot \ldots \cdot A_{i_m j_m} \cdot (e^{i_1}_* \wedge e^{j_1}_* \wedge \ldots \wedge e^{i_m}_* \wedge e^{j_m}_*)(e_1, \ldots, e_{2m}) \\ &= \frac{1}{2^m \cdot m!} \sum_{\sigma \in S_{2m}} (-1)^{|\sigma|} A_{\sigma(1),\sigma(2)} \cdot \ldots \cdot A_{\sigma(2m-1),\sigma(2m)}. \end{aligned}$$

In this last sum, there is quite a lot of overcounting: each individual term depends on the way that the permutation $\sigma \in S_{2m}$ partitions the set $\{1, \ldots, 2m\}$ into pairs $\{\sigma(2j-1), \sigma(2j)\}$ for $j = 1, \ldots, m$, but any two permutations that define the same partition into pairs make the same contribution. There are exactly $2^m \cdot m!$ permutations corresponding to each partition, allowing for the freedom to reorder the pairs and to flip each pair individually. We can therefore choose a subset $\widehat{S}_{2m} \subset S_{2m}$ that contains exactly one permutation for every possible partition and write a formula for $\mathrm{Pf}(A)$ that is free of combinatorial factors:

$$\mathrm{Pf}(A) = \sum_{\sigma \in \widehat{S}_{2m}} (-1)^{|\sigma|} A_{\sigma(1),\sigma(2)} \cdot \ldots \cdot A_{\sigma(2m-1),\sigma(2m)} \in \mathbb{R}.$$

Setting $V = \mathbb{R}^{2m}$, we take this as a definition of $\mathrm{Pf}(\mathbf{A})$ for $\mathbf{A} \in \mathfrak{so}(2m)$. The Pfaffian is thus a polynomial function of the entires in $\mathbf{A}$ with integer coefficients; the fact that the coefficients are integers will not matter to us in any direct way, but algebraists like it.

The fact that $\mathrm{Pf} : \mathfrak{so}(V) \to \mathbb{R}$ can be defined without a choice of basis but depends on the inner product and orientation implies that it is automatically invariant under conjugation by

definition of the Pfaffian. I have filled in some of those details here for your information. Thanks to Gerard Bargalló for suggesting the particular definition of $\mathrm{Pf} : \mathfrak{so}(2m) \to \mathbb{R}$ given here.

transformations that preserve these structures, i.e. $\mathrm{SO}(V)$. More generally, for any $B \in \mathrm{GL}(V)$, we have

$$B^*\omega_A(v,w) = \omega_A(Bv,Bw) = \langle Bv, ABw\rangle = \langle v, B^T ABw\rangle = \omega_{B^TAB}(v,w),$$

and $B^* d\mathrm{vol} = \det(B)\cdot d\mathrm{vol}$, thus $\frac{1}{m!}B^*(\omega_A^m) = \frac{1}{m!}\omega_{B^TAB}^m = \mathrm{Pf}(B^TAB)\cdot d\mathrm{vol} = \mathrm{Pf}(A)\cdot B^*d\mathrm{vol}$, implying

$$\mathrm{Pf}(B^TAB) = \det(B)\cdot \mathrm{Pf}(A) \qquad \text{for all } A \in \mathfrak{so}(V),\ B \in \mathrm{GL}(V).$$

This shows explicitly that the Pfaffian can be regarded as an Ad-invariant polynomial on $\mathfrak{so}(V)$, but it is not invariant under conjugation by elements in the larger group $\mathrm{O}(V)$, and thus cannot be considered an Ad-invariant polynomial on $\mathfrak{o}(V)$. This is good news if the goal is to define a characteristic class that is sensitive to orientations.

Aside from conjugation invariance, the most important fact to know about the Pfaffian is that it is a "square root" of the determinant—this property gives a second convincing justification for the combinatorial factor we put into the definition.

PROPOSITION 48.15. *For all $A \in \mathfrak{so}(V)$, $[\mathrm{Pf}(A)]^2 = \det(A)$.*

PROOF. Both Pf^2 and $\det$ are conjugation-invariant polynomials of degree $2m$ on $\mathfrak{so}(V)$, and since every $A \in \mathfrak{so}(V)$ is represented in some positively-oriented orthonormal basis by a matrix of the form

$$\mathbf{A} = \begin{pmatrix} 0 & \lambda_1 & \cdots & 0 & 0 \\ -\lambda_1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \lambda_m \\ 0 & 0 & \cdots & -\lambda_m & 0 \end{pmatrix} \in \mathfrak{so}(2m)$$

for some $\lambda_1,\dots,\lambda_m \in \mathbb{R}$, it suffices to check that the relation holds for linear transformations on $\mathbb{R}^{2m}$ of this form. One then computes from the definition that $\mathrm{Pf}(\mathbf{A}) = \lambda_1\cdot\ldots\cdot\lambda_m$, while $\det(\mathbf{A}) = (\lambda_1\cdot\ldots\cdot\lambda_m)^2$. □

Taking

$$f : \mathfrak{so}(2m) \to \mathbb{R} : \mathbf{A} \mapsto \mathrm{Pf}\left(\frac{1}{2\pi}\mathbf{A}\right)$$

as an Ad-invariant polynomial of degree m, we define the **Euler class** of any principal $\mathrm{SO}(2m)$-bundle $\pi : E \to M$ as the associated characteristic class

$$e(E) := c_f(E) \in H^{2m}_{\mathrm{dR}}(M).$$

Similarly, we associate to any oriented Euclidean vector bundle $E \to M$ of rank $2m$ a principal $\mathrm{SO}(2m)$-bundle $F^{\mathrm{SO}}(E) \to M$, the bundle of *oriented* orthonormal frames, and define

$$e(E) := e(F^{\mathrm{SO}}(E)).$$

While this definition of $e(E)$ explicitly requires a choice of bundle metric for E, the following exercise shows that it does not really depend on this choice, though it does depend on the orientation.

EXERCISE 48.16. Assume $E \to M$ is an oriented real vector bundle.

(a) Show that for any two choices $\langle\ ,\ \rangle_0$ and $\langle\ ,\ \rangle_1$ of positive bundle metric on E, there exists an orientation-preserving vector bundle isomorphism $A : E \to E$ satisfying $\langle Av, Aw\rangle_1 = \langle v, w\rangle_0$ for all $(v,w) \in E \oplus E$.
Hint: Define a bundle metric on the pullback of $E \to M$ via the projection $[0,1]\times M \to M$ that interpolates between $\langle\ ,\ \rangle_0$ and $\langle\ ,\ \rangle_1$. Then use parallel transport.

(b) Show that for the same vector bundle $\bar{E} \to M$ with reversed orientation, $e(\bar{E}) = -e(E)$. *Hint: The fact that* $\mathrm{Pf} : \mathfrak{so}(2m) \to \mathbb{R}$ *changes sign under conjugation with elements of* $\mathrm{O}(2m)\backslash \mathrm{SO}(2m)$ *is relevant here.*

REMARK 48.17. Exercise 48.16(b) reveals the reason why we do not mind the lack of a Chern-Weil construction of $e(E)$ when $\mathrm{rank}(E)$ is odd. The topological version of the Euler class also satisfies $e(\bar{E}) = -e(E)$, but when the rank is odd, the antipodal map $v \mapsto -v$ defines an orientation-preserving bundle isomorphism $E \to \bar{E}$, implying $e(E) = e(\bar{E})$, hence $2e(E) = 0$. This can happen in cohomology with integer coefficients without $e(E)$ being trivial, because $H^*(M; \mathbb{Z})$ is generally an abelian group rather than a vector space, but de Rham cohomology $H^*_{\mathrm{dR}}(M)$ is a vector space, so $2e(E) = 0$ in $H^*_{\mathrm{dR}}(M)$ can only mean that $e(E) = 0$.

We conclude this brief survey by stating (without proof) the most famous result about the Euler class on smooth manifolds: a generalization of the Gauss-Bonnet theorem (cf. Lecture 29 from last semester) to higher dimensions, which can be interpreted as a computation of the Euler class of any tangent bundle. The setting for the theorem is a Riemannian manifold (M, g) of even dimension $2n$, with a connection ∇ on TM that need not be symmetric, but is required to be compatible with the metric. Note that any such connection is automatically also compatible with the $\mathrm{SO}(2n)$-structure of TM, because parallel transport maps can always be deformed continuously to the identity and therefore automatically preserve orientations.

THEOREM 48.18 (generalized Gauss-Bonnet). *Assume (M, g) is an oriented Riemannian manifold of dimension $2n \geqslant 2$, ∇ is an affine connection on M compatible with the metric, and*

$$\mathrm{Pf}(\Omega/2\pi) \in \Omega^{2n}(M)$$

denotes the $2n$-form that locally matches $\mathrm{Pf}(F_\alpha/2\pi)$ for the local curvature 2-form $F_\alpha \in \Omega^2(\mathcal{U}_\alpha, \mathfrak{so}(2m))$ associated to any choice of oriented orthonormal frame over an open subset $\mathcal{U}_\alpha \subset M$. Then integrating $\mathrm{Pf}(\Omega/2\pi)$ gives the Euler characteristic of M,

$$\int_M \mathrm{Pf}(\Omega/2\pi) = \chi(M) \in \mathbb{Z}.$$

EXERCISE 48.19. Check that in the $n = 1$ case of Theorem 48.18, if ∇ is the Levi-Cività connection on (M, g), then $\mathrm{Pf}(\Omega/2\pi) = \frac{1}{2\pi} K_G \, d\mathrm{vol}$, where $K_G : M \to \mathbb{R}$ is the Gaussian curvature.

EXERCISE 48.20. Any complex line bundle $E \to M$ can be regarded as a real bundle of rank 2 with a canonical orientation determined by its complex structure, i.e. we orient each fiber E_p such that for $v \neq 0 \in E_p$, the real basis (v, iv) is considered positively oriented. Show that for this choice of orientation, $e(E) = c_1(E) \in H^2_{\mathrm{dR}}(M)$.

49. Affine transformations and isometries

49.1. Lie transformation groups. It is time to prove a fact that has been mentioned already on a few occasions: for any pseudo-Riemannian manifold (M, g), the group of isometries

$$\mathrm{Isom}(M, g) := \left\{ \psi \in \mathrm{Diff}(M) \;\middle|\; \psi^* g = g \right\}$$

is a Lie group. Actually, it is something better: $\mathrm{Isom}(M, g)$ is a *Lie transformation group*. In general, for a given smooth manifold M, we say that a subgroup

$$G \subset \mathrm{Diff}(M)$$

in the group of diffeomorphisms $M \to M$ is a **Lie transformation group** on M if it admits a smooth structure compatible with the topology it inherits from $\mathrm{Diff}(M)$ (i.e. the C^∞_{loc}-topology) such that it becomes a Lie group and, additionally, the obvious left action

$$G \times M \to M : (\psi, p) \mapsto \psi(p)$$

becomes smooth. In this situation, the G-action on M is automatically effective, and it follows that the Lie algebra antihomomorphism defined via fundamental vector fields (see Theorem 40.8)

$$\mathfrak{g} \to \mathfrak{X}(M) : X \mapsto X^F$$

is injective. Indeed, if $X^F \equiv 0$ for some $X \in \mathfrak{g}$, then it follows via Proposition 40.7 that $\exp(tX) = \mathrm{Id} \in \mathrm{Diff}(M)$ for every $t \in \mathbb{R}$, thus $X = 0$. As a consequence, the Lie algebra of any Lie transformation group has a natural identification with a finite-dimensional Lie subalgebra of the space of vector fields $\mathfrak{X}(M)$, namely the space of so-called **infinitesimal transformations** defined by G,

$$\mathfrak{g} = \left\{ X \in \mathfrak{X}(M) \;\middle|\; X \text{ has a global flow and } \varphi_X^t \in G \text{ for all } t \in \mathbb{R} \right\}.$$

A set of this kind can be defined for arbitrary subgroups $G \subset \mathrm{Diff}(M)$, but if G is not a Lie transformation group, then $\mathfrak{g}$ will not always be a vector space, and certainly not a finite-dimensional one.

Here is an example that is very different from the isometry group.

EXAMPLE 49.1. Recall (cf. Lecture 14 from last semester) that a **symplectic form** on a manifold M of dimension $2n \geqslant 2$ is a 2-form $\omega \in \Omega^2(M)$ that can be identified in some choice of coordinates $(q^1, p^1, \dots, q^n, p^n)$ near any given point with the local model

$$\omega = \sum_{j=1}^{n} dp^j \wedge dq^j.$$

A diffeomorphism $\psi : M \to M$ is then called a **symplectomorphism** $\psi : (M, \omega) \to (M, \omega)$ if it satisfies $\psi^*\omega = \omega$. The symplectomorphisms on a symplectic manifold (M, ω) form a subgroup

$$\mathrm{Symp}(M, \omega) \subset \mathrm{Diff}(M),$$

but the following observation shows that it cannot be a Lie transformation group. Every compactly supported smooth function $H \in C^\infty(M)$ gives rise to a Hamiltonian vector field $X_H \in \mathfrak{X}(M)$, uniquely determined by the condition $\omega(X_H, \cdot) = -dH$. These vector fields all have global flows since they have compact support, and by Cartan's magic formula, $\mathcal{L}_{X_H}\omega = 0$, implying that $\varphi_{X_H}^t \in \mathrm{Symp}(M, \omega)$ for every $t \in \mathbb{R}$. The space of infinitesimal symplectomorphisms thus contains the infinite-dimensional space of all compactly supported Hamiltonian vector fields—one can show that it is a Lie subalgebra, but it is clearly not finite dimensional.

Our main objective in this lecture is to prove that the isometry group $\mathrm{Isom}(M, g)$ and various other closely related objects are Lie transformation groups. One sees a strong hint of this if one examines the corresponding space of infinitesimal transformations. These are called *Killing vector fields*: we will see in §49.3 that they satisfy an overdetermined first-order linear PDE which guarantees the uniqueness (though not the existence) of solutions having any given value and first covariant derivative at one point. As a consequence, the space of Killing vector fields is finite dimensional. This is an encouraging sign, though it does not imply on its own that $\mathrm{Isom}(M, g)$ is a Lie transformation group. The horror scenario one could imagine is that $\mathrm{Isom}(M, g)$ fails to be a manifold near $\mathrm{Id} \in \mathrm{Isom}(M, g)$ because there exist sequences $\psi_k \in \mathrm{Isom}(M, g)$ with $\psi_k \to \mathrm{Id}$ in the C^∞_{loc}-topology such that ψ_k is not expressible as a flow $\varphi^1_{X_k}$ for any sequence of Killing vector fields $X_k \to 0$. This danger is not unlike the scenario that needed to be ruled out when we proved in Theorem 41.2 that closed subgroups of Lie groups are also smooth submanifolds, and the strategy by which we will rule it out also has something in common with the argument used in that theorem. If indeed such sequences can be excluded, then a neighborhood of 0 in the space of Killing vector fields provides a natural parametrization of $\mathrm{Isom}(M, g)$ near the identity map, for which the action on M is manifestly smooth. Composing this parametrization with arbitrary left translations then endows the rest of $\mathrm{Isom}(M, g)$ as well with the structure of a Lie transformation group.

In reality, we will not make the argument quite so directly, but will instead prove that $\mathrm{Isom}(M, g)$ is a closed subgroup of another group that defines a Lie transformation group on the orthonormal frame bundle $F^{\mathrm{O}}(TM)$. As a closed subgroup, $\mathrm{Isom}(M, g)$ will then be a Lie subgroup according to Theorem 41.2, and will therefore also act smoothly on $F^{\mathrm{O}}(TM)$, implying that its action on M is also smooth.

49.2. Affine transformations. Let us place the group $\mathrm{Isom}(M, g)$ into a wider context. It turns out that the main feature causing $\mathrm{Isom}(M, g)$ to be finite dimensional is not the fact that isometries $\psi \in \mathrm{Isom}(M, g)$ preserve the metric, but rather that they preserve the geodesic equation. The set of diffeomorphisms with this property deserves closer examination.

Definition 49.2. Suppose $\psi : M \to N$ is a diffeomorphism and ∇ is an affine connection on N. The affine connection $\psi^*\nabla$ on M is then defined via the condition

$$(\psi^*\nabla)_X Y := \psi^* \left(\nabla_{\psi_* X}(\psi_* Y)\right) \qquad \text{for every } X, Y \in \mathfrak{X}(M).$$

Exercise 49.3. Verify that $\psi^*\nabla$ as given in the definition above is an affine connection on M.

Remark 49.4. The object $\psi^*\nabla$ in Definition 49.2 is not the same thing as what we have previously called the *pullback connection*, which would in this case be a connection on the pullback bundle $\psi^*TN \to M$ rather than $TM \to M$. Pullback connections can be defined for any bundle over N and any smooth map $M \to N$, whereas Definition 49.2 is specific to tangent bundles and only makes sense when the smooth map $\psi : M \to N$ is a diffeomorphism.

Definition 49.5. Given two manifolds M, N equipped with affine connections ∇^M and ∇^N respectively, a diffeomorphism $\psi : M \to N$ is called an **affine transformation** and written $\psi : (M, \nabla^M) \to (N, \nabla^N)$ if

$$\psi^*\nabla^N = \nabla^M.$$

For a single manifold M with affine connection ∇, the set of affine transformations $(M, \nabla) \to (M, \nabla)$ defines a subgroup of $\mathrm{Diff}(M)$ which we will denote by

$$\mathrm{Aff}(M, \nabla) := \left\{\psi \in \mathrm{Diff}(M) \;\middle|\; \psi^*\nabla = \nabla\right\}.$$

Affine transformations $(M, \nabla^M) \to (N, \nabla^N)$ have the property that for any path $\gamma(t) \in M$ and vector field $X(t) \in T_{\gamma(t)}M$ along that path, X is parallel along γ if and only if the vector field $T\psi(X(t)) \in T_{\psi\circ\gamma(t)}N$ along the path $\psi \circ \gamma$ in N is parallel. It follows in particular that γ is a geodesic with respect to ∇^M if and only if $\psi \circ \gamma$ is a geodesic with repsect to ∇^N. The terminology is motivated by the example of $\mathbb{R}^n$ with the trivial connection, for which geodesics are straight lines, thus affine transformations map straight lines to straight lines (cf. Exercise 49.7 below). This basic observation leads easily to the conclusion that, in general, there cannot be very many affine transformations:

Theorem 49.6 (Rigidity of affine transformations). *For any two connected manifolds M and N with affine connections ∇^M and ∇^N respectively, any two points $p \in M$, $q \in N$ and an isomorphism $\Phi : T_pM \to T_qN$, there exists at most one affine transformation $\psi : (M, \nabla^M) \to (N, \nabla^N)$ satisfying*

$$\psi(p) = q, \qquad \text{and} \qquad T_p\psi = \Phi.$$

Proof. Given a pair of affine transformations $\varphi, \psi : (M, \nabla^M) \to (N, \nabla^N)$, let $\mathcal{U} \subset M$ denote the set of all points $p \in M$ such that $\varphi(p) = \psi(p)$ and $T_p\varphi = T_p\psi$. If $p \in \mathcal{U}$, then we can parametrize neighborhoods of $p \in M$ and $q := \varphi(p) = \psi(p) \in N$ via the geodesics through those points and observe that since φ and ψ both map geodesics to geodesics, they must be identical maps on some neighborhood of p. This shows that $\mathcal{U} \subset M$ is an open set. Since it obviously is also closed and M was assumed connected, it follows that $\mathcal{U} = M$, so $\varphi \equiv \psi$. $\square$

EXERCISE 49.7. Show that for the trivial connection ∇ on $\mathbb{R}^n$, $\mathrm{Aff}(\mathbb{R}^n, \nabla)$ is precisely the set of maps of the form $\mathbf{x} \mapsto \mathbf{A}\mathbf{x} + \mathbf{b}$ for $\mathbf{A} \in \mathrm{GL}(n, \mathbb{R})$ and $\mathbf{b} \in \mathbb{R}^n$, i.e. it is the group $\mathrm{Aff}(\mathbb{R}^n)$ from Example 37.9.

EXERCISE 49.8. Show that for the unit sphere $S^n \subset \mathbb{R}^{n+1}$ with its standard metric and Levi-Cività connection ∇, $\mathrm{Aff}(S^n, \nabla) = \mathrm{O}(n+1)$, where the action of $\mathrm{O}(n+1)$ on S^n is defined as the obvious restriction of its natural action on $\mathbb{R}^{n+1}$.

It is important to understand that Theorem 49.6 is a uniqueness result without existence: it says that there is *at most* one affine transformation with a particular value and derivative at one point, but there may also be none at all. The set of affine transformations $(M, \nabla^M) \to (N, \nabla^N)$ can very well be empty, and the group $\mathrm{Aff}(M, \nabla^M)$ might contain nothing other than the identity map.

Exercises 49.7 and 49.8 above demonstrate that for the Levi-Cività connection on a Riemannian manifold, affine transformations need not preserve the metric in general. However, on any manifold whose tangent bundle carries extra geometric structure such as a bundle metric, one can restrict to connections that are compatible with that structure and then consider the group of transformations that preserve both the geometric structure and the connection. This will turn out to be the most useful way to understand isometry groups.

Observe first that every diffeomorphism $\psi : M \to N$ between two manifolds induces a diffeomorphism between the frame bundles of their tangent bundles,

$$\psi_* : F(TM) \to F(TN),$$

sending each frame $(X_1, \dots, X_n) \in F(T_pM)$ to $(T\psi(X_1), \dots, T\psi(X_n)) \in F(T_{\psi(p)}N)$, or equivalently, sending the isomorphism $\phi \in \mathrm{Hom}(\mathbb{R}^n, T_pM)$ to

$$\psi_*\phi := T\psi \circ \phi \in \mathrm{Hom}(\mathbb{R}^n, T_{\psi(p)}N).$$

Expressed in this way, the map ψ_* is clearly equivariant with respect to the right $\mathrm{GL}(n, \mathbb{R})$-actions on the principal bundles $F(TM)$ and $F(TN)$. If TM is also equipped with a G-structure for some Lie subgroup $G \subset \mathrm{GL}(n, \mathbb{R})$, then we can consider the restriction of ψ_* to the G-frame bundle

$$F^G(TM) \subset F(TM),$$

which is a submanifold of $F(TM)$ and also a principal G-bundle. Recall that the fiber $F^G(T_pM)$ over a point p consists of all frames for T_pM that arise from G-compatible trivializations. The standard example to keep in mind is $G = \mathrm{O}(n)$, for which $F^G(TM)$ is the bundle of orthonormal frames. If $G = \mathrm{GL}(n, \mathbb{R})$, then $F^G(TM)$ is simply the bundle of *all* frames $F(TM) \to M$.

DEFINITION 49.9. Suppose M and N are n-manifolds whose tangent bundles are endowed with G-structures for some fixed Lie subgroup $G \subset \mathrm{GL}(n, \mathbb{R})$. We say that a diffeomorphism $\psi : M \to N$ **preserves the G-structures** and write

$$\psi \in \mathrm{Diff}^G(M)$$

if the map $\psi_* : F(TM) \to F(TN)$ sends $F^G(TM)$ to $F^G(TN)$.

EXAMPLE 49.10. For $G = \mathrm{O}(n)$, a G-structure on $TM \to M$ is equivalent to a (positive) bundle metric $g = \langle\ ,\ \rangle$ and thus makes (M, g) a Riemannian manifold. A diffeomorphism $\psi : M \to N$ between two Riemannian manifolds (M, g) and (N, h) then preserves the $\mathrm{O}(n)$-structures if and only if it is an isometry, thus

$$\mathrm{Isom}(M, g) = \mathrm{Diff}^{\mathrm{O}(n)}(M).$$

This observation becomes equally valid for *pseudo*-Riemannian manifolds of arbitrary signature (k,ℓ) if we replace $\mathrm{O}(n)$ with the indefinite orthogonal group $\mathrm{O}(k,\ell)$.

EXAMPLE 49.11. Every manifold M carries a $\mathrm{GL}(n,\mathbb{R})$-structure on its tangent bundle by default, and $\mathrm{Diff}^{\mathrm{GL}(n,\mathbb{R})}(M)$ is just the usual diffeomorphism group $\mathrm{Diff}(M)$. For $G := \mathrm{GL}_+(n,\mathbb{R}) := \{\mathbf{A} \in \mathrm{GL}(n,\mathbb{R}) \mid \det(\mathbf{A}) > 0\}$, a G-structure is equivalent to an orientation of M, and $\mathrm{Diff}^G(M)$ is then the group of orientation-preserving diffeomorphisms.

Example 49.11 shows that the transformation groups $\mathrm{Diff}^G(M)$ need not be finite-dimensional Lie groups in general. Here is a less basic example:

EXAMPLE 49.12. The **linear symplectic group**

$$\mathrm{Sp}(2n) \subset \mathrm{GL}(2n,\mathbb{R})$$

consists of all linear transformations $\mathbf{A} : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ that preserve the **standard symplectic form** $\omega_{\mathrm{std}} \in \Omega^2(\mathbb{R}^{2n})$, meaning $\omega_{\mathrm{std}}(\mathbf{Av},\mathbf{Aw}) = \omega_{\mathrm{std}}(\mathbf{v},\mathbf{w})$ for all $\mathbf{v},\mathbf{w} \in \mathbb{R}^{2n}$, where ω_{std} is defined in global coordinates $(p^1,q^1,\dots,p^n,q^n)$ on $\mathbb{R}^{2n}$ by $\omega_{\mathrm{std}} := \sum_{j=1}^n dp^j \wedge dq^j$. Recall that if (M,ω) is a $2n$-dimensional symplectic manifold, then M admits an atlas of charts that identify $\omega \in \Omega^2(M)$ with ω_{std} in local coordinates. The local frames defined via any atlas of this form endow the tangent bundle $TM \to M$ with an $\mathrm{Sp}(2n)$-structure, and if (M_1,ω_1) and (M_2,ω_2) are two symplectic manifolds, a diffeomorphism $\psi : M_1 \to M_2$ then preserves the $\mathrm{Sp}(2n)$-structure if and only if $\psi^*\omega_2 = \omega_1$, i.e. it is a symplectomorphism, so for any symplectic manifold (M,ω), we have

$$\mathrm{Symp}(M,\omega) = \mathrm{Diff}^{\mathrm{Sp}(2n)}(M).$$

As we saw in Example 49.1, $\mathrm{Symp}(M,\omega)$ is typically a very large group, and certainly not a finite-dimensional Lie group except in trivial cases, e.g. when M is a single point.

DEFINITION 49.13. Given a manifold M with a G-structure and a G-compatible connection ∇ on its tangent bundle, we define the group of **G-compatible affine transformations**

$$\mathrm{Aff}^G(M,\nabla) := \mathrm{Diff}^G(M) \cap \mathrm{Aff}(M,\nabla) \subset \mathrm{Diff}(M).$$

Theorem 49.6 implies that if M is connected, then for any frame $\phi \in F^G(TM)$, the map

$$\mathrm{Aff}^G(M,\nabla) \to F^G(TM) : \psi \mapsto \psi_*\phi$$

is injective. The main result of this lecture, Theorem 49.23, will show in fact that the image of this injection is always a smooth submanifold, whose dimension cannot be predicted by any general formula, but clearly is no larger than the dimension of the frame bundle $F^G(TM)$. For groups like $\mathrm{GL}_+(n,\mathbb{R})$ and $\mathrm{Sp}(2n)$ such that $\mathrm{Diff}^G(M)$ is not finite dimensional, this means that $\mathrm{Aff}^G(M,\nabla) \subset \mathrm{Diff}^G(M)$ is a drastically smaller subgroup. The situation is very different however for $\mathrm{O}(n)$ and $\mathrm{O}(k,\ell)$:

PROPOSITION 49.14. *Every isometry $\psi : (M,g) \to (N,h)$ between pseudo-Riemannian manifolds is also an affine transformation for the respective Levi-Cività connections. In particular, for any pseudo-Riemannian metric g of signature (k,ℓ) on M with Levi-Cività connection ∇, $\mathrm{Isom}(M,g) = \mathrm{Aff}^{\mathrm{O}(k,\ell)}(M,\nabla)$.*

PROOF. If $\psi : M \to N$ satisfies $\psi^*h = g$ and ∇ is the Levi-Cività connection on (N,h), one checks easily that the affine connection $\psi^*\nabla$ on M is symmetric and compatible with g, so the result follows from the uniqueness of the Levi-Cività connection. □

49.3. Infinitesimal transformations and the Killing equation. Let's take a closer look at the spaces of infinitesimal transformations corresponding to the groups $\mathrm{Aff}(M, \nabla)$ and $\mathrm{Isom}(M, g)$. Since the manifold M may be noncompact, it will be important to distinguish between vector fields that admit global flows and those that do not.

Definition 49.15. A vector field $X \in \mathfrak{X}(M)$ is called **complete** if it has a globally-defined flow $\varphi_X^t : M \to M$ for all $t \in \mathbb{R}$.

Assume M is a smooth manifold with an affine connection ∇. A practical characterization of the space of infinitesimal affine transformations on (M, ∇) requires defining the notion of the **Lie derivative of** ∇ with respect to a vector field $X \in \mathfrak{X}(M)$:

$$\mathcal{L}_X \nabla \in \Gamma(T_2^1 M), \qquad (\mathcal{L}_X \nabla)(Y, Z) := \frac{d}{dt} \left[(\varphi_X^t)^* \nabla \right]_Y Z \bigg|_{t=0} .$$

Notice that since the difference between two connections is always a tensor, the Lie derivative of a connection is not another connection, but instead a tensor, i.e. the expression $(\mathcal{L}_X \nabla)(Y, Z)$ defined above is C^∞-linear in both Y and Z. One easily verifies from the definition that it satisfies the Leibniz rule

$$\mathcal{L}_X (\nabla_Y Z) = (\mathcal{L}_X \nabla)(Y, Z) + \nabla_{\mathcal{L}_X Y} Z + \nabla_Y (\mathcal{L}_X Z), \tag{49.1}$$

which can be used in practice as an alternative definition for $\mathcal{L}_X \nabla$. With this notion in place, we have:

Proposition 49.16. *A complete vector field $X \in \mathfrak{X}(M)$ satisfies $\varphi_X^t \in \mathrm{Aff}(M, \nabla)$ for all $t \in \mathbb{R}$ if and only if $\mathcal{L}_X \nabla \equiv 0$.* □

If $X \in \mathfrak{X}(M)$ satisfies $\mathcal{L}_X \nabla = 0$ but is not complete, then its flow defines an affine transformation

$$\varphi_X^t : (\mathcal{O}_X^t, \nabla) \to (\mathcal{O}_X^{-t}, \nabla)$$

between two open subsets $\mathcal{O}_X^{\pm t} \subset M$ for each $t \in \mathbb{R}$, where $\mathcal{O}_X^t \subsetneq M$ for each $t \neq 0$ but $\bigcup_{t>0} \mathcal{O}_X^t = \bigcup_{t<0} \mathcal{O}_X^t = M$. As a mild abuse of terminology and notation, we will dispense with the completeness condition and define the space of **infinitesimal affine transformations**

$$\mathfrak{aff}(M, \nabla) := \left\{ X \in \mathfrak{X}(M) \;\middle|\; \mathcal{L}_X \nabla = 0 \right\}.$$

Note that if $\mathrm{Aff}(M, \nabla)$ is a Lie transformation group and M is noncompact, then $\mathfrak{aff}(M, \nabla)$ might be a strictly larger space than the actual Lie algebra of $\mathrm{Aff}(M, \nabla)$, since it might contain vector fields that are not complete. We will see however in the main theorem that this does not happen if (M, ∇) is geodesically complete.

Exercise 49.17. Derive the following alternative formulas for $\mathcal{L}_X \nabla$:

(a) $(\mathcal{L}_X \nabla)(Y, Z) = [X, \nabla_Y Z] - \nabla_{[X,Y]} Z - \nabla_Y [X, Z]$
(b) $(\mathcal{L}_X \nabla)(Y, Z) = \nabla_Y \nabla_Z X + \nabla_Y (T(X, Z)) - \nabla_{\nabla_Y Z} X - T(X, \nabla_Y Z) + R(X, Y)Z$, where T and R denote the torsion and Riemann tensor respectively for ∇.

Exercise 49.18. Show that if $X \in \mathfrak{aff}(M, \nabla)$, then along any geodesic $\gamma(t) \in M$, $X(t) := X(\gamma(t))$ satisfies the Jacobi equation

$$\nabla_t^2 X + \nabla_t (T(X, \dot\gamma)) + R(X, \dot\gamma)\dot\gamma = 0.$$

Deduce from this a linearized analogue of Theorem 49.6: if M is connected, then any infinitesimal affine transformation is determined by its value and first covariant derivative at one point. Derive from this the bound

$$\dim \mathfrak{aff}(M, \nabla) \leqslant n + n^2 = n(n+1),$$

assuming $\dim M = n$.

On a pseudo-Riemannian manifold (M, g), linearizing the condition $\psi^* g = g$ similarly produces the so-called **Killing equation**

$$\mathcal{L}_X g := \frac{d}{dt}(\varphi_X^t)^* g\bigg|_{t=0} = \frac{d}{dt} g\bigg|_{t=0} = 0,$$

and we define the space of **Killing vector fields** by

$$\mathfrak{isom}(M, g) := \left\{ X \in \mathfrak{X}(M) \;\middle|\; \mathcal{L}_X g = 0 \right\}.$$

Once again, this definition comes with the caveat that $\mathfrak{isom}(M, g)$ may turn out to be strictly larger than the actual Lie algebra of $\mathrm{Isom}(M, g)$, because it may contain vector fields without global flows. In general, the flow of a Killing vector field $X \in \mathfrak{isom}(M, g)$ will define a family of isometries

$$\varphi_X^t : (\mathcal{O}_X^t, g) \to (\mathcal{O}_X^{-t}, g)$$

between open subsets $\mathcal{O}_X^{\pm t} \subset M$, with $\mathcal{O}_X^t = M$ for $t \neq 0$ if and only if X is complete.

For the Levi-Cività connection ∇ on (M, g), Proposition 49.14 implies that isometries are also affine transformations, and it follows that Killing vector fields are also infinitesimal affine transformations:

$$\mathfrak{isom}(M, g) \subset \mathfrak{aff}(M, \nabla).$$

By Exercise 49.18, it follows that $\mathfrak{isom}(M, g)$ is finite dimensional.

Exercise 49.19. Suppose M is a smooth manifold with a symmetric connection ∇, and the associated tensor bundles $T^k_\ell M \to M$ are equipped with the connections naturally determined by ∇. Assuming $X, Y, Z \in \mathfrak{X}(M)$, prove:

(a) For any 1-form $\lambda \in \Omega^1(M)$, $d\lambda(X, Y) = (\nabla_X \lambda)(Y) - (\nabla_Y \lambda)(X)$.

(b) For any 1-form $\lambda \in \Omega^1(M)$, $(\mathcal{L}_X \lambda)(Y) = (\nabla_X \lambda)(Y) + \lambda(\nabla_Y X)$.
Hint: Recall Cartan's magic formula $\mathcal{L}_X \omega = d(\iota_X \omega) + \iota_X(d\omega)$ for the Lie derivative of a differential form (see §14.2 from last semester).

(c) For any $S \in \Gamma(T^0_2 M)$, $(\mathcal{L}_X S)(Y, Z) = (\nabla_X S)(Y, Z) + S(\nabla_Y X, Z) + S(Y, \nabla_Z X)$.
Hint: It suffices (why?) to verify this for tensor fields of the form $\lambda \otimes \mu \in \Gamma(T^0_2 M)$ with $\lambda, \mu \in \Omega^1(M)$. How does the operator $\mathcal{L}_X$ to behave under tensor products?

(d) If ∇ is the Levi-Cività connection for a pseudo-Riemannian metric $g = \langle\ ,\ \rangle$ with associated musical isomorphisms $TM \to T^*M : X \mapsto X_\flat$ and $T^*M \to TM : \lambda \mapsto \lambda^\sharp$, then for $X \in \mathfrak{X}(M)$, the type $(0, 2)$ tensor field $\nabla(X_\flat) \in \Gamma(T^0_2 M)$ is antisymmetric (i.e. it is a differential 2-form) if and only if X satisfies the Killing equation.

The punchline of Exercise 49.19 is that the Killing equation $\mathcal{L}_X g = 0$ is equivalent to the condition

$$\langle \nabla_Y X, Z \rangle + \langle Y, \nabla_Z X \rangle \equiv 0 \qquad \text{for all } Y, Z \in \mathfrak{X}(M), \tag{49.2}$$

which can be expressed more succinctly as the condition that the $(0, 2)$-tensor $\nabla X_\flat$ is antisymmetric. In local coordinates, writing $X = X^i \partial_i$ and $X_\flat = X_i \, dx^i$, the latter condition becomes

$$\nabla_i X_j - \nabla_j X_i = 0. \tag{49.3}$$

Exercise 49.20. Give a second proof that (49.2) is equivalent to $\mathcal{L}_X g = 0$ by deriving a Leibniz rule for $\mathcal{L}_X \langle Y, Z \rangle$ in terms of the Lie derivatives $\mathcal{L}_X Y$ and $\mathcal{L}_X Z$, and using the formula $\mathcal{L}_X Y = [X, Y]$ for $X, Y \in \mathfrak{X}(M)$ (see §6.3 from last semester).

The best possible bound on $\dim \mathfrak{isom}(M, g)$ is now obtained from the observation that if X satisfies (49.2), then the linear map $\nabla X : T_p M \to T_p M$ is antisymmetric at every point $p \in M$. If

$\dim M = n$, the space of antisymmetric maps $T_pM \to T_pM$ has dimension $1 + 2 + \ldots + (n-1) = \frac{(n-1)n}{2}$, and so allowing the additional freedom to specify $X(p)$, Exercise 49.18 implies

$$\dim \mathfrak{isom}(M, g) \leqslant n + \frac{(n-1)n}{2} = \frac{n(n+1)}{2} = \dim \mathrm{O}(n+1)$$

whenever M is connected. In general the dimension of $\mathfrak{isom}(M, g)$ cannot be predicted any more precisely than this: the examples of $\mathbb{R}^n$ and S^n with their standard metrics show that the inequality can sometimes be an equality, but for generic metrics without any special symmetry, one typically expects $\mathrm{Isom}(M, g)$ to be at most a discrete group, so that $\mathfrak{isom}(M, g)$ is trivial.

The next exercise shows that in addition to being finite-dimensional subspaces, $\mathfrak{aff}(M, \nabla)$ and $\mathfrak{isom}(M, g)$ are both Lie subalgebras of $\mathfrak{X}(M)$.

EXERCISE 49.21. Throughout this exercise, fix $X, Y \in \mathfrak{X}(M)$. The Lie bracket $[X, Y] \in \mathfrak{X}(M)$ is traditionally defined via the property that the operators $\mathcal{L}_{[X,Y]}$ and $\mathcal{L}_X\mathcal{L}_Y - \mathcal{L}_Y\mathcal{L}_X$ should match when applied to smooth real-valued functions. The goal of this exercise is to show that they also match when applied to general tensor fields and connections. One immediate consequence of this is that for any two vector fields $X, Y \in \mathfrak{X}(M)$ on a pseudo-Riemannian manifold (M, g) satisfying $\mathcal{L}_X g = 0$ and $\mathcal{L}_Y g = 0$, one also has $\mathcal{L}_{[X,Y]} g = 0$, hence the space of Killing vector fields is a Lie subalgebra of $\mathfrak{X}(M)$.

(a) Deduce from the Jacobi identity and the formula $\mathcal{L}_X Y = [X, Y]$ that $\mathcal{L}_{[X,Y]} Z = \mathcal{L}_X\mathcal{L}_Y Z - \mathcal{L}_Y\mathcal{L}_X Z$ for all $Z \in \mathfrak{X}(M)$.
(b) Show that for $\lambda \in \Omega^1(M)$ and $Z \in \mathfrak{X}(M)$, the Leibniz rule $\mathcal{L}_X(\lambda(Z)) = (\mathcal{L}_X\lambda)(Z) + \lambda(\mathcal{L}_X Z)$ holds, and deduce from this and part (a) that $\mathcal{L}_{[X,Y]}\lambda = \mathcal{L}_X\mathcal{L}_Y\lambda - \mathcal{L}_Y\mathcal{L}_X\lambda$.
(c) Show that for any two tensor fields S, T on M, the Leibniz rule $\mathcal{L}_X(S \otimes T) = \mathcal{L}_X S \otimes T + S \otimes \mathcal{L}_X T$ holds. Deduce from this and parts (a)–(b) via an inductive argument that $\mathcal{L}_{[X,Y]} S = \mathcal{L}_X\mathcal{L}_Y S - \mathcal{L}_Y\mathcal{L}_X S$ holds for tensor fields S of arbitrary rank.
(d) Use part (a) and the Leibniz rule (49.1) to prove $\mathcal{L}_{[X,Y]}\nabla = \mathcal{L}_X\mathcal{L}_Y\nabla - \mathcal{L}_Y\mathcal{L}_X\nabla$ for an affine connection ∇ on M.

If M is endowed with a G-structure on its tangent bundle $TM \to M$, then the condition $\psi_*(F^G(TM)) = F^G(TM)$ on diffeomorphisms $\psi \in \mathrm{Diff}(M)$ can also be linearized to produce a condition on vector fields $X \in \mathfrak{X}(M)$. This relies on the following lemma.

LEMMA 49.22. *There exists a linear map $\mathfrak{X}(M) \to \mathfrak{X}(F(TM)) : X \mapsto F(X)$ such that for all $X \in \mathfrak{X}(M)$, if $p \in M$ is in the domain of the flow φ_X^t, then every frame $\phi \in F(T_pM)$ at p is in the domain of $\varphi_{F(X)}^t$ and*

$$\varphi_{F(X)}^t(\phi) = (\varphi_X^t)_*\phi.$$

PROOF. Given $X \in \mathfrak{X}(M)$, we will use a symmetric connection ∇ on M to write down a formula for $F(X) \in \mathfrak{X}(F(TM))$. We adopt the notational convention of writing elements of $F(TM)$ as pairs (p, ϕ), where $p \in M$ and ϕ belongs to the fiber $F(T_pM)$ over p. Our connection ∇ determines a principal connection on $F(TM)$, and thus a horizontal-vertical splitting of each tangent space $T_{(p,\phi)}F(TM)$ such that the derivative of the projection $F(TM) \to M$ gives a natural isomorphism between $H_{(p,\phi)}F(TM)$ and T_pM. If we regard frames $\phi \in F(T_pM)$ as invertible linear maps $\mathbb{R}^n \to T_pM$, then the vertical subspace at (p, ϕ) has a natural identification with the space of all (not necessarily invertible) linear maps $\mathbb{R}^n \to T_pM$, and we thus obtain an isomorphism

$$T_{(p,\phi)}F(TM) \cong T_pM \times \mathrm{Hom}(\mathbb{R}^n, T_pM),$$

in which the two factors on the right hand side correspond to the horizontal and vertical subspaces respectively. With this identification in place, we claim that the desired vector field $F(X)$ on

$F(TM)$ is given by

$$F(X)(p,\phi) = (X(p), \nabla X(p) \circ \phi) \in T_pM \times \mathrm{Hom}(\mathbb{R}^n, T_pM) = T_{(p,\phi)}F(TM),$$

meaning $F(X)(p,\phi) = \frac{d}{dt}(\varphi_X^t)_*\phi\big|_{t=0}$, where we recall that by definition, $(\varphi_X^t)_*\phi := T_p\varphi_X^t \circ \phi \in F(T_{\varphi_X^t(p)}M) \subset \mathrm{Hom}(\mathbb{R}^n, T_{\varphi_X^t(p)}M)$. Fixing $p \in M$ and $\phi \in F(T_pM)$ and defining the path $\gamma(t) := \varphi_X^t(p)$ in M, the claim is equivalent to the statement that the section $T_p\varphi_X^t \circ \phi \in F(T_{\gamma(t)}M)$ of $F(TM)$ along γ has covariant derivative

$$\nabla_t\big(T_p\varphi_X^t \circ \phi)\big)\big|_{t=0} = \nabla X(p) \circ \phi \in \mathrm{Hom}(\mathbb{R}^n, T_pM) = V_{(p,\phi)}F(TM).$$

In light of the relationship between the connections on $F(TM) \to M$ and $TM \to M$, this in turn is equivalent to the relation

$$\nabla_t\big(T_p\varphi_X^t(Y)\big)\big|_{t=0} = \nabla_Y X \qquad \text{for every } Y \in T_pM.$$

To prove the latter, choose a smooth path $\alpha(s) \in M$ with $\alpha(0) = p$ and $\partial_s\alpha(0) = Y$; using the symmetry of the connection, we then find

$$\nabla_t\big(T_p\varphi_X^t(Y)\big)\big|_{t=0} = \nabla_t\partial_s\varphi_X^t(\alpha(s))\big|_{s=t=0} = \nabla_s\partial_t\varphi_X^t(\alpha(s))\big|_{s=t=0} = \nabla_s\big(X(\alpha(s))\big)\big|_{s=0} = \nabla_Y X.$$

□

Using the linear map $F : \mathfrak{X}(M) \to \mathfrak{X}(F(TM))$ from Lemma 49.22, we can now define

$$\mathfrak{diff}^G(M) := \left\{ X \in \mathfrak{X}(M) \;\middle|\; F(X)(\phi) \in T_\phi(F^G(TM)) \text{ for all } \phi \in F^G(TM) \right\},$$

which is the vector space of all vector fields X for which the induced vector field $F(X)$ on $F(TM)$ restricts to a vector field on the G-frame bundle $F^G(TM)$. In other words, these vector fields are distinguished by the property that the maps $(\varphi_X^t)_*$ on $F(TM)$ preserve $F^G(TM)$ wherever they are defined. The usual caveat applies: vector fields in $\mathfrak{diff}^G(M)$ might not be complete in general, and the same goes for

$$\mathfrak{aff}^G(M,\nabla) := \mathfrak{aff}(M,\nabla) \cap \mathfrak{diff}^G(M) \subset \mathfrak{X}(M),$$

which might end up being strictly larger than the Lie algebra of $\mathrm{Aff}^G(M,\nabla)$, but we will see that this does not happen if (M,∇) is geodesically complete.

49.4. The main result. As is standard in the world of Riemannian manifolds, an arbitrary affine connection ∇ on a manifold M can be called **geodesically complete** if every solution to the geodesic equation $\nabla_t\dot\gamma = 0$ exists for all time $t \in \mathbb{R}$. Equivalently, this means that the domain of the exponential map determined by ∇ is TM, rather than a smaller open subset.

Theorem 49.23. *Assume G is a Lie subgroup of $\mathrm{GL}(n,\mathbb{R})$, M is a connected smooth n-manifold whose tangent bundle $TM \to M$ is equipped with a G-structure, and ∇ is a G-compatible affine connection on M that is geodesically complete. Then the group $\mathrm{Aff}^G(M,\nabla) \subset \mathrm{Diff}(M)$ of affine transformations of M preserving the G-structure admits a Lie group structure for which the action*

$$\mathrm{Aff}^G(M,\nabla) \times F^G(TM) \to F^G(TM) : (\psi,\phi) \mapsto \psi_*\phi$$

is free and proper.[103] *It follows in particular that $\mathrm{Aff}^G(M,\nabla)$ is a Lie transformation group on M, and for any fixed frame $\phi \in F^G(TM)$, the map*

$$\mathrm{Aff}^G(M,\nabla) \to F^G(TM) : \psi \mapsto \psi_*\phi$$

[103] The statement and proof that the action on the frame bundle is free and proper were omitted when we discussed this theorem in the lecture. We did prove that the map from $\mathrm{Aff}^G(M,\nabla)$ to any of its orbits in the frame bundle is an injective immersion, though strictly speaking, one needs properness in order to ensure that its image is also a closed subset and submanifold, thus showing for instance that $\mathrm{Aff}^G(M,\nabla)$ is compact whenever $F^G(TM)$ is compact.

is an embedding sending $\mathrm{Aff}^G(M, \nabla)$ *diffeomorphically onto a closed subset and smooth submanifold of* $F^G(TM)$*. Moreover, every vector field in the space* $\mathfrak{aff}^G(M, \nabla)$ *of infinitesimal affine transformations preserving the* G*-structure is complete, hence* $\mathfrak{aff}^G(M, \nabla)$ *is naturally isomorphic to the Lie algebra of* $\mathrm{Aff}^G(M, \nabla)$*.*

In light of Proposition 49.14, an immediate consequence of Theorem 49.23 is the following fundamental result in Riemannian geometry, originally due to Myers and Steenrod [**MS39**]. It can also be stated for general pseudo-Riemannian manifolds at the cost of replacing $\mathrm{O}(n)$ with $\mathrm{O}(k, \ell)$ for some $k + \ell = n$. In order to state both results together, let us abbreviate

$$F^{\mathrm{O}}(TM) := F^{\mathrm{O}(k,\ell)}(TM)$$

when (M, g) is a pseudo-Riemannian manifold of signature (k, ℓ).

Corollary 49.24 (in light of Proposition 49.14). *On any geodesically complete connected pseudo-Riemannian manifold* (M, g)*, the group of isometries* $\mathrm{Isom}(M, g)$ *is a Lie transformation group with*

$$\dim \mathrm{Isom}(M, g) \leqslant \dim F^{\mathrm{O}}(TM) = \frac{n(n+1)}{2},$$

and its Lie algebra is the space $\mathfrak{isom}(M, g)$ *of Killing vector fields, all of which are complete.* □

Remark 49.25. Theorem 49.23 and Corollary 49.24 have obvious extensions to disconnected manifolds, at least if the number of connected components is finite. The caveat about the case where M has infinitely-many connected components is that $\mathrm{Aff}^G(M, \nabla)$ and $\mathrm{Isom}(M, g)$ may then have *uncountably* many connected components, due to the fact that the set of permutations of a countably-infinite set is uncountable. In this case they will fail to satisfy the second countability axiom (or equivalently, they will not be separable), and thus cannot be called "manifolds" according to our definitions, but the cardinality of the set of connected components is really the only issue—they are still locally Euclidean and have natural smooth structures for which the action on M is smooth.

For a compact Riemannian manifold (M, g), one obtains a further corollary from the statement that $\mathrm{Isom}(M, g)$ embeds into $F^{\mathrm{O}}(TM)$ as a *closed* subset.[104] The key point here is that since $\mathrm{O}(n)$ is a compact group, the frame bundle $F^{\mathrm{O}}(TM)$ is also compact; this is false however for indefinite metrics, since $\mathrm{O}(k, \ell)$ is not compact when $k, \ell \geqslant 1$.

Corollary 49.26. *For any compact Riemannian manifold* (M, g)*, the group* $\mathrm{Isom}(M, g)$ *is compact.* □

Exercise 49.27. The following gives a counterexample to the pseudo-Riemannian analogue of Corollary 49.26:

(a) Show that for any two linearly-independent vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^2$, there exists a nontrivial symmetric bilinear form $\langle\ ,\ \rangle$ on $\mathbb{R}^2$ with $\langle \mathbf{v}, \mathbf{v} \rangle = \langle \mathbf{w}, \mathbf{w} \rangle = 0$, and it is unique up to scaling, and is nondegenerate with signature $(1, 1)$.

(b) A matrix $\mathbf{A} \in \mathrm{SL}(2, \mathbb{Z})$ with $\mathrm{tr}(\mathbf{A}) > 2$ necessarily has two distinct real eigenvalues of the form $\lambda > 1$ and $1/\lambda < 1$. Show that for any such matrix, the bilinear form in part (a) can be chosen so that $\mathbf{A}$ preserves it.
Hint: $\mathbf{A}$ *has two linearly-independent eigenvectors.*

[104] Corollary 49.26 was never mentioned in lecture since it depends on the properness of the action of $\mathrm{Isom}(M, g)$ on $F^{\mathrm{O}}(TM)$, which also was not mentioned.

(c) Use the matrix $\mathbf{A}$ and bilinear form $\langle\ ,\ \rangle$ from parts (a) and (b) to construct a pseudo-Riemannian metric g on $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$ of signature $(1,1)$ such that the stabilizer $G_p \subset \mathrm{Isom}(\mathbb{T}^2, g)$ of some point $p \in \mathbb{T}^2$ is not compact.
Hint: G_p acts on $T_p\mathbb{T}^2$ by linear transformations—show that this transformation group contains an unbounded subgroup of $\mathrm{SL}(2,\mathbb{Z})$.

Remark 49.28. We will not prove it here, but the geodesic completeness condition in Theorem 49.23 and Corollary 49.24 is not necessary. We will make use of it in our proof so that we do not have to worry about the distinction between infinitesimal transformations that do or do not admit global flows. The original proof of Corollary 49.24 in [MS39] does not require completeness. One can easily deduce from it the $G = \{\mathrm{Id}\}$ case of Theorem 49.23, from which the general case follows, as we will show below.

To set up the proof of Theorem 49.23, we will show that the result follows essentially from the special case in which G is the trivial group. For $G = \{\mathbb{1}\} \subset \mathrm{GL}(n,\mathbb{R})$, a G-structure on the tangent bundle of an n-manifold M is equivalent to a global trivialization Φ of the tangent bundle, also known as a **parallelization** of M. Such a structure determines a global frame $X_1, \dots, X_n \in \mathfrak{X}(M)$, and thus a distinguished n-dimensional space

$$\mathfrak{V} \subset \mathfrak{X}(M)$$

consisting of the linear combinations of $X_1, \dots, X_n$ with constant coefficients; equivalently, $\mathfrak{V}$ is the space of vector fields that look constant in the trivialization Φ. An affine connection ∇ on M is compatible with this structure if and only if the vector fields in $\mathfrak{V}$ are all parallel, and it follows that there is only one such connection, namely the *trivial* connection with respect to Φ. As a consequence, this is another situation in which diffeomorphisms compatible with the G-structure are automatically also affine transformations: we will refer to these maps as **automorphisms** of (M,Φ) and denote the group by

$$\mathrm{Aut}(M,\Phi) := \mathrm{Diff}^{\{\mathrm{Id}\}}(M) = \mathrm{Aff}^{\{\mathrm{Id}\}}(M,\nabla) = \left\{\psi \in \mathrm{Diff}(M) \;\middle|\; \psi^*V = V \text{ for all } V \in \mathfrak{V}\right\}.$$

Differentiating the relation $(\varphi_X^t)^*V = V$ with respect to t, one finds that the corresponding space of infinitesimal transformations is

$$\mathfrak{aut}(M,\Phi) := \mathfrak{diff}^{\{\mathrm{Id}\}}(M) = \mathfrak{aff}^{\{\mathrm{Id}\}}(M,\nabla) = \left\{X \in \mathfrak{X}(M) \;\middle|\; [X,V] = 0 \text{ for all } V \in \mathfrak{V}\right\}.$$

Proposition 49.29. *If $TM \to M$ is endowed with a G-structure for some Lie subgroup $G \subset \mathrm{GL}(n,\mathbb{R})$ and ∇ is a G-compatible affine connection, then the manifold $F^G(TM)$ admits a parallelization Φ such that the image of the injective map*

$$\mathrm{Aff}^G(M,\nabla) \hookrightarrow \mathrm{Diff}(F^G(TM)) : \psi \mapsto \psi_*$$

is a closed subgroup of $\mathrm{Aut}(F^G(TM),\Phi) \subset \mathrm{Diff}(F^G(TM))$. Moreover, if (M,∇) is geodesically complete, then every vector field in $\mathfrak{aut}(F^G(TM),\Phi)$ is complete.

The proof of this proposition has three main steps, the first of which is to identify which diffeomorphisms $F^G(TM) \to F^G(TM)$ take the form ψ_* for some $\psi \in \mathrm{Diff}^G(M)$. To simplify notation, let us write

$$\pi : E \to M$$

for the G-frame bundle $F^G(TM)$ of TM. Regarding frames $\phi \in E_p$ as linear maps $\phi : \mathbb{R}^n \to T_pM$, there is an $\mathbb{R}^n$-valued **tautological 1-form** (cf. Exercise 46.16)

$$\theta \in \Omega^1(E,\mathbb{R}^n)$$

defined by

$$\theta_\phi := \phi^{-1} \circ \pi_* : T_\phi E \to \mathbb{R}^n, \qquad \text{for each } \phi \in E_p \subset \mathrm{Hom}(\mathbb{R}^n, T_pM),\ p \in M.$$

It is straightforward to verify that θ is equivariant with respect to the canonical representation of $G \subset \mathrm{GL}(n, \mathbb{R})$ on $\mathbb{R}^n$, meaning

$$R_g^*\theta = g^{-1} \circ \theta \qquad \text{for all } g \in G. \tag{49.4}$$

In fact, θ is the equivariant 1-form corresponding under the isomorphism of Theorem 46.3 to the bundle-valued form $\mathbb{1} \in \Omega^1(M, TM)$, i.e. the identity map $TM \to TM$.

In the following, a diffeomorphism $\Psi : E \to E$, will be called **fiber preserving** if the image of every fiber is contained in a (possibly different) fiber.

Lemma 49.30. *A fiber-preserving diffeomorphism $\Psi : E \to E$ satisfies $\Psi^*\theta = \theta$ if and only if $\Psi = \psi_*$ for some $\psi \in \mathrm{Diff}^G(M)$.*

Proof. If Ψ is fiber preserving, we can write $\Psi(E_p) = E_{\psi(p)}$ for each $p \in M$, defining in this way a diffeomorphism $\psi : M \to M$. For $p \in M$ and $\phi \in E_p$, $\Psi(\phi)$ is an isomorphism $\mathbb{R}^n \to T_{\psi(p)}M$, and since $\pi \circ \Psi = \psi \circ \pi$,

$$\begin{aligned}(\Psi^*\theta)_\phi &= \theta_{\Psi(\phi)} \circ T\Psi = \Psi(\phi)^{-1} \circ \pi_* \circ T\Psi = \Psi(\phi)^{-1} \circ T(\pi \circ \Psi) = \Psi(\phi)^{-1} \circ T(\psi \circ \pi) \\ &= \left(\Psi(\phi)^{-1} \circ T\psi\right) \circ \pi_*.\end{aligned}$$

This matches $\theta_\phi = \phi^{-1} \circ \pi_*$ if and only if $\phi^{-1} = \Psi(\phi)^{-1} \circ T\psi$, or equivalently $\Psi(\phi) = T\psi \circ \phi$, which means $\Psi = \psi_*$ and $\psi \in \mathrm{Diff}^G(M)$. $\square$

Remark 49.31. If the fibers of E are connected, then the assumption in Lemma 49.30 that $\Psi : E \to E$ is fiber preserving is redundant, because any two points in the same fiber can be connected via a finite sequence of flow lines of vertical vector fields, and $\Psi^*\theta = \theta$ guarantees that Ψ preserves verticality since $\ker\theta = VE$. This could be relevant for instance if M is oriented and one restricts attention to orientation-preserving isometries, since the fibers of $F^{\mathrm{SO}(n)}(TM)$ are then connected. However, it would be too restrictive to assume that E has connected fibers in general.

In the second step, we suppose ∇ is a G-compatible affine connection on M, which corresponds to a principal connection on the G-frame bundle $E \to M$ and thus gives rise to a horizontal subbundle $HE \subset TE$ and connection 1-form $A \in \Omega^1(E, \mathfrak{g})$.

Lemma 49.32. *A transformation $\psi \in \mathrm{Diff}^G(M)$ is affine with respect to ∇ if and only if the induced transformation $\Psi := \psi_* : E \to E$ satisfies $\Psi^*A = A$.*

Proof. Assuming $\Psi = \psi_* : E \to E$ for some $\psi \in \mathrm{Diff}^G(M)$, the map Ψ is G-equivariant, thus writing an arbitrary vertical tangent vector at $\phi \in E_p$ as the value of a fundamental vector field $Z^F(\phi) = \partial_t \left(\phi \cdot \exp(tZ)\right)|_{t=0}$ for some $Z \in \mathfrak{g}$, we have

$$\begin{aligned}(\Psi^*A)_\phi\left(Z^F(\phi)\right) &= A_{\Psi(\phi)} \circ T\Psi\left(\partial_t \left(\phi \cdot \exp(tZ)\right)|_{t=0}\right) \\ &= A_{\Psi(\phi)}\left(\partial_t \left(\Psi(\phi \cdot \exp(tZ))\right)|_{t=0}\right) \\ &= A_{\Psi(\phi)}\left(\partial_t \left(\Psi(\phi) \cdot \exp(tZ)\right)|_{t=0}\right) = A_{\Psi(\phi)}\left(Z^F(\Psi(\phi))\right) = Z = A_\phi\left(Z^F(\phi)\right).\end{aligned}$$

This shows on the one hand that A and Ψ^*A match on the vertical subbundle VE, and since Ψ^*A is also equivariant due to the equivariance of Ψ and A, it also shows that Ψ^*A is another connection 1-form on $E \to M$, thus defining another G-compatible connection ∇' on $TM \to M$. We claim $\nabla' = \psi^*\nabla$. This will hold if and only if for every path $\gamma(t) \in M$ and every vector field $Y(t) \in T_{\gamma(t)}M$ along γ,

$$T\psi(\nabla'_t Y(t)) = \nabla_t\left(T\psi(Y(t))\right) \in T_{\psi(\gamma(t))}M \tag{49.5}$$

for all t. To prove (49.5), choose a frame $\phi(t) \in E_{\gamma(t)}$ along γ that is parallel with respect to ∇', and observe that since $T\Psi$ maps $\ker \Psi^* A$ to $\ker A$, the frame $\Psi(\phi(t)) \in E_{\psi(\gamma(t))}$ along $\psi \circ \gamma$ is likewise parallel with respect to ∇. There is then a unique function $f(t) \in \mathbb{R}^n$ such that $Y(t) = \phi(t) f(t)$, and $\phi(t)$ being parallel implies $\nabla'_t Y(t) = \phi(t)\dot{f}(t)$. Since $\Psi = \psi_*$, we also have $T\psi(Y(t)) = \Psi(\phi(t)) f(t)$, implying

$$\nabla_t \left(T\psi(Y(t))\right) = \Psi(\phi(t))\dot{f}(t) = T\psi\left(\phi(t)\dot{f}(t)\right) = T\psi\left(\nabla'_t Y(t)\right),$$

which proves the claim. We conclude $\psi^*\nabla = \nabla$ if and only if $\Psi^* A = A$. □

We can now define the parallelization Φ on $E = F^G(TM)$ promised by Proposition 49.29: it is determined by the two vector-valued 1-forms $\theta \in \Omega^1(E, \mathbb{R}^n)$ and $A \in \Omega^1(E, \mathfrak{g})$. Indeed, since θ takes values in $\mathbb{R}^n$, it can be regarded as a collection of n real-valued 1-forms $\theta^1, \ldots, \theta^n \in \Omega^1(E)$, which are linearly independent at every point $\phi \in E_p \subset \operatorname{Hom}(\mathbb{R}^n, T_p M)$ since for the standard basis $\mathbf{e}_1, \ldots, \mathbf{e}_n \in \mathbb{R}^n$, $\theta^i(\operatorname{Hor}_\phi(\phi(\mathbf{e}_j))) = \delta^i_j$ by definition. Moreover, all of these vanish on the vertical subspace $V_\phi E$. Choosing a basis $Z_1, \ldots, Z_m$ of $\mathfrak{g}$, we can similarly turn $A \in \Omega^1(E, \mathfrak{g})$ into a collection of real-valued 1-forms $A^j \in \Omega^1(E)$ with $A = A^j Z_j$, which are linearly independent at every point and all vanish on HE. It follows that $(\theta^1, \ldots, \theta^n, A^1, \ldots, A^m)$ defines a global frame for T^*E. Its dual frame

$$\eta_1, \ldots, \eta_n, \xi_1, \ldots, \xi_m \in \mathfrak{X}(E),$$

defined via the conditions $\theta^i(\eta_j) = A^i(\xi_j) = \delta^i_j$ and $\theta^i(\xi_j) = A^i(\eta_j) = 0$ for all i, j, is a collection of nowhere-vanishing vector fields that form a basis of TE at every point and thus define a parallelization Φ of E. A diffeomorphism $\Psi : E \to E$ then belongs to $\operatorname{Aut}(E, \Phi)$ if and only if it preserves all of the 1-forms $\theta^1, \ldots, \theta^n$ and $A^1, \ldots, A^m$, which means it satisfies $\Psi^*\theta = \theta$ and $\Psi^* A = A$. By Lemmas 49.30 and Lemma 49.32, the image of $\operatorname{Aff}^G(M, \nabla)$ under the map $\psi \mapsto \psi_* \in \operatorname{Diff}(E)$ is thus the closed subgroup

$$\left\{\Psi \in \operatorname{Aut}(E, \Phi) \;\middle|\; \Psi \text{ is fiber preserving}\right\} \subset \operatorname{Aut}(E, \Phi).$$

To complete the proof of Proposition 49.29, we still need to show that for the parallelization Φ defined on E via the vector fields $\eta_1, \ldots, \eta_n$ and $\xi_1, \ldots \xi_m$, all vector fields in $\mathfrak{aut}(E, \Phi)$ are complete if (M, ∇) is geodesically complete. The relation to geodesic completeness comes from the horizontal vector fields η_i and their linear combinations, whose flow lines turn out to be horizontal lifts of geodesics:

LEMMA 49.33. *For any constants $c^1, \ldots, c^n \in \mathbb{R}$, every flow line of the vector field $V := c^i \eta_i \in \mathfrak{X}(E)$ is of the form*

$$\phi(t) \in E_{\gamma(t)} \subset E,$$

where $\gamma(t) \in M$ is a geodesic with respect to ∇ and $\phi(t)$ is a parallel frame along γ. Conversely, every path in E of this form is a flow line of some vector field that is a linear combination of $\eta_1, \ldots, \eta_n$ with constant coefficients.

PROOF. Let $V = c^i \eta_i \in \mathfrak{X}(E)$ for constants $c^1, \ldots, c^n \in \mathbb{R}$. Since V always points in horizontal directions, a flow line of V is a parallel frame $\phi(t) = (X_1(t), \ldots, X_n(t))$ along some path $\gamma(t) \in M$, and γ itself is then determined by the condition $\dot{\gamma}(t) = c^i X_i(t)$. It follows that $\dot{\gamma}$ is also parallel along γ, so γ is a geodesic. Conversely, if γ is a geodesic and $\phi(t)$ is any parallel frame along γ, then there exist unique functions $c^i(t)$ for $i = 1, \ldots, n$ such that $\dot{\gamma}(t) = c^i(t) X_i(t)$, and the geodesic equation then implies that the functions $c^i(t)$ are constant, making $\phi(t) \in E$ a flow line of $c^i \eta_i$. □

Here is a useful tool for proving that a vector field is complete.

PROPOSITION 49.34. *Suppose $S \subset \mathfrak{X}(M)$ is a set of vector fields on a manifold M with the following properties:*

(1) *Every* $V \in S$ *is complete.*
(2) *Any two points* $p, q \in M$ *in the same connected component of* M *are related to each other by* $q = \varphi_{V_1}^{t_1} \circ \ldots \circ \varphi_{V_N}^{t_N}(p)$ *for some* $N \in \mathbb{N}$, $V_1, \ldots, V_N \in S$ *and* $t_1, \ldots, t_N \in \mathbb{R}$.

Then any vector field $X \in \mathfrak{X}(M)$ *that commutes with every* $V \in S$ *is complete.*

Proof. Since the flow of X preserves connected components, we can assume without loss of generality that M is connected. Pick a point $p \in M$ and $\epsilon > 0$ such that $\varphi_X^t(p)$ is defined for all $t \in [-\epsilon, \epsilon]$. Then for any $V \in S$, the condition $[X, V] = 0$ implies via Theorem 6.9 from the first semester that for every $s \in \mathbb{R}$, $\varphi_X^t(\varphi_V^s(p))$ is also defined for all $t \in [-\epsilon, \epsilon]$ and matches $\varphi_V^s(\varphi_X^t(p))$. This shows that for every $t \in [-\epsilon, \epsilon]$, φ_X^t is defined on every point that lies on a flow line of any $V \in S$ through p. Continuing inductively from these points via flow lines of other vector fields in S, this leads to the conclusion that φ_X^t is defined everywhere on M for all $t \in [-\epsilon, \epsilon]$, and by iteration, it follows that $\varphi_t : M \to M$ is defined for all $t \in \mathbb{R}$. □

Conclusion of the proof of Proposition 49.29. We have already deduced from Lemmas 49.30 and 49.32 that the map $\psi \mapsto \psi_*$ identifies $\mathrm{Aff}^G(M, \nabla)$ with a closed subgroup of $\mathrm{Aut}(E, \Phi)$. If ∇ on M is geodesically complete, then it follows from Lemma 49.33 that the linear combinations of $\eta_1, \ldots, \eta_n$ with constant coefficients are all complete vector fields on E. The fundamental vector fields $Z^F \in \mathfrak{X}(E)$ for $Z \in \mathfrak{g}$ are also complete, since their flows are obtained by acting on E with $\exp(tZ) \in G$. Now if $p, q \in M$ are two sufficiently close points with frames $\phi \in E_p$ and $\phi' \in E_q$ that are also sufficiently close in E, then q can be reached from p by a geodesic, and it follows in turn from Lemma 49.33 that

$$\phi' = \varphi_{Z^F}^1 \circ \varphi_{c^i \eta_i}(\phi)$$

for some $Z \in \mathfrak{g}$ and $c^1, \ldots, c^n \in \mathbb{R}$. It follows that the set $S \subset \mathfrak{X}(E)$ consisting of all fundamental vector fields and all linear combinations $c^i \eta_i$ satisfies the two hypotheses of Proposition 49.34. Any $\xi \in \mathfrak{aut}(E, \Phi)$ commutes with all of these vector fields, and is therefore complete. □

The proof of the following will be the main topic of the next section.

Theorem 49.35. *Assume* M *is a smooth connected* n*-manifold equipped with a parallelization* Φ *such that every infinitesimal automorphism* $X \in \mathfrak{aut}(M, \Phi) \subset \mathfrak{X}(M)$ *is complete. Then* $\mathrm{Aut}(M, \Phi)$ *is a Lie transformation group acting freely and properly on* M*.*

Proof of Theorem 49.23 modulo Theorem 49.35. Using the parallelization Φ of $E = F^G(TM)$ constructed above, the hypothesis that (M, ∇) is geodesically complete establishes via Proposition 49.29 the main hypothesis of Theorem 49.35 that vector fields in $\mathfrak{aut}(E, \Phi)$ are complete, though we need to be a bit careful about the possibility that E may be disconnected. Let us ignore this detail for now and just assume E is connected. Theorem 49.35 then gives $\mathrm{Aut}(E, \Phi)$ the structure of a Lie transformation group acting freely and properly on E, and it contains a closed subgroup that is the image of $\mathrm{Aff}^G(M, \nabla)$ under the injection $\mathrm{Aff}^G(M, \nabla) \hookrightarrow \mathrm{Diff}(E) : \psi \mapsto \psi_*$. Since closed subgroups are Lie subgroups, this endows $\mathrm{Aff}^G(M, \nabla)$ with the structure of a Lie transformation group, which also acts freely and properly on $F^G(TM)$, and it follows from this that $\mathrm{Aff}^G(M, \nabla)$ also acts smoothly on M. The completeness of every $X \in \mathfrak{aff}^G(M, \nabla)$ now follows from the fact that the map $F : \mathfrak{X}(M) \to \mathfrak{X}(E)$ in Lemma 49.22 sends each $X \in \mathrm{Aff}^G(M, \nabla)$ to something in $\mathfrak{aut}(E, \Phi)$. □

Exercise 49.36. Fix the proof of Theorem 49.23 above to work in cases where M is connected but $E = F^G(TM)$ is not.
Hint: All that really matters is the structure of the identity component of $\mathrm{Aff}^G(M, \nabla)$.

49.5. Automorphisms of a parallelization. Throughout this section, assume M is a smooth n-manifold equipped with a parallelization Φ, thus defining a distinguished n-dimensional subspace

$$\mathfrak{V} \subset \mathfrak{X}(M)$$

consisting of vector fields that look constant in the trivialization Φ. Let us restate Theorem 49.35, which still needs to be proved.

THEOREM 49.37. *If every vector field in $\mathfrak{aut}(M, \Phi)$ is complete, then $\mathrm{Aut}(M, \Phi)$ is a Lie transformation group acting freely and properly on M, so in particular, for any point $p \in M$, the map $\mathrm{Aut}(M, \Phi) \to M : \psi \mapsto \psi(p)$ is an embedding that maps $\mathrm{Aut}(M, \Phi)$ diffeomorphically to a closed subset and smooth submanifold of M.*[105]

We already know from the results of §49.3 that $\dim \mathfrak{aut}(M, \Phi) < \infty$, though we will see an easier way to prove this below. As outlined in §49.1, the crucial step in the proof is then to show that every element of $\mathrm{Aut}(M, \Phi)$ in some neighborhood of the identity arises from the flow of a vector field in $\mathfrak{aut}(M, \Phi)$.

The proof of this will necessarily seem a bit technical, so we will start with an outline and postpone the proofs of some lemmas until the next subsection. The argument can be summarized with three main claims and a few subsequent comments.

Claim 1: Two maps $\varphi, \psi \in \mathrm{Aut}(M, \Phi)$ are identical if and only if they match at one point. This will be an easy consequence of the fact that the diffeomorphisms $\psi : M \to M$ in $\mathrm{Aut}(M, \Phi)$ preserve all of the vector fields $V \in \mathfrak{V}$, whose flows can be used to connect any point in M to any other. This shows that the action of $\mathrm{Aut}(M, \Phi)$ on M is free, and the same observation will also imply that this action is proper.

The next claim is essentially a linearization of claim 1, and the best way to express it is via an intelligent choice of affine connection on M. The parallelization Φ determines a natural connection ∇, for which the vector fields in $\mathfrak{V}$ are parallel, implying that their flow lines are the geodesics. But Φ also determines another distinguished connection ∇^Φ on TM, defined by

$$\nabla^\Phi_Y X := \nabla_Y X + T(X, Y),$$

where T is the torsion tensor of ∇.

Claim 2: A vector field $X \in \mathfrak{X}(M)$ is in $\mathfrak{aut}(M, \Phi)$ if and only if $\nabla^\Phi X = 0$, and every flow line of a vector field in $\mathfrak{aut}(M, \Phi)$ is a geodesic with respect to ∇. We can prove these statements right away: since $\nabla V = 0$ for $V \in \mathfrak{V}$ and the set of all $V \in \mathfrak{V}$ spans TM at every point, another vector field $X \in \mathfrak{X}(M)$ satisfies $\nabla^\Phi X = 0$ if and only if $\nabla^\Phi_V X = 0$ for all $V \in \mathfrak{V}$, which then means

$$0 = \nabla^\Phi_V X = \nabla_V X + T(X, V) = \nabla_V X + \nabla_X V - \nabla_V X - [X, V] = -[X, V],$$

and thus $X \in \mathfrak{aut}(M, \Phi)$. The antisymmetry of the torsion tensor implies moreover that ∇ and ∇^Φ have the same geodesics, which therefore include the flow lines of any vector field that is parallel with respect to either connection.

An immediate consequence of claim 2 is that nontrivial vector fields in $\mathfrak{aut}(M, \Phi)$ are nowhere zero, so they span a smooth subbundle $E \subset TM$ with the property that at every point $p \in M$, every $Y \in E_p$ is the value of a unique vector field in $\mathfrak{aut}(M, \Phi)$. We can choose a complementary subbundle $E^\perp \subset TM$ and write

$$TM = E \oplus E^\perp.$$

[105] The version of this result proved in lecture did not mention properness, and strictly speaking, it only proved that the bijection of $\mathrm{Aut}(M, \Phi)$ to each of its orbits in M is an injective immersion, not necessarily an embedding.

Now comes the main event: suppose $\psi_k \in \mathrm{Aut}(M, \Phi)$ is a sequence converging in C^∞_{loc} to the identity map with $\psi_k \neq \mathrm{Id}$ for all k. Using the connection ∇, we can write

$$\psi_k = \exp \circ X_k,$$

for a sequence of nontrivial vector fields $X_k \in \mathfrak{X}(M)$ that converge in C^∞_{loc} to 0. A slightly subtle issue here is that since the convergence of ψ_k is uniform only on compact subsets, it may only be possible to write $\psi_k = \exp \circ X_k$ on a nested sequence of compact subsets $\mathcal{U}_k \subset \mathcal{U}_{k+1} \subset \ldots M$ whose union is M. This detail will need to be handled with some care, but for this outline we shall ignore it.

Claim 3: After passing to a subsequence, there exists a sequence of positive numbers $\tau_k \to 0$ such that X_k/τ_k converges in C^∞_{loc} to a nontrivial element

$$X_\infty \in \mathfrak{aut}(M, \Phi).$$

The proof of this statement is the hardest step, but we can give it a quick summary: it is based on the Arzelà-Ascoli theorem. The main reason it works is that the connection ∇^Φ characterizes the linearization at $\mathrm{Id} : M \to M$ of the nonlinear condition satisfied by ψ_k, implying that the vector fields X_k *nearly* satisfy $\nabla^\Phi X_k = 0$. What actually happens is that there exists a sequence of connections ∇^k, which depend on the choice of maps ψ_k but converge to ∇^Φ as a result of the convergence $\psi_k \to \mathrm{Id}$, such that $\nabla^k X_k = 0$ for all k. Since the ∇^k are linear operators, we can then introduce the rescaling constants $\tau_k > 0$ and write

$$\nabla^k \widehat{X}_k = 0, \qquad \text{where} \qquad \widehat{X}_k := \frac{X_k}{\tau_k}$$

for all k. If you write down what this equation looks like in local coordinates, you find a convergent sequence of Christoffel symbols $(\Gamma_{(k)})^\ell_{ij}$ such that the components $\widehat{X}^\ell_k$ of $\widehat{X}_k$ satisfy

$$\partial_i \widehat{X}^\ell_k + (\Gamma_{(k)})^\ell_{ij} \widehat{X}^j_k = 0.$$

This is the strongest type of first-order PDE one could ever hope to consider: it determines all of the first partial derivatives of $\widehat{X}_k$ in terms of the values of $\widehat{X}_k$. (Compare this with e.g. the Cauchy-Riemann equations, which constrain certain linear combinations of the first partial derivatives, but do not completely determine them.) In particular, if we now choose the sequence $\tau_k > 0$ so that $\widehat{X}_k$ on some compact subset is C^0-bounded but not converging to 0, then it follows that $\widehat{X}$ is also C^1-bounded, and plugging this new information into the PDE implies that it is also C^2-bounded, and so forth. The conclusion is that $\widehat{X}_k$ will be uniformly C^m-bounded on every compact subset for every $m \in \mathbb{N}$, so by the Arzelà-Ascoli theorem, it has a C^∞_{loc}-convergent subsequence. The limit of this sequence will satisfy $\nabla^\Phi X_\infty = 0$ since $\nabla^k \widehat{X}_k = 0$ and $\nabla^k \to \nabla^\Phi$.

With claim 3 in place, we can use the splitting $TM = E \oplus E^\perp$ established by claim 2 to write

$$X_k = Y_k + Z_k,$$

where Y_k and Z_k are sections of E and $E^\perp$ respectively, and both converge in C^∞_{loc} to 0 as $k \to \infty$. Choosing a point $p \in M$, there is a unique sequence $\widehat{Y}_k \in \mathfrak{aut}(M, \Phi)$ such that $\widehat{Y}_k(p) = Y_k$, and it converges in C^∞_{loc} to 0. Since every vector field in $\mathfrak{aut}(M, \Phi)$ is assumed complete, we can use its flow to define another sequence $f_k \in \mathrm{Aut}(M, \Phi)$ converging to the identity by

$$f_k := \varphi^1_{\widehat{Y}_k} = \exp \circ \widehat{Y}_k.$$

If we assume that ψ_k cannot similarly be expressed as the time 1 flow of a sequence of vector fields in $\mathfrak{aut}(M, \Phi)$ converging to 0, then it follows via claim 1 that

$$Z_k(p) \neq 0 \in E^\perp_p \text{ for all } k.$$

This is the step where it is important to know that vector fields in $\mathfrak{aut}(M,\Phi)$ are complete—if we did not know this, then the maps f_k could only be defined on a nested sequence of domains $\mathcal{U}_k \subset \mathcal{U}_{k+1} \subset M$ exhausting M, and ψ_k could potentially match f_k on these domains (implying $Z_k(p) = 0$) without being globally expressible via the flows of vector fields in $\mathfrak{aut}(M,\Phi)$.

The final step will be to consider the sequence

$$f_k^{-1} \circ \psi_k \in \mathrm{Aut}(M,\Phi),$$

which also converges to the identity and can be written as $\exp \circ V_k$ for a sequence of vector fields V_k that, by claim 3, will necessarily converge to a nontrivial element $V_\infty \in \mathfrak{aut}(M,\Phi)$ after suitable rescaling. But the fact that $\psi_k = \exp \circ (Y_k + Z_k)$ and $f_k = \exp \circ \widehat{Y}_k$ with $Y_k(p) = \widehat{Y}_k(p)$ and $Z_k(p) \neq 0$ will imply in this situation that $V_\infty(p)$ is a nontrivial vector in $E_p^\perp$, which contradicts the definition of the subbundle E.

You may at this point want to go back and reread the proof of the closed subgroup theorem (Theorem 41.2), as the proof outlined above is similar to it in many respects. The fact that our maps ψ_k are not already known to live inside some finite-dimensional Lie group makes some steps harder, and in particular, the complete proof of claim 3 will require more serious analysis. The approach taken below is heavily inspired by methods that are standard in the theory of elliptic PDEs, where one often uses analytical estimates for certain differential operators to establish compactness results via the Arzelà-Ascoli theorem. The estimates required are much easier in our situation than in actual elliptic theory, because our PDE is overdetermined to the point that everything we need to know can be deduced from the theory of *ordinary* differential equations. It is possible however to generalize this approach to other situations in which more serious elliptic estimates (typically in Sobolev spaces) would be required, e.g. in order to prove that the group of holomorphic diffeomorphisms on a compact complex manifold is a Lie transformation group.

We now proceed with the details of the outline given above.

49.6. The technical part. Here are the remaining details in the proof of Theorem 49.37.

We already observed that the space $\mathfrak{aut}(M,\Phi)$ of infinitesimal automorphisms can be characterized as the solution set of the first-order linear PDE $\nabla^\Phi X = 0$, where ∇^Φ is a particular affine connection determined by the parallelization Φ. Let us write down the corresponding nonlinear PDE satisfied by maps $\psi \in \mathrm{Aut}(M,\Phi)$. We will continue to denote by ∇ the affine connection on M that is trivial with respect to Φ, though let us add a word of caution about this: being a "trivial" connection implies that its curvature vanishes, but its torsion can be nonzero, and *must* indeed be nonzero outside of the exceptional situation where $\nabla^\Phi = \nabla$ and thus $\mathfrak{aut}(M,\Phi) = \mathfrak{V}$. In any case, the triviality of the connection ∇ implies that its parallel transport map

$$P_{(p,q)} := P_\gamma^1 : T_pM \to T_qM$$

is independent of the choice of smooth path γ from $\gamma(0) = p$ to $\gamma(1) = q$. The parallel vector fields $V \in \mathfrak{V}$ are then characterized by the condition $V(q) = P_{(p,q)}(V(p))$ for every $p, q \in M$, so writing the condition $\psi^*V = V$ in the form $T_p\psi(V(p)) = V(\psi(p))$, we find that a diffeomorphism $\psi : M \to M$ belongs to $\mathrm{Aut}(M,\Phi)$ if and only if it satisfies

$$T_p\psi = P_{(p,\psi(p))} \in \mathrm{Hom}(T_pM, T_{\psi(p)}M) \qquad \text{for all} \qquad p \in M. \tag{49.6}$$

In a local chart $x = (x^1, \dots, x^n)$, if we write $x \circ \psi = (\psi^1, \dots, \psi^n)$, (49.6) is equivalent to a system of n^2 first-order nonlinear PDEs of the form

$$\partial_i \psi^j(p) - P_i^{\ j}(p, \psi(p)) = 0, \qquad i, j \in \{1, \dots, n\},$$

where the parallel transport is now represented by a matrix-valued function with entries $P_i^{\ j}$ that depend smoothly on both p and $\psi(p)$. As first-order PDEs go, this one is rather unsubtle: it implies that all the first partial derivatives of ψ at any point are completely determined by the value of ψ at that point. This makes it a highly *overdetermined* PDE: we will see for instance that its solutions are completely determined by their values at one point, giving a very strong uniqueness result, though for the same reason, one should not typically expect interesting solutions to exist.[106] A convenient side benefit is that in order to prove what must be proved about this equation, we will have no need for fancy PDE methods: the theory of *ordinary* differential equations is completely sufficient. The essential analytical properties of the equation are of a purely local nature, thus for much of the discussion, we will consider solutions defined only on an open subset $\mathcal{U} \subset M$, which could be all of M but could also just be a small neighborhood of a point.

We claim that in some sense to be made precise below, the linear equation $\nabla^\Phi X = 0$ is the linearization of (49.6). To say what this means, let us first rewrite the nonlinear equation by defining for any open subset $\mathcal{U} \subset M$ a map

$$\mathbf{F} : C^\infty(\mathcal{U}, M) \to \Gamma(\operatorname{End}(T\mathcal{U})), \qquad \mathbf{F}(\psi)(p) := P_{(\psi(p),p)} \circ T_p\psi - \mathbb{1} \in \operatorname{End}(T_pM), \tag{49.7}$$

so that if we take $\mathcal{U} := M$ and consider a diffeomorphism $\psi \in \operatorname{Diff}(M) \subset C^\infty(M, M)$, (49.6) becomes

$$\mathbf{F}(\psi) = 0 \in \Gamma(\operatorname{End}(TM)).$$

Heuristically, we can regard $C^\infty(\mathcal{U}, M)$ as an infinite-dimensional manifold on which $\mathbf{F}$ is a smooth vector-valued function, and in the special case $\mathcal{U} = M$, there is a distinguished open subset $\operatorname{Diff}(M) \subset C^\infty(M, M)$ on which the zero-set of $\mathbf{F}$ is precisely $\operatorname{Aut}(M, \Phi)$. (It will not be necessary for our present purposes to make these notions precise, and we will not attempt to do so—infinite-dimensional differential geometry involves a multitude of tricky technical details that are irrelevant to this heuristic discussion.) From this perspective, we think of $\mathfrak{X}(\mathcal{U})$ as the tangent space $T_{\mathrm{Id}} C^\infty(\mathcal{U}, M)$ to this infinite-dimensional manifold at the inclusion map $\mathrm{Id} \in C^\infty(\mathcal{U}, M)$, and linearizing (49.6) then means computing the directional derivative of $\mathbf{F}$ at Id in the direction of an arbitrary $X \in \mathfrak{X}(\mathcal{U})$. Suppose indeed that $\{\psi_s : \mathcal{U} \to M\}_{s \in (-\epsilon,\epsilon)}$ is a smooth 1-parameter family of maps with $\psi_0 = \mathrm{Id}$ and $\partial_s \psi_s|_{s=0} = X \in \mathfrak{X}(\mathcal{U})$, and given $p \in \mathcal{U}$ and $Y \in T_pM$, choose a smooth path $\gamma(t) \in \mathcal{U}$ with $\gamma(0) = p$ and $\dot\gamma(0) = Y$. Then

$$\begin{aligned}
\left.\frac{d}{ds}\mathbf{F}(\psi_s)(p)Y\right|_{s=0} &= \partial_s \left[P_{(\psi_s(p),p)} \circ T_p\psi_s(Y) - Y\right]\big|_{s=0} = \partial_s \left[P_{(\psi_s(p),p)}\left(\partial_t \left[\psi_s(\gamma(t))\right]\big|_{t=0}\right)\right]\big|_{s=0} \\
&= \nabla_s \partial_t \left[\psi_s(\gamma(t))\right]\big|_{s=t=0} \\
&= \nabla_t \partial_s \left[\psi_s(\gamma(t))\right]\big|_{s=t=0} + T\left(\partial_s\left[\psi_s(\gamma(t))\right], \partial_t\left[\psi_s(\gamma(t))\right]\right)\big|_{s=t=0} \\
&= \nabla_t \left[X(\gamma(t))\right]\big|_{t=0} + T(X(p), Y) = \nabla_Y X + T(X(p), Y) = \nabla^\Phi_Y X.
\end{aligned}$$

This calculation leads us to define a linear operator $D\mathbf{F}(\mathrm{Id}) : \mathfrak{X}(\mathcal{U}) \to \Gamma(\operatorname{End}(T\mathcal{U}))$, called the **linearization** of $\mathbf{F}$ at $\mathrm{Id} \in C^\infty(\mathcal{U}, M)$, by

$$D\mathbf{F}(\mathrm{Id})X = \left.\frac{d}{ds}\mathbf{F}(\psi_s)\right|_{s=0} = \nabla^\Phi X \in \Gamma(\operatorname{End}(T\mathcal{U})).$$

Unlike ∇, the connection ∇^Φ will not usually be flat, so the space $\mathfrak{aut}(M, \Phi)$ of ∇^Φ-parallel vector fields may be trivial, or it may have any dimension up to n:

LEMMA 49.38. *For any point $p \in M$ and any connected open subset $\mathcal{U} \subset M$ containing p, the space of vector fields $X \in \mathfrak{X}(\mathcal{U})$ that are parallel with respect to ∇^Φ injects into T_pM via the map $X \mapsto X(p)$. In particular,* $\dim \mathfrak{aut}(M, \Phi) \leqslant n$.

[106] The algebraic analogue of an overdetermined system of PDEs is a system of m algebraic equations in n variables where $m > n$. Special examples of such systems may very well have nontrivial solutions, but for a generic choice of equations, it is much more likely for the solution set to be empty.

Proof. If $\nabla^\Phi X = 0$, then $X(p)$ uniquely determines $X(q)$ for every $q \in \mathcal{U}$ via parallel transport with respect to the connection ∇^Φ along an arbitrary path in $\mathcal{U}$ from p to q. □

We will next prove some analytical results about solutions of the nonlinear PDE (49.6) and its linearization $\nabla^\Phi X = 0$. Given a smooth vector bundle $E \to M$, a sequence of sections $s_k \in \Gamma(E)$ and an integer $\ell \geqslant 0$, we will say that the sequence s_k is:

- **uniformly C^ℓ_{loc}-bounded** if for every local trivialization $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^m$, the sequence of functions $f_k : \mathcal{U}_\alpha \to \mathbb{F}^m$ determined by $\Phi_\alpha(s_k(p)) = (p, f_k(p))$ has all derivatives up to order ℓ bounded independently of k on each compact subset of $\mathcal{U}_\alpha$;
- **C^ℓ_{loc}-convergent** to a section $s \in \Gamma(E)$ if for every local trivialization $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^m$, the sequence of functions $f_k : \mathcal{U}_\alpha \to \mathbb{F}^m$ determined by $\Phi_\alpha(s_k(p)) = (p, f_k(p))$ has all derivatives up to order ℓ convergent uniformly on compact subsets of $\mathcal{U}_\alpha$ to the corresponding derivatives of $f : \mathcal{U}_\alpha \to \mathbb{F}^m$, where $\Phi_\alpha(s(p)) = (p, f(p))$.

The C^∞_{loc}-topology on $\Gamma(E)$ is thus characterized by the condition that a sequence converges if and only if it is C^ℓ_{loc}-convergent for every $\ell \in \mathbb{N}$. Since uniform C^1-bounds imply equicontinuity, the Arzelà-Ascoli theorem implies that uniformly $C^{\ell+1}_{\mathrm{loc}}$-bounded sequences always have C^ℓ_{loc}-convergent subsequences, so in particular, any sequence that is uniformly C^ℓ_{loc}-bounded for every $\ell \in \mathbb{N}$ has a C^∞_{loc}-convergent subsequence.

The following lemma is an analogue of Theorem 49.6 on the rigidity of affine transformations for the setting of a manifold with a parallelization; in fact, up to minor details concerning the connectedness of the frame bundle $F^G(TM)$, it gives us a second proof of the rigidity of affine transformations in light of Proposition 49.29. It also implies the statement in Theorem 49.37 that the action of $\mathrm{Aut}(M, \Phi)$ on M is free and proper.

Lemma 49.39. *Fix a point $p \in M$ and a connected open neighborhood $\mathcal{U} \subset M$ of p. For any $q \in M$, there exists at most one map $\psi : \mathcal{U} \to M$ satisfying Equation (49.6) and $\psi(p) = q$. Moreover, if $p_k \in \mathcal{U}$ is a sequence of points converging to p and $\psi_k : \mathcal{U} \to M$ is a sequence of maps satisfying (49.6) such that $q_k := \psi_k(p_k)$ converges to some point $q \in M$, then ψ_k is C^∞_{loc}-convergent on $\mathcal{U}$ to a map $\psi : \mathcal{U} \to M$ satisfying (49.6) and $\psi(p) = q$.*

Proof. For $V \in \mathfrak{V}$, a map $\psi : \mathcal{U} \to M$ satisfying (49.6) satisfies $T_q\psi(V(q)) = V(\psi(q))$ for every $q \in \mathcal{U}$, and it follows that for every flow line γ of V with image in $\mathcal{U}$, $\psi \circ \gamma$ is also a flow line of V. Since the vector fields in $\mathfrak{V}$ span TM at every point, it follows that if $\psi(p) = q$, ψ is determined on some neighborhood of p by the formula

$$\psi\big(\varphi^1_V(p)\big) = \varphi^1_V(q) \tag{49.8}$$

for all $V \in \mathfrak{V}$ in some neighborhood of 0. This proves that if $\psi, \psi' : \mathcal{U} \to M$ are any two maps satisfying (49.6), the set of points at which they match is open. It is clearly also closed, so if it is nonempty, then it is all of $\mathcal{U}$ since $\mathcal{U}$ is connected.

A similar argument using smooth dependence of solutions to ODEs on initial conditions proves the convergence of any sequence $\psi_k : \mathcal{U} \to M$ satisfying (49.6) with $q_k = \psi_k(p_k) \to q$ and $p_k \to p$. Indeed, (49.8) in this case becomes

$$\psi_k\big(\varphi^1_V(p_k)\big) = \varphi^1_V(q_k)$$

for all $V \in \mathfrak{V}$ in some neighborhood of 0, so ψ_k converges uniformly with all derivatives on some compact neighborhood of p to the unique map ψ satisfying (49.8) on that neighborhood. Since all of the ψ_k also satisfy the stronger condition (49.6) and that condition is closed with respect to the C^∞_{loc}-topology, it follows that ψ also satisfies it. Finally, one argues via the connectedness of $\mathcal{U}$ that the C^∞_{loc}-convergence $\psi_k \to \psi$ is valid not just near p but on a nonempty open and closed subset of $\mathcal{U}$, which is therefore all of $\mathcal{U}$. □

For the linearized analogue of the previous lemma, we will consider arbitrary connections on $TM \to M$, since any connection could in principle be ∇^Φ for some choice of trivialization Φ. Recall that the set of all connections on $TM \to M$ is an affine space over the space of smooth bundle maps $\Gamma(\mathrm{Hom}(TM, \mathrm{End}(TM)))$, so for a sequence ∇^k of connections, we will say that ∇^k is:

- **uniformly C^ℓ_{loc}-bounded** if $\nabla^k = \nabla + A_k$ for some fixed connection ∇ and a uniformly C^ℓ_{loc}-bounded sequence $A_k \in \Gamma(\mathrm{Hom}(TM, \mathrm{End}(TM)))$;
- **C^ℓ_{loc}-convergent** to a connection ∇^∞ if $\nabla^k = \nabla^\infty + A_k$ for a sequence of bundle maps $A_k \in \Gamma(\mathrm{Hom}(TM, \mathrm{End}(TM)))$ that is C^ℓ_{loc}-convergent to 0.

We will then say that $\nabla^k \to \nabla^\infty$ in C^∞_{loc} if it is C^ℓ_{loc}-convergent for every $\ell \in \mathbb{N}$.

LEMMA 49.40. *Fix a point $p \in M$ and a connected open neighborhood $\mathcal{U} \subset M$ of p, suppose ∇^k is a sequence of connections on $T\mathcal{U} \to \mathcal{U}$ and $X_k \in \mathfrak{X}(\mathcal{U})$ is a sequence of vector fields satisfying $\nabla^k X_k = 0$.*

(1) If the sequence $X_k(p) \in T_pM$ is bounded and the connections ∇^k are uniformly C^m_{loc}-bounded for every $m \in \mathbb{N}$, then X_k has a C^∞_{loc}-convergent subsequence.

(2) If the sequence $X_k(p) \in T_pM$ converges and the connections ∇^k are C^∞_{loc}-convergent to some connection ∇^∞ on $T\mathcal{U} \to \mathcal{U}$, then X_k is C^∞_{loc}-convergent to a vector field $X \in \mathfrak{X}(\mathcal{U})$ satisfying $\nabla^\infty X = 0$.

PROOF. We claim first that if $X_k(p)$ is bounded and ∇^k is uniformly C^0_{loc}-bounded, then X_k is also uniformly C^0_{loc}-bounded. Indeed, in local coordinates near p, the equation $\nabla^k X_k = 0$ takes the form

$$\partial_i X_k^j + (\Gamma_{(k)})^j_{i\ell} X_k^\ell = 0, \tag{49.9}$$

where we denote the Christoffel symbols of ∇^k by $(\Gamma_{(k)})^j_{i\ell}$ and observe that these are uniformly C^0-bounded on compact subsets by assumption. It follows that for any smooth path $\gamma(t)$ in this coordinate neighborhood, the components $X_k^j(\gamma(t))$ along γ satisfy a linear ODE of the form

$$\frac{d}{dt} X_k^j(\gamma(t)) = (A_k(t))^j{}_i X_k^i(\gamma(t))$$

for a uniformly C^0_{loc}-bounded sequence of matrix-valued functions $A_k(t)$ with entries $(A_k(t))^j{}_i$. A standard argument using e.g. the Grönwall inequality then establishes a bound of the form $|X_k^j(\gamma(t))| \leqslant e^{C|t|} \cdot \max\{|X_k^1(\gamma(0))|, \dots, |X_k^n(\gamma(0))|\}$ for some constant $C > 0$ independent of k, thus a bound on $X_k(p)$ implies a uniform C^0-bound on the components of X_k in some compact neighborhood of p. Note that there was nothing special about the point p in this discussion; the same conclusion would result if p were replaced by any other point in $\mathcal{U}$. Now for $q \in \mathcal{U}$, a bound on $X_k(p)$ implies a bound on $X_k(q)$ by choosing a path $\gamma : [0, 1] \to \mathcal{U}$ from p to q and breaking it up into finitely-many small segments $[t_j, t_{j+1}]$ such that any bound on $X_k(t_j)$ implies a uniform C^0-bound on X_k along $\gamma([t_j, t_{j+1}])$. From this follows a uniform C^0-bound for X_k on some neighborhood of q, and since arbitrary compact subsets of $\mathcal{U}$ are finite unions of neighborhoods of this form, the claim follows.

A similar application of the Grönwall inequality shows that if $X_k(p) \to Y$ and the connections ∇^k are C^0_{loc}-convergent to a connection ∇^∞, then X_k is C^0_{loc}-convergent to a vector field $X \in \mathfrak{X}(\mathcal{U})$ satisfying $\nabla^\infty X = 0$, which is uniquely determined by the condition $X(p) = Y$.

Next, we claim that for any $m \in \mathbb{N}$, if both X_k and ∇^k are uniformly C^{m-1}_{loc}-bounded, then X_k is also uniformly C^m_{loc}-bounded. This follows from the equation $\nabla^k X_k = 0$, which in local coordinates looks like (49.9), with the sequence of Christoffel symbols $(\Gamma_{(k)})^j_{i\ell}$ assumed to satisfy a uniform C^{m-1}-bound on compact subsets. Indeed, $\partial_i X_k^j$ then also satisfies a uniform C^{m-1}-bound, and X_k^j is thus C^m-bounded on compact subsets. If we make the stronger assumption that X_k

and ∇^k are both C^{m-1}_{loc}-convergent, then this argument shows that the first derivatives of X_k are C^{m-1}_{loc}-convergent, hence X_k itself is C^m_{loc}-convergent.

Both statements in the lemma now follow by induction on m. $\square$

With this technical preparation out of the way, we can now begin studying the structure of $\text{Aut}(M, \Phi)$ near the identity map. Let

$$\exp : \mathcal{O} \to M$$

denote the exponential map for the trivial connection ∇. Here $\mathcal{O} \subset TM$ is an open neighborhood of the zero section; we are not assuming the vector fields in $\mathfrak{V}$ are complete, so geodesics may escape to infinity in finite time.

Lemma 49.41. *Suppose $\mathcal{U} \subset M$ is an open subset and $X_k \in \mathfrak{X}(\mathcal{U})$ is a sequence of vector fields that are C^∞_{loc}-convergent to zero and take values in the domain $\mathcal{O} \subset TM$ of* exp*, such that the maps*

$$\varphi_k := \exp \circ X_k : \mathcal{U} \to M$$

satisfy Equation (49.6). Then there exists a sequence of connections ∇^k on $T\mathcal{U}$ that are C^∞_{loc}-convergent to ∇^Φ and satisfy $\nabla^k X_k = 0$.

Proof. The idea is to decompose the nonlinear PDE (49.6) into its linear part plus a remainder term, and then absorb the remainder into the notation as a zeroth-order perturbation of the connection ∇^Φ. The tricky detail is that we need to write down the derivative of $\varphi_k = \exp \circ X_k$ in terms of the covariant derivative of X_k, and that requires differentiating the exponential map $\exp : \mathcal{O} \to M$.

For bookkeeping purposes, let us write elements of TM as pairs (p, Y) where $p \in M$ and $Y \in T_pM$. For each $(p, Y) \in TM$, the connection ∇ determines an isomorphism

$$T_{(p,Y)}(TM) = T_pM \times T_pM,$$

where the first factor corresponds to the horizontal subspace and the second to the vertical subspace, so under this identification, a vector field $Y(t) \in T_{\gamma(t)}M$ along a path $\gamma(t) \in M$ is viewed as a path in TM with tangent vector

$$\dot{Y}(t) = (\dot{\gamma}(t), \nabla_t Y(t)) \in T_{\gamma(t)}M \times T_{\gamma(t)}M = T_{(\gamma(t),Y(t))}(TM).$$

With this identification in place, we associate to each $(p, Y) \in \mathcal{O}$ the linear map

$$G(p, Y) := P_{(\exp_p(Y),p)} \circ T_{(p,Y)}(\exp) : T_pM \times T_pM \to T_pM.$$

For each fixed $p \in M$, $Y \mapsto G(p, Y)$ is a smooth function on an open subset of T_pM with values in the fixed vector space $\text{Hom}(T_pM \times T_pM, T_pM)$, thus it is subject to the methods of first-year analysis: in particular, we can write

$$G(p, Y) = G(p, 0) + D_2G(p, 0)Y + R(p, Y)Y,$$

where $D_2G(p, Y) : T_pM \to \text{Hom}(T_pM \times T_pM, T_pM)$ denotes the derivative of the map $Y \mapsto G(p, Y)$, and the remainder function $R : \mathcal{O} \to \text{Hom}(T_pM, \text{Hom}(T_pM \times T_pM, T_pM))$ is given by

$$R(p, Y) = \int_0^1 [D_2G(p, \tau Y) - D_2G(p, 0)] \, d\tau,$$

so it depends smoothly on $(p, Y) \in \mathcal{O}$ and satisfies $R(p, 0) = 0$ for all $p \in M$. For a vector field $X \in \mathfrak{X}(\mathcal{U})$ taking values in the domain of exp, the result of feeding the map

$$\varphi := \exp \circ X \in C^\infty(\mathcal{U}, M)$$

into the operator $\mathbf{F} : C^\infty(\mathcal{U}, M) \to \Gamma(\mathrm{End}(T\mathcal{U}))$ from (49.7) can now be written as follows: for each $p \in \mathcal{U}$ and $Y \in T_pM$,

$$\begin{aligned} \mathbf{F}(\varphi)(p)Y &= P_{(\varphi(p),p)} \circ T_p(\exp \circ X)(Y) - Y \\ &= P_{(\exp_p(X(p)),p)} \circ T_{(p,X(p))}(\exp) \circ T_pX(Y) - Y \\ &= G(p, X(p))\,(Y, \nabla_Y X) - Y \\ &= \left[G(p,0) + D_2G(p,0)X(p) + R(p, X(p))X(p)\right](Y, \nabla_Y X) - Y. \end{aligned} \tag{49.10}$$

Since $P_{(p,p)}$ is the identity map on T_pM, we have $G(p,0)(Y,Z) = Y + Z$ for any $Y, Z \in T_pM$, so that (49.10) becomes

$$\mathbf{F}(\varphi)(p)Y = \nabla_Y X + \left[D_2G(p,0)X(p)\right](Y, \nabla_Y X) + \left[R(p, X(p))X(p)\right](Y, \nabla_Y X)\,. \tag{49.11}$$

In order to understand the second term in this expression, let us consider what it implies about the linearization $\nabla^\Phi X = D\mathbf{F}(\mathrm{Id})X$, which can be found by setting $\varphi_s(p) := \exp_p(sX(s))$ and differentiating $\mathbf{F}(\varphi_s)$ with respect to the parameter s: computing this via (49.11) gives

$$\begin{aligned} \nabla^\Phi_Y X &= \frac{d}{ds}\mathbf{F}(\varphi_s)(p)Y\Big|_{s=0} \\ &= \frac{d}{ds}\left(\nabla_Y(sX) + \left[D_2G(p,0)sX(p)\right](Y, \nabla_Y(sX)) + \left[R(p, sX(p))sX(p)\right](Y, \nabla_Y(sX))\right)\Big|_{s=0} \\ &= \nabla_Y X + \left[D_2G(p,0)X(p)\right](Y, 0)\,, \end{aligned}$$

so that (49.11) can now be rewritten as

$$\mathbf{F}(\varphi)(p)Y = \nabla^\Phi_Y X + \left[D_2G(p,0)X(p)\right](0, \nabla_Y X) + \left[R(p, X(p))X(p)\right](Y, \nabla_Y X)\,. \tag{49.12}$$

Now if $\varphi_k = \exp \circ X_k$ for a sequence $X_k \in \mathfrak{X}(\mathcal{U})$, we can define a sequence of bundle maps $A_k : T\mathcal{U} \to \mathrm{End}(T\mathcal{U})$ by

$$A_k(Z)Y := \left[D_2G(p,0)Z\right](0, \nabla_Y X_k) + \left[R(p, X_k(p))Z\right](Y, \nabla_Y X_k)\,, \qquad \text{for } p \in \mathcal{U},\ Y, Z \in T_pM,$$

and observe that if X_k is C^∞_{loc}-convergent to zero, then A_k is as well. Equation (49.12) then implies that if φ_k satisfy $\mathbf{F}(\varphi_k) = 0$, the vector fields X_k satisfy

$$(\nabla^\Phi + A_k)X_k = 0,$$

and the desired sequence of connections is thus $\nabla^k := \nabla^\Phi + A_k$. □

Here is the main technical result in the background of Theorem 49.37.

PROPOSITION 49.42. *In the setting of Lemma 49.41, assume the open subset $\mathcal{U} \subset M$ is connected and $\varphi_k \ne \mathrm{Id}$ for all k. Then after passing to a subsequence, there exists a sequence $\tau_k > 0$ with $\tau_k \to 0$ such that the vector fields $\frac{1}{\tau_k}X_k \in \mathfrak{X}(\mathcal{U})$ are C^∞_{loc}-convergent to a nontrivial solution $X_\infty \in \mathfrak{X}(\mathcal{U})$ of the equation $\nabla^\Phi X_\infty = 0$.*

PROOF. Fix a point $p \in \mathcal{U}$ and observe that by Lemma 49.39, $\varphi_k(p) \ne p$ and thus $X_k(p) \ne 0$ for all k. Choose any sequence $\tau_k > 0$ such that the sequence $\frac{1}{\tau_k}X_k(p) \in T_pM$ is bounded and also stays outside some fixed neighborhood of $0 \in T_pM$ for all k; note that this requires $\tau_k \to 0$ since $X_k \to 0$ in C^∞_{loc}. The rescaled vector fields $\frac{1}{\tau_k}X_k$ also satisfy the linear equations $\nabla^k\left(\frac{1}{\tau_k}X_k\right) = \frac{1}{\tau_k}\nabla^k X_k = 0$ for the C^∞_{loc}-convergent sequence of connections $\nabla^k \to \nabla^\Phi$ from Lemma 49.41, so Lemma 49.40 implies that they have a subsequence C^∞_{loc}-convergent to some vector field $X_\infty \in \mathfrak{X}(\mathcal{U})$ satisfying $\nabla^\Phi X_\infty = 0$. We have $X_\infty(p) \ne 0$ since $\frac{1}{\tau_k}X_k(p)$ was bounded away from zero. □

Proof of Theorem 49.37. We need to show that some C^∞_{loc}-neighborhood of the identity map in $\mathrm{Aut}(M,\Phi)$ contains only maps of the form φ^1_X for vector fields $X \in \mathfrak{aut}(M,\Phi)$ that are C^∞_{loc}-close to zero. Assume the contrary, namely that there exists a sequence $\psi_k \in \mathrm{Aut}(M,\Phi)$ converging to the identity that cannot be expressed as $\varphi^1_{Y_k}$ for any sequence $Y_k \in \mathfrak{aut}(M,\Phi)$ converging to zero. For each $p \in M$, we can write $\psi_k(p) = \exp_p(X_k(p))$ for all k sufficiently large, which uniquely defines a sequence of vector fields X_k defined on a nested sequence of open subsets $\mathcal{U}_k \subset M$ whose union is M, such that $X_k \to 0$ in the C^∞_{loc}-topology on M. (Note that this notion of convergence does not require each X_k to be defined globally on M; it suffices that each individual compact subset of M is contained in $\mathcal{U}_k$ for k large enough.) By Lemma 49.38, there exists a smooth subbundle $E \subset TM$ such that

$$E_p = \left\{ Y(p) \in T_pM \;\middle|\; Y \in \mathfrak{X}(M) \text{ with } \nabla^\Phi Y \equiv 0 \right\},$$

and we can then choose a complementary subbundle $E^\perp \subset TM$, so

$$TM = E \oplus E^\perp.$$

This produces a decomposition

$$X_k = Y_k + Z_k, \qquad Y_k \in \Gamma(E|_{\mathcal{U}_k}),\ Z_k \in \Gamma(E^\perp|_{\mathcal{U}_k}),$$

where Y_k and Z_k are both C^∞_{loc}-convergent to zero. Fixing a point $p \in M$, define

$$\widehat{Y}_k \in \mathfrak{aut}(M,\Phi)$$

for each k as the unique ∇^Φ-parallel vector field satisfying $\widehat{Y}_k(p) = Y_k(p)$. We have $\widehat{Y}_k \to 0$ in C^∞_{loc} since $Y_k(p) \to 0$, thus we can define another sequence in $\mathrm{Aut}(M,\Phi)$ converging in C^∞_{loc} to the identity by

$$f_k := \varphi^1_{\widehat{Y}_k} \in \mathrm{Aut}(M,\Phi).$$

By assumption ψ_k and f_k cannot be identical for any k, so by Lemma 49.39, $\psi_k(p) \ne f_k(p)$ and thus $Z_k(p) \ne 0$ for all k. The sequence $f_k^{-1} \circ \psi_k \in \mathrm{Aut}(M,\Phi)$ is now also C^∞_{loc}-convergent to the identity, so there exist unique vector fields V_k, defined on another nested sequence of open subsets exhausting M, such that

$$f_k^{-1} \circ \psi_k = \exp \circ V_k \qquad \text{on } \mathcal{U}_k$$

and $V_k \to 0$ in C^∞_{loc} on M. By Proposition 49.42, we can pass to a subsequence and find a sequence of positive numbers $\tau_k \to 0$ such that $\frac{1}{\tau_k} V_k$ converges in C^∞_{loc} to a nontrivial vector field in $\mathfrak{aut}(M,\Phi)$.

Let us now examine the behavior of $V_k(p) \in T_pM$ as $k \to \infty$ more closely. Consider the smooth function

$$F : \mathfrak{aut}(M,\Phi) \oplus E_p^\perp \overset{\text{open}}{\supset} \mathcal{V} \to T_pM, \qquad \text{such that} \qquad \exp_p\left(F(Y,Z)\right) = \varphi^1_{-Y} \circ \exp_p(Y(p) + Z),$$

where $\mathcal{V} \subset \mathfrak{aut}(M,\Phi) \oplus E_p^\perp$ is a small enough neighborhood of 0 so that $(Y,Z) \mapsto \exp_p(Y(p)+Z)$ sends $\mathcal{V}$ diffeomorphically to a neighborhood of p in M. We have,

$$F(Y,0) = 0 \text{ for all } Y, \qquad \text{and} \qquad D_2F(0,0)Z = Z,$$

and can therefore write

$$F(Y,Z) = Z + Q(Y,Z)Z$$

for some smooth function $Q : \mathcal{V} \to \mathrm{Hom}(E_p^\perp, T_pM)$ that vanishes at $(0,0)$; indeed, computing the integral $\int_0^1 \frac{d}{d\tau} F(Y,\tau Z)\, d\tau$ leads to the formula

$$Q(Y,Z) = D_2F(Y,0) - D_2F(0,0) + \int_0^1 D_2F(Y,\tau Z)\, d\tau.$$

It follows that the sequence of vector fields V_k defined above satisfies

$$V_k(p) = F(\widehat{Y}_k, Z_k(p)) = Z_k(p) + Q(\widehat{Y}_k, Z_k(p))Z_k(p),$$

where $Q(\widehat{Y}_k, Z_k(p)) \to 0$, thus the vector field $\lim_{k\to\infty} \frac{1}{\tau_k} V_k \in \mathfrak{aut}(M, \Phi)$ has a nontrivial value in $E_p^\perp$ at p, and that is a contradiction. $\square$

50. Spin structures

In Lecture 42 we took pains to define the notion of a G-structure on a fiber bundle without assuming that G acts effectively on the standard fiber, but as yet we've seen very few actual examples where non-effective actions arise. One such example flew under the radar in Lecture 46: the adjoint bundle $\mathrm{Ad}(E) \to M$ of a principal G-bundle has fiber $\mathfrak{g}$ with G acting via the adjoint representation, which for instance is trivial if G is abelian, and that is the reason why the connection that $\mathrm{Ad}(E)$ inherits from an arbitrary principal connection on E will sometimes be canonically trivial. We will see some less trivial examples in this lecture, and try also to get a general idea of what the subject known as *gauge theory* is about.

50.1. The case of dimension three. Let's dive right in and define what a spin structure is on an oriented Euclidean vector bundle of rank 3. The reason to start with this case is that the Lie group we will end up calling $\mathrm{Spin}(3)$ is already familiar to us: it can be identified with $\mathrm{SU}(2)$. The key fact to recall about $\mathrm{SU}(2)$ is that it serves as a double cover of $\mathrm{SO}(3)$, i.e. as we saw in §39.2, there is a Lie group homomorphism

$$\Phi : \mathrm{SU}(2) \to \mathrm{SO}(3)$$

that is a covering map of degree 2, defined by taking the adjoint representation $\mathrm{Ad} : \mathrm{SU}(2) \to \mathrm{SO}(\mathfrak{su}(2))$ and choosing an orthonormal (with respect to an Ad-invariant inner product) basis of $\mathfrak{su}(2)$ so as to identify it with $\mathbb{R}^3$.

Definition 50.1. A **spin structure** on an oriented Euclidean vector bundle $E \to M$ of rank 3 consists of an equivalence class of $\mathrm{SO}(3)$-bundle atlases with transition functions $\{g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{SO}(3)\}_{(\alpha,\beta)\in I\times I}$, together with a system of $\mathrm{SU}(2)$-valued transition functions $\{h_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{SU}(2)\}_{(\alpha,\beta)\in I\times I}$ such that

$$g_{\beta\alpha} = \Phi \circ h_{\beta\alpha}$$

for all $\alpha, \beta \in I$.

Another way of saying this is that a spin structure on $E \to M$ is an $\mathrm{SU}(2)$-structure for which the action of $\mathrm{SU}(2)$ on the standard fiber $\mathbb{R}^3$ is via the homomorphism $\Phi : \mathrm{SU}(2) \to \mathrm{SO}(3)$. Every vector bundle with an $\mathrm{SU}(2)$-structure in this sense inherits from this an $\mathrm{SO}(3)$-structure, defined simply by replacing the $\mathrm{SU}(2)$-valued transition functions $h_{\beta\alpha}$ by $\Phi \circ h_{\beta\alpha}$ and letting $\mathrm{SO}(3)$ act on $\mathbb{R}^3$ in ths obvious way; if E is given with an orientation and bundle metric in the first place as in Definition 50.1, then we require the $\mathrm{SO}(3)$-structures determined by this data and by its $\mathrm{SU}(2)$-structure to be the same. Note that since the $\mathrm{SU}(2)$-action on $\mathbb{R}^3$ is not effective, it is important in Definition 50.1 that the transition functions $h_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{SU}(2)$ be required to satisfy the cocycle condition

$$h_{\alpha\alpha} = \mathbb{1}, \qquad \text{and} \qquad h_{\alpha\beta} h_{\beta\gamma} h_{\gamma\alpha} = \mathbb{1}.$$

In most examples we've considered until now, this condition followed automatically from the fact that any transition function was uniquely determined by the two corresponding local trivializations—but that is not true anymore, so the cocycle condition must be required explicitly.

Remark 50.2. The definition of the Čech cohomology groups $\check{H}^k(M;\mathbb{Z}_2)$ with $\mathbb{Z}_2$ coefficients was outlined in Remark 32.6, where we saw that there is a characteristic class $w_1(E) \in \check{H}^1(M;\mathbb{Z}_2)$ of real vector bundles $E \to M$ that vanishes if and only if the bundle is orientable. The **second Stiefel-Whitney class** $w_2(E) \in \check{H}^2(M;\mathbb{Z}_2)$ similarly vanishes if and only if E admits a spin structure. To define it for an SO(3)-bundle, suppose $\{\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{R}^3\}_{\alpha\in I}$ is a bundle atlas with transition functions $\{g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{SO}(3)\}_{(\alpha,\beta)\in I\times I}$, and assume the open covering $\mathfrak{U} := \{\mathcal{U}_\alpha\}_{\alpha\in I}$ of M has been chosen so that all nonempty intersections of up to three of the sets $\mathcal{U}_\alpha$ are connected. (This can always be achieved for instance by choosing a "good" cover—see Exercise 34.19.) For each α, β, we can make an arbitrary choice of function $h_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{SU}(2)$ satisfying $\Phi \circ h_{\beta\alpha} = g_{\beta\alpha}$; when $\alpha = \beta$, let's assume $h_{\alpha\alpha} = \mathbb{1}$. The resulting system of SU(2)-valued functions will not generally satisfy the cocycle condition, and this failure can be measured by a Čech 2-cocycle: for each $\alpha, \beta, \gamma \in I$ such that $\mathcal{U}_\alpha \cap \mathcal{U}_\beta \cap \mathcal{U}_\gamma \neq \varnothing$, we define $c_{\alpha\beta\gamma} \in \mathbb{Z}_2$ via the condition

$$h_{\alpha\beta} h_{\beta\gamma} h_{\gamma\alpha} = (-1)^{c_{\alpha\beta\gamma}},$$

where we are using the fact that the system of transition functions $\{g_{\beta\alpha}\}$ does satisfy the cocycle condition and $\mathcal{U}_\alpha \cap \mathcal{U}_\beta \cap \mathcal{U}_\gamma$ is connected, so that

$$\Phi(h_{\alpha\beta} h_{\beta\gamma} h_{\gamma\alpha}) = g_{\alpha\beta} g_{\beta\gamma} g_{\gamma\alpha} = \mathbb{1},$$

implying that $h_{\alpha\beta} h_{\beta\gamma} h_{\gamma\alpha}$ is a constant equal to ± 1. The next exercise shows that the function $(\alpha,\beta,\gamma) \mapsto c_{\alpha\beta\gamma} \in \mathbb{Z}_2$ is a cocycle and thus represents an element $w_2(E) \in \check{H}^2(M;\mathbb{Z}_2)$, which turns out to be independent of the choices, and vanishes if and only if it is possible to choose the functions $h_{\beta\alpha}$ so that they define a spin structure.

Exercise 50.3. For the Čech cochain $f \in \check{C}^2(\mathfrak{U};\mathbb{Z}_2)$ defined by $f(\alpha,\beta,\gamma) := c_{\alpha\beta\gamma} \in \mathbb{Z}_2$ in Remark 50.2, prove:

(a) For all $\alpha,\beta,\gamma,\delta \in I$, $c_{\beta\gamma\delta} - c_{\alpha\gamma\delta} + c_{\alpha\beta\delta} - c_{\alpha\beta\gamma} = 0$. In other words, $\delta f = 0 \in \check{C}^3(\mathfrak{U};\mathbb{Z}_2)$.

(b) If $f' \in \check{C}^2(\mathfrak{U};\mathbb{Z}_2)$ is a different cochain $f'(\alpha,\beta,\gamma) = c'_{\alpha\beta\gamma}$ defined by making different choices of SU(2)-valued functions $h'_{\beta\alpha}$ with $\Phi \circ h'_{\beta\alpha} = g_{\beta\alpha}$ and $h'_{\alpha\alpha} = \mathbb{1}$, then there exists a function $F(\alpha,\beta) \in \mathbb{Z}_2$ defined whenever $\mathcal{U}_\alpha \cap \mathcal{U}_\beta \neq \varnothing$ such that $F(\beta,\gamma) - F(\alpha,\gamma) + F(\alpha,\beta) = c'_{\alpha\beta\gamma} - c_{\alpha\beta\gamma}$ whenever $\mathcal{U}_\alpha \cap \mathcal{U}_\beta \cap \mathcal{U}_\gamma \neq \varnothing$. In other words, $F \in \check{C}^1(\mathfrak{U};\mathbb{Z}_2)$ satisfies $\delta F = f' - f \in \check{C}^2(\mathfrak{U};\mathbb{Z}_2)$.

(c) If there exists a function $F(\alpha,\beta) \in \mathbb{Z}_2$ defined whenever $\mathcal{U}_\alpha \cap \mathcal{U}_\beta \neq \varnothing$ such that $c_{\alpha\beta\gamma} = F(\beta\gamma) - F(\alpha\gamma) + F(\alpha\beta)$ holds whenever $\mathcal{U}_\alpha \cap \mathcal{U}_\beta \cap \mathcal{U}_\gamma \neq \varnothing$, then the choice of the functions $h_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{SU}(2)$ can be modified so that they satisfy the cocycle condition.

While Definition 50.1 is quite natural, one does not find it stated very often in the literature. The reason is that there is an equivalent way to formulate the same notion in terms of a principal bundle, which does not require talking about local trivializations, transition functions or the cocycle condition. In the following, we denote the bundle of positively-oriented orthonormal frames for fibers of E by

$$F^{\mathrm{SO}}(E) := F^{\mathrm{SO}(3)}E.$$

Definition 50.4. A **spin structure** on an oriented Euclidean vector bundle $E \to M$ of rank 3 consists of a principal SU(2)-bundle $P \to M$ together with a smooth map

$$\Psi : P \to F^{\mathrm{SO}}(E)$$

that sends each fiber P_p for $p \in M$ to the corresponding fiber $F^{\mathrm{SO}}(E_p)$ and is equivariant in the sense that

$$\Psi(\phi\mathbf{A}) = \Psi(\phi)\Phi(\mathbf{A}) \qquad \text{for all} \qquad \phi \in P,\ \mathbf{A} \in \mathrm{SU}(2).$$

Since SU(2) and SO(3) act freely and transitively on the fibers of their respective principal bundles, the equivariance condition in Definition 50.4 implies that the map $\Psi : P \to F^{\mathrm{SO}}(E)$ defines for each $p \in M$ a two-fold covering map $P_p \to F^{\mathrm{SO}}(E_p)$, and the map $P \to F^{\mathrm{SO}}(E)$ itself is therefore also a covering map of degree 2.

Why are Definitions 50.1 and 50.4 equivalent? To go from the second definition to the first, one can think of an SO(3)-bundle atlas for $E \to M$ as a family of local sections $\{s_\alpha \in \Gamma(F^{\mathrm{SO}}(E)|_{\mathcal{U}_\alpha})\}_{\alpha \in I}$ of its SO(3)-frame bundle, and observe that after possibly shrinking the domains $\mathcal{U}_\alpha$, all of these sections can be lifted to local sections $t_\alpha \in \Gamma(P|_{\mathcal{U}_\alpha})$ of the principal SU(2)-bundle with $\Phi \circ t_\alpha = s_\alpha$. These systems of local sections uniquely determine systems of transition functions $g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{SO}(3)$ and $h_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{SU}(2)$ such that

$$s_\alpha = s_\beta g_{\beta\alpha} \qquad \text{and} \qquad t_\alpha = t_\beta h_{\beta\alpha},$$

which automatically satisfy the cocycle condition, and the equivariance of $\Psi : P \to F^{\mathrm{SO}}(E)$ then implies $\Phi \circ h_{\beta\alpha} = g_{\beta\alpha}$.

To go the other direction: recall that every fiber bundle is uniquely determined by its system of transition functions and the way that the structure group acts on the fiber, so in particular, $F^{\mathrm{SO}}(E)$ is isomorphic to the unique SO(3)-bundle that can be constructed via Theorem 42.21 out of the transition functions $\{g_{\beta\alpha}\}$ with standard fiber SO(3) acted upon via left translation. One can use the lifts $\{h_{\beta\alpha}\}$ to construct an SU(2)-bundle $P \to M$ in the same manner, and then define the map $\Psi : \mathrm{SU}(2) \to \mathrm{SO}(3)$ so that it takes the form $\mathcal{U}_\alpha \times \mathrm{SU}(2) \to \mathcal{U}_\alpha \times \mathrm{SO}(3) : (p, \mathbf{A}) \mapsto (p, \Phi(\mathbf{A}))$ when expressed in corresponding local trivializations of both bundles. We leave it as an exercise to check that the map constructed in this way is well defined and satisfies the equivariance condition.

50.2. Motivation from quantum mechanics. Before proceeding to generalize spin structures beyond dimension three, I'd like to address the question of why anyone might ever think such a notion is useful. This requires a short digression on quantum mechanics.

Every physical system comes with certain natural symmetries, and it is always important to make sure that the laws describing the behavior of that system are invariant under those symmetries. In classical mechanics, which takes place in $\mathbb{R}^3$ with the Euclidean metric g_{E}, the symmetry group is usually $\mathrm{Isom}_+(\mathbb{R}^3, g_{\mathrm{E}}) \subset \mathrm{Isom}(\mathbb{R}^3, g_{\mathrm{E}})$, which consists of all orientation-preserving isometries of $(\mathbb{R}^3, g_{\mathrm{E}})$, and is generated by the rotation group SO(3) and the translations $\mathbf{x} \mapsto \mathbf{x} + \mathbf{v}$ for $\mathbf{v} \in \mathbb{R}^3$. For simplicity, we will ignore the translations in this discussion and just talk about rotations. The SO(3)-invariance of Newton's laws of mechanics means for instance that if a path $\mathbb{R} \to \mathbb{R}^{3N} : t \mapsto (\mathbf{x}_1(t), \dots, \mathbf{x}_N(t))$ satisfies the system of second-order differential equations describing the possible motion of N objects in space exerting gravitational forces on each other, then for every $\mathbf{R} \in \mathrm{SO}(3)$, $t \mapsto (\mathbf{R}\mathbf{x}_1(t), \dots, \mathbf{R}\mathbf{x}_N(t))$ is also a solution to that system of equations. This must be true, because the second path can also be interpreted as the same solution but observed in a different reference frame that has been rotated relative to the first one. If preferred, one can simultaneously consider the evolution of the momenta $\mathbf{p}_1(t), \dots, \mathbf{p}_N(t) \in \mathbb{R}^3$ of the N particles (which determine their velocities), so that the motion of the path

$$\mathbb{R} \to \mathbb{R}^{6N} : t \mapsto (\mathbf{x}_1(t), \dots, \mathbf{x}_N(t), \mathbf{p}_1(t), \dots, \mathbf{p}_N(t))$$

is determined by a first-order differential equation. The "physical state" of the system at a given time is then described as a point in the so-called **phase space** $\mathbb{R}^{6N}$, and if that state is known at some reference time t_0, the rest of the path in phase space is determined by it. The rotation group acts on phase space in a completely straightforward way, as the product of $2N$ copies of the canonical linear representation of SO(3) on $\mathbb{R}^3$.

Nonrelativistic quantum mechanics also lives in $(\mathbb{R}^3, g_{\mathrm{E}})$ and thus needs to be invariant under rotations, but there is now a major difference in the meaning of the term "physical state", which causes a somewhat subtle difference in the way that symmetries can be represented. As any physicist will tell you, a physical state in quantum mechanics is determined by a vector of norm 1 in a complex Hilbert space $\mathcal{H}$. For the following discussion, I will occasionally pretend that $\mathcal{H}$ is finite dimensional so that some basic principles of linear algebra can be applied—in most situations $\mathcal{H}$ would be infinite dimensional, and this introduces technical complications that have motivated the development of entire subfields within functional analysis, but these technical complications are tangential to our discussion, so we will ignore them. The possible outcomes of measurements of a quantum mechanical system are determined by the absolute values of inner products $|\langle v, w\rangle|$ of vectors $v, w \in \mathcal{H}$ with norm 1. In more detail, if the system is in the state determined by a normalized vector $v \in \mathcal{H}$ at time t_0, and we wish to measure some physical quantity of the system at that time, say its energy, then it may or may not be possible to predict a precise answer based on the state v. In particular, quantum mechanics associates to every observable quantity such as energy a Hermitian operator on $\mathcal{H}$, and the state $v \in \mathcal{H}$ will have a definite energy if and only if it is an eigenvector of that operator, whose eigenvalue is then the value of the energy. If v is not an eigenvector, but we are given a normalized eigenvector w with eigenvalue $E \in \mathbb{R}$, then the probability of getting E as the answer when we measure the energy of the system in state v is defined to be

$$P(E) := |\langle w, v\rangle|^2 \in [0, 1].$$

Everything that can ever be predicted about the results of measurements in quantum mechanics is therefore determined in this way by absolute values of inner products of $\mathcal{H}$.

Now observe: the normalized vectors $v, w \in \mathcal{H}$ can each be multiplied by $e^{i\theta}$ for any $\theta \in \mathbb{R}$ without changing the probability $P(E)$ defined above. This means that as far as physical measurements are concerned, a state of the system does not actually correspond to a normalized vector in $\mathcal{H}$, but rather to an element of its **projectivization**

$$\mathbb{P}(\mathcal{H}) := (\mathcal{H}\backslash\{0\})/\mathbb{C}^* = \{v \in \mathcal{H} \mid \|v\| = 1\} \Big/ S^1,$$

where $S^1 \subset \mathbb{C}^*$ is the unit circle and both groups act on $\mathcal{H}$ by scalar multiplication. The symmetry question now becomes: what are the possible ways for $\mathrm{SO}(3)$ to act on $\mathbb{P}(\mathcal{H})$ so that products $|\langle v, w\rangle|$ are invariant?

One obvious answer is that $\mathrm{SO}(3)$ could act directly on $\mathcal{H}$ via a unitary representation $\mathrm{SO}(3) \to \mathrm{U}(\mathcal{H})$, but other scenarios are also possible. For instance, if we are instead given a unitary representation of $\mathrm{SU}(2)$,

$$\rho : \mathrm{SU}(2) \to \mathrm{U}(\mathcal{H}),$$

then we can use the double cover $\Phi : \mathrm{SU}(2) \to \mathrm{SO}(3)$ to define an $\mathrm{SO}(3)$-action on $\mathbb{P}(\mathcal{H})$ by

$$\mathbf{A}[v] := [\pm\rho(\widetilde{\mathbf{A}})v], \qquad \text{where} \qquad \widetilde{\mathbf{A}} \in \Phi^{-1}(\mathbf{A}) \subset \mathrm{SU}(2).$$

Each $\mathbf{A} \in \mathrm{SO}(3)$ lifts to two possible choices of $\widetilde{\mathbf{A}} \in \mathrm{SU}(2)$, hence the $\pm$ sign, but since we are working in the projectivization, it does not matter which sign we pick. An $\mathrm{SO}(3)$ action on $\mathbb{P}(\mathcal{H})$ in this sense is called a **projective representation** of $\mathrm{SO}(3)$ on $\mathcal{H}$. In general, every unitary representation of $\mathrm{SO}(3)$ naturally gives rise to a non-faithful unitary representation of $\mathrm{SU}(2)$, simply by composing the map $\mathrm{SO}(3) \to \mathrm{U}(\mathcal{H})$ with the double cover $\mathrm{SU}(2) \to \mathrm{SO}(3)$, but there also exist faithful representations of $\mathrm{SU}(2)$, such as its canonical unitary representation on $\mathbb{C}^2$, which do not correspond to any honest representation of $\mathrm{SO}(3)$, but do give rise to projective representations. The basic principles of quantum mechanics therefore dictate that in order to understand the symmetries of a physical system, one might sometimes need to consider representations of $\mathrm{SU}(2)$ instead of $\mathrm{SO}(3)$. A fundamental result known as *Wigner's theorem* guarantees moreover that this is the most general thing that can happen: every admissible $\mathrm{SO}(3)$-action on $\mathbb{P}(\mathcal{H})$ can be defined via a unitary representation of $\mathrm{SU}(2)$ on $\mathcal{H}$.

The result can also be expressed in terms of the Lie algebra $\mathfrak{so}(3)$. Recall that the derivative of the covering map $\Phi : \mathrm{SU}(2) \to \mathrm{SO}(3)$ at $\mathbb{1}$ is a Lie algebra isomorphism $\Phi_* : \mathfrak{su}(2) \to \mathfrak{so}(3)$. Any unitary representation $\rho : \mathrm{SU}(2) \to \mathrm{U}(\mathcal{H})$ similarly induces a Lie algebra representation $\rho_* : \mathfrak{su}(2) \to \mathfrak{u}(\mathcal{H}) \subset \mathrm{End}(\mathcal{H})$, and in light of the isomorphism Φ_*, this can equally well be viewed as a representation of $\mathfrak{so}(3)$, acting on $\mathcal{H}$ by anti-Hermitian transformations. The group $\mathrm{SU}(2)$ is diffeomorphic to S^3 and thus simply connected, so by Theorem 39.23, every Lie algebra representation of $\mathfrak{su}(2)$ arises from a group representation of $\mathrm{SU}(2)$, and it follows that every Lie algebra representation of $\mathfrak{su}(2) \cong \mathfrak{so}(3)$ by anti-Hermitian transformations on $\mathcal{H}$ could conceivably occur in a quantum mechanical system. On the other hand, $\mathrm{SO}(3) \cong \mathrm{SU}(2)/\mathbb{Z}_2$ is diffeomorphic to $\mathbb{RP}^3$, which is not simply connected, so a representation of its Lie algebra $\mathfrak{so}(3)$ does not always lift to a representation of $\mathrm{SO}(3)$, but does always lift to a *projective* representation, obtained by identifying $\mathfrak{so}(3)$ with $\mathfrak{su}(2)$ and then lifting the $\mathfrak{su}(2)$-representation to $\mathrm{SU}(2)$.

Experiments with elementary particles revealed in the early 20th century that, in fact, one *must* allow certain Lie algebra representations of $\mathfrak{so}(3)$ that do not lift to representations of $\mathrm{SO}(3)$. This is where the word "spin" comes into the story. Spin is a form of angular momentum that elementary particles have, which can be measured in experiments, but was found to have some properties completely different from anything in the world of classical mechanics: for example, for each type of particle, the spin about a given axis can take only finitely many possible values. Perhaps I should back up and clarify more precisely what "angular momentum" means.

There is a basic principle called *Noether's theorem*, valid in both classical and quantum mechanics, which states that every 1-parameter family of symmetries of a physical system gives rise to an observable quantity that is conserved as the system evolves. The laws of conservation of energy, momentum and angular momentum are all examples of this principle, resulting from the invariance of physical laws under time translations, spatial translations and rotations respectively. As mentioned above, each observable quantity in quantum mechanics is represented by a Hermitian operator on the Hilbert space $\mathcal{H}$, and the operator in question can be derived from Noether's theorem as an infinitesimal generator of the corresponding symmetry. More precisely, if G is a Lie group acting on $\mathcal{H}$ via a unitary representation $\rho : G \to \mathrm{U}(\mathcal{H})$, then since the Lie algebra of $\mathrm{U}(\mathcal{H})$ consists of anti-Hermitian linear transformations $\mathcal{H} \to \mathcal{H}$, we can associate to every 1-parameter group of symmetries $\exp(tX) \in G$ generated by a Lie algebra element $X \in \mathfrak{g}$ a unique Hermitian linear transformation $\mathbf{A} : \mathcal{H} \to \mathcal{H}$,[107] such that

$$\rho(\exp(tX)) = e^{it\mathbf{A}},$$

or in other words,

$$\rho_*(X) = i\mathbf{A}.$$

Up to a factor of the physical constant $\hbar$, which can be set equal to 1 in the right units of measurement, the Hermitian operator $\mathbf{A}$ on $\mathcal{H}$ is defined to represent the observable quantity corresponding to the symmetries generated by $X \in \mathfrak{g}$.

By definition, the angular momentum of a physical system about a particular axis in $\mathbb{R}^3$ is the conserved observable quantity that corresponds via Noether's theorem to the rotational symmetry about that axis. Concretely, this means that if $\rho_* : \mathfrak{so}(3) \cong \mathfrak{su}(2) \to \mathfrak{u}(\mathcal{H})$ is the Lie algebra representation by which the generators of rotations act on physical states, and we write the rotation by angle θ about the x^j-axis for $j = 1, 2, 3$ as

$$e^{\theta \mathbf{R}_j} \in \mathrm{SO}(3), \qquad \mathbf{R}_j \in \mathfrak{so}(3),$$

[107] This is where the discussion becomes much more complicated if $\mathcal{H}$ is infinite dimensional, but even in that case, one can use a functional-analytical result called *Stone's theorem* to derive $\mathbf{A}$ from $\rho(\exp(tX))$ as an unbounded self-adjoint operator on $\mathcal{H}$; see e.g. [RS80].

then the angular momentum about the x^j-axis is represented by the unique Hermitian operator $\mathbf{L}_j$ on $\mathcal{H}$ satisfying

$$\rho_*(\mathbf{R}_j) = i\mathbf{L}_j.$$

One of the surprising insights revealed by experiments in the early days of quantum mechanics was that every electron has, aside from its so-called *orbital* angular momentum that depends on its motion through $\mathbb{R}^3$ and corresponds directly to the classical notion of angular momentum, an additional *intrinsic* angular momentum that is independent of its motion in space and can only take two values when measured about any given axis. This intrinsic angular momentum is what is called the *spin* of a particle. If one ignores the orbital angular momentum, e.g. by imagining an electron that is confined to a fixed position in space, then the existence of spin indicates that $\mathfrak{so}(3)$ must be acting on $\mathcal{H}$ via a representation in which each of the standard generators of rotations correspond to anti-Hermitian transformations on $\mathcal{H}$ that have only two eigenvalues. This can only be true if $\mathcal{H}$ is $\mathbb{C}^2$, with $\mathfrak{so}(3) \cong \mathfrak{su}(2)$ acting on it by something equivalent to the canonical representation of $\mathfrak{su}(2)$ on $\mathbb{C}^2$, in which the rotation generators $\mathbf{R}_j \in \mathfrak{so}(3)$ correspond to $-\frac{i}{2}\boldsymbol{\sigma}_j \in \mathfrak{su}(2)$ for the Pauli matrices $\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2, \boldsymbol{\sigma}_3$ mentioned in §39.2. This is the simplest example of a representation of the Lie algebra $\mathfrak{so}(3) \cong \mathfrak{su}(2)$ that does not lift to a group representation of SO(3). It does of course lift to a unitary representation of SU(2), given by the inclusion

$$\iota : \mathrm{SU}(2) \hookrightarrow \mathrm{U}(2).$$

Let us briefly describe how to relax the fictional assumption that our electron is fixed in place. In the usual presentation of quantum mechanics for a single particle in $\mathbb{R}^3$, ignoring spin, the Hilbert space $\mathcal{H}$ is taken to be the space $L^2(\mathbb{R}^3, \mathbb{C})$ of square-integrable complex-valued functions on $\mathbb{R}^3$, and the so-called *wave function* $\psi \in L^2(\mathbb{R}^3, \mathbb{C})$ of a particle is then interpreted as a probability distribution, where the probability of finding the particle to be located in a particular region $\mathcal{U} \subset \mathbb{R}^3$ is defined as

$$P(\mathcal{U}) := \int_{\mathcal{U}} |\psi(\mathbf{x})|^2 \, dx^1 \wedge dx^2 \wedge dx^3.$$

Rotations $\mathbf{R} \in \mathrm{SO}(3)$ act on $L^2(\mathbb{R}^3, \mathbb{C})$ in a fairly obvious way, namely

$$(\mathbf{R} \cdot \psi)(\mathbf{x}) := \psi(\mathbf{R}^{-1}\mathbf{x}),$$

thus defining a unitary (not just projective) representation $\rho : \mathrm{SO}(3) \to \mathrm{U}(\mathcal{H})$, which can also be regarded as a representation $\mathrm{SU}(2) \to \mathrm{U}(\mathcal{H})$ after composing it with the double cover $\Phi : \mathrm{SU}(2) \to \mathrm{SO}(3)$. The Hermitian operators $\mathbf{L}_j$ on $L^2(\mathbb{R}^3, \mathbb{C})$ defined via $\rho_*(\mathbf{R}_j) = i\mathbf{L}_j$ represent what is called the *orbital* angular momentum of the particle. In order to incorporate spin into this picture and thus describe the electron, one can generalize by replacing $L^2(\mathbb{R}^3, \mathbb{C})$ with

$$\mathcal{H} := L^2(\mathbb{R}^3, \mathbb{C}^2) = L^2(\mathbb{R}^3, \mathbb{C}) \otimes \mathbb{C}^2,$$

with the SU(2)-action on $\mathcal{H}$ defined as the tensor product $\rho \otimes \iota$ of the two representations described above, so in terms of vector-valued functions $\psi : \mathbb{R}^3 \to \mathbb{C}^2$,

$$(\mathbf{R} \cdot \psi)(\mathbf{x}) := \mathbf{R}\psi(\Phi(\mathbf{R}^{-1})\mathbf{x}), \qquad \mathbf{R} \in \mathrm{SU}(2).$$

Differentiating the representation $\rho \otimes \iota$ at $\mathbb{1}$ yields a Lie algebra representation $(\rho \otimes \iota)_* : \mathfrak{so}(3) \cong \mathfrak{su}(2) \to \mathfrak{u}(L^2(\mathbb{R}^3, \mathbb{C}) \otimes \mathbb{C}^2))$ that is a sum of two terms,

$$(\rho \otimes \iota)_*(\mathbf{A}) = \rho_*(\mathbf{A}) \otimes \mathbb{1} + \mathbb{1} \otimes \iota_*(\mathbf{A}), \qquad \mathbf{A} \in \mathfrak{su}(2) \cong \mathfrak{so}(3),$$

and the two corresponding Hermitian operators on $L^2(\mathbb{R}^3, \mathbb{C}) \otimes \mathbb{C}^2$ are interpreted as representing the orbital angular momentum and the spin respectively. The existence of spin thus acquires a precise mathematical explanation: it results from the fact that wave functions in quantum mechanics can be vector valued, and the range of possible values of the spin is determined by the choice of unitary representation with which SU(2) acts on these vector values.

The transformation of the space $L^2(\mathbb{R}^3, \mathbb{C}^2)$ of electron wave functions under rotations has the following peculiar property. If $\psi_0 \in L^2(\mathbb{R}^3, \mathbb{C}^2)$ represents the state of an electron in a particular reference frame, and we define a smooth family of wave functions $\{\psi_\theta \in L^2(\mathbb{R}^3, \mathbb{C}^2)\}_{\theta \in \mathbb{R}}$ by requiring ψ_θ to be the same state but viewed from a reference frame that is rotated by an angle θ about a fixed axis in $\mathbb{R}^3$, then

$$\psi_{2\pi} = -\psi_0.$$

In other words, after the observer has made one full rotation, the wave function appears to have changed its sign. This fact has caused intense confusion for generations of physics undergraduates,[108] but it's really just another symptom of the fact that SO(3) does not act directly on the space of wave functions, but only on its projectivization—the actual group action in the picture is an action of SU(2), and under the double cover $\mathrm{SU}(2) \to \mathrm{SO}(3)$, a single loop of rotations in SO(3) lifts to a path in SU(2) from $\mathbb{1}$ to $-\mathbb{1}$, not a closed loop. This causes no theoretical problem in quantum mechanics since the wave function itself is not something that can be measured, only the absolute values of its inner products with other wave functions can be, and these do not see the difference between ψ_0 and $-\psi_0$.

Since $\mathbb{C}^2$-valued wave functions were introduced as a way to represent particles with spin, physicists refer to the vectors in $\mathbb{C}^2$ in this context as **spinors**. This word is also used more generally to refer to representations of the special orthogonal Lie algebra that do not lift to representations of the special orthogonal group.

50.3. The Dirac equation. The motivation for what comes next warrants a second digression on quantum mechanics.

The quantum theory of Schrödinger and Heisenberg was hugely successful at describing ordinary phenomena at the atomic scale, but it was always known to be incomplete, because it was not consistent with Einstein's theory of special relativity. Unlike nonrelativistic mechanics, which lives on $\mathbb{R}^3$ with the Euclidean metric, special relativity lives in $\mathbb{R}^4$ with the **Minkowski metric**

$$g_M = \eta_{\mu\nu}\, dx^\mu\, dx^\nu := (dx^0)^2 - (dx^1)^2 - (dx^2)^2 - (dx^3)^3,$$

which has signature $(1,3)$. In the following we will use the physicists' notational convention in which the coordinates on Minkowski space are labelled with a Greek index ranging from 0 to 3, while Latin indices are reserved for the spatial coordinates, thus x^j for $j = 1,2,3$ are the usual coordinates of 3-dimensional space, $t := x^0$ represents time, and all four can be denoted by x^μ for $\mu = 0, \dots, 3$. The analogue of the rotation group SO(3) for Minkowski space is the **Lorentz group** $\mathrm{SO}(1,3) := \{\mathbf{A} \in \mathrm{O}(1,3) \mid \det(\mathbf{A}) = 1\}$, which contains a copy of SO(3) as the subgroup

$$\mathrm{SO}(3) \cong \left\{ \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{R} \end{pmatrix} \in \mathrm{GL}(4,\mathbb{R}) \;\middle|\; \mathbf{R} \in \mathrm{SO}(3) \right\} \subset \mathrm{SO}(1,3)$$

acting on $\{t\} \times \mathbb{R}^3$ for each constant t by rotations, but also includes transformations known as *Lorentz boosts* that mix the time and spatial coordinates.

In the nonrelativistic quantum mechanics of a single particle, the state of the system is described by a function $\psi(t, \mathbf{x})$ of $(t, \mathbf{x}) \in \mathbb{R} \times \mathbb{R}^3$, such that $|\psi(t,\cdot)|^2 : \mathbb{R}^3 \to [0,\infty)$ is interpreted as a probability distribution for the location of the particle at time t. The evolution of the wave function over time is governed by the Schrödinger equation, a PDE that is first-order in time, so that if $\psi(t_0, \cdot)$ for a given t_0 is known, then it determines ψ for all t. The Schrödinger equation is not invariant under the action of SO(1,3), thus the first challenge in developing a relativistic

[108] I know this from personal experience.

formulation of quantum mechanics was to find an SO(1,3)-invariant replacement for it. One PDE that seems natural to consider in this context is the *Klein-Gordon equation*[109]

$$\left(\partial_t^2 - \sum_{j=1}^3 \partial_j^2 + m^2\right)\psi = \left(\partial^\mu \partial_\mu + m^2\right)\psi = 0,$$

where $m \geqslant 0$ is a constant representing the mass of the particle, and we abbreviate

$$\partial^\mu := \eta^{\mu\nu}\partial_\nu.$$

This equation seems natural for a few reasons: to start with, one of the fundamental equations of classical relativity is a relation between the mass m, kinetic energy E and momentum $\mathbf{p}$ of a freely moving particle that takes the form

$$E^2 - |\mathbf{p}|^2 = m^2,$$

and if one replaces the observables E and $\mathbf{p}$ in this equation with the Hermitian operators on $L^2(\mathbb{R}^3)$ that represent them in standard nonrelativistic quantum mechanics, the Klein-Gordon equation is what comes out. Another reason is the experimental observation that freely moving particles of mass $m \geqslant 0$ can be described quantum-mechanically in terms of wave-like functions of the form

$$\psi(t, \mathbf{x}) = e^{i(\omega t - \langle \mathbf{k}, \mathbf{x}\rangle)},$$

where the *frequency* $\omega \in \mathbb{R}$ and *wave vector* $\mathbf{k} \in \mathbb{R}^3$ are related to each other by $\omega^2 - |\mathbf{k}|^2 = m^2$. The Klein-Gordon equation is the natural SO(1,3)-invariant wave equation that has these particular traveling waves as solutions.

The problem with the Klein-Gordon equation is that it is second-order in time, thus knowledge of the solution at a given time t_0 does not determine the rest of the solution, and for this reason, this equation was not considered suitable for a formulation of relativistic quantum mechanics. This definciency motivates the question: can we find an SO(1,3)-invariant PDE that is *first-order* in time and admits the same wave-like solutions as the Klein-Gordon equation?

The second-order differential operator $\partial^\mu\partial_\mu$ is essentially a variation on the standard Laplace operator—if we were using the Euclidean metric on $\mathbb{R}^4$ instead of the Minkowski metric, it would be (up to a sign) precisely the Laplace operator. Dirac's original idea for relativistic quantum mechanics was to look for a first-order differential operator that would be a square root of this generalized Laplacian. He proposed as an ansatz to write such an operator in the form

$$\not{\partial} := \gamma^\mu \partial_\mu,$$

and to choose the symbols $\gamma^0, \gamma^1, \gamma^2, \gamma^3$ to have algebraic properties such that $\not{\partial}^2 = \partial^\mu\partial_\mu$, so that solutions to the equation

$$(i\not{\partial} - m)\psi = 0 \tag{50.1}$$

would then automatically also satisfy the Klein-Gordon equation,

$$-(i\not{\partial} + m)(i\not{\partial} - m)\psi = (\not{\partial}^2 + m^2)\psi = (\partial^\mu\partial_\mu + m^2)\psi = 0.$$

Using the commutativity of the operators ∂_μ and ∂_ν, we find

$$\not{\partial}^2 = \gamma^\mu\gamma^\nu\partial_\mu\partial_\nu = \frac{1}{2}(\gamma^\mu\gamma^\nu + \gamma^\nu\gamma^\mu)\partial_\mu\partial_\nu,$$

and Dirac's idea works if and only if this operator matches $\eta^{\mu\nu}\partial_\mu\partial_\nu = \partial^\mu\partial_\mu$, which is true if

$$\gamma^\mu\gamma^\nu + \gamma^\nu\gamma^\mu = 2\eta^{\mu\nu}. \tag{50.2}$$

[109]We are formulating this discussion in what particle physicists refer to as "natural units", chosen so that both Planck's constant $\hbar$ and the speed of light c are equal to 1. If the speed of light were not 1, then factors of c^2 would need to appear in the Klein-Gordon equation and several other relations.

This identity obviously cannot be satisfied if the γ^μ are assumed to be mere scalars, but one can find sets of matrices that satisfy it if the right hand side is interpreted as the product of the scalar $2\eta^{\mu\nu}$ with the identity matrix, e.g. Dirac came up with the particular set of complex 4-by-4 matrices

$$\gamma^0 := \begin{pmatrix} \mathbb{1}_{2\times 2} & 0 \\ 0 & -\mathbb{1}_{2\times 2} \end{pmatrix}, \qquad \gamma^j := \begin{pmatrix} 0 & \boldsymbol{\sigma}_j \\ -\boldsymbol{\sigma}_j & 0 \end{pmatrix} \quad \text{for } j = 1,2,3, \tag{50.3}$$

where $\boldsymbol{\sigma}_j$ denote the 2-by-2 Pauli matrices introduced in §39.2. This is not the only possible choice: another popular convention is to keep $\gamma^1, \gamma^2, \gamma^3$ as defined above but set

$$\gamma^0 := \begin{pmatrix} 0 & \mathbb{1}_{2\times 2} \\ \mathbb{1}_{2\times 2} & 0 \end{pmatrix} \tag{50.4}$$

With any such choice of matrices in place, the **Dirac equation** (50.1) can be viewed as a PDE for $\mathbb{C}^4$-valued functions on $\mathbb{R}^4$, which is indeed first-order and implies the Klein-Gordon equation. The invariance of the Dirac equation under $\mathrm{SO}(1,3)$ is a slightly tricky business, because here it once again turns out that the right symmetry group to consider is not $\mathrm{SO}(1,3)$ itself, but instead the universal cover of $\mathrm{SO}(1,3)$, which happens to be a double cover. We will not get into the details here, as they can be derived from the more general discussion in §50.5 below, but suffice it to say that the functions $\psi : \mathbb{R}^4 \to \mathbb{C}^4$ satisfying the Dirac equation are known as *Dirac spinors*, and the vector space $\mathbb{C}^4$ in which they take their values is acted upon via a Lie algebra representation of $\mathfrak{so}(1,3)$ that does not lift to a representation of $\mathrm{SO}(1,3)$, but does determine a projective representation.

REMARK 50.5. The fact that both the domain and the target of the functions fed into Dirac's equation (50.1) are 4-dimensional (real and complex respectively) is a coincidence, nothing more. The use of $\mathbb{C}^4$ results from having chosen a particular representation of the algebraic relation (50.2) via 4-by-4 matrices. One can also find other representations with N-by-N matrices for some $N > 4$, and the Dirac equation then becomes an equation for functions $\psi : \mathbb{R}^4 \to \mathbb{C}^N$. It is not even strictly necessary for the target vector space to be complex, but this is traditional in quantum mechanics.

50.4. Clifford algebras and spin groups. Our goal for the remainder of this lecture will be to write down a coordinate-invariant version of the Dirac equation that makes sense on any pseudo-Riemannian manifold endowed with certain extra structure. The extra structure required turns out to be a spin structure.

The general definition of the spin groups requires a short excursion into the world of Clifford algebras, which can be thought of as a formalization of Dirac's matrix relation (50.2), though they also were known to mathematicians before Dirac. Throughout the following, V denotes a real n-dimensional vector space that is equipped with a symmetric nondegenerate bilinear form $\langle\ ,\ \rangle$. We denote by $\mathrm{O}(V) \subset \mathrm{GL}(V)$ the group of linear transformations that preserve $\langle\ ,\ \rangle$, define the subgroup consisting of orientation-preserving transformations,

$$\mathrm{SO}(V) := \left\{ R \in \mathrm{O}(V) \;\middle|\; \det(R) = 1 \right\}$$

and let $\mathfrak{so}(V)$ denote their common Lie algebra. Note that $\mathrm{SO}(V)$ is the identity component of $\mathrm{O}(V)$ if $\langle\ ,\ \rangle$ is a positive inner product, but in the indefinite case, $\mathrm{SO}(V)$ is generally a union of multiple connected components of $\mathrm{O}(V)$.

DEFINITION 50.6. Let $T(V) := \bigoplus_{k=0}^{\infty} V^{\otimes k}$ denote the tensor algebra of V, which is an associative unital algebra with respect to the product $\otimes$. The **Clifford algebra** of V is

$$\mathrm{Cl}(V) := T(V)/I,$$

where $I \subset T(V)$ is the two-sided ideal generated by all elements of the form $v \otimes v + \langle v, v\rangle$ for $v \in V$. The product of two elements $x = [v]$ and $y = [w]$ in $\mathrm{Cl}(V)$ will be written simply as

$$xy := [v \otimes w] \in \mathrm{Cl}(V).$$

We identify V itself with the subspace of $\mathrm{Cl}(V)$ consisting of all elements of the form $[v] \in \mathrm{Cl}(V)$ for $v \in V = V^{\otimes 1} \subset T(V)$, and $\mathbb{R}$ with the subspace consisting of elements of the form $[t] \in \mathrm{Cl}(V)$ for $t \in \mathbb{R} = V^{\otimes 0} \subset T(V)$.

EXERCISE 50.7. Check that the natural maps $\mathbb{R} \to \mathrm{Cl}(V) : t \mapsto [t]$ and $V \to \mathrm{Cl}(V) : v \mapsto [v]$ are injective.
Hint: If you get stuck, see [**LM89**, page 8].

By definition, every $v \in V \subset \mathrm{Cl}(V)$ satisfies the algebraic relation $v^2 = -\langle v, v\rangle$, so plugging $v + w$ into this relation for arbitrary $v, w \in V$ and expoiting bilinearity gives

$$vw + wv = -2\langle v, w\rangle.$$

Up to a sign, this generalizes the relation (50.2) satisfied by Dirac's γ-matrices if one imagines them as forming an orthonormal basis of the 4-dimensional vector space that they span. The extra sign is a convention, and is not universal—some treatments define the ideal I in $\mathrm{Cl}(V) = T(V)/I$ as generated by elements of the form $v \otimes v - \langle v, v\rangle$, which is more consistent with Dirac's choice of matrices, but our convention translates easily into Dirac's by introducing factors of i where appropriate. For mathematical reasons, some things (such as Examples 50.9 and 50.10 below) work out a bit prettier if one assumes $vw + wv$ is $-2\langle v, w\rangle$ instead of $2\langle v, w\rangle$, so that is what we will do, but there is no deep significance to this choice.

EXERCISE 50.8. Show that $\mathrm{Cl}(V)$ has the same dimension as the exterior algebra $\Lambda^* V = \bigoplus_{k=0}^n \Lambda^k V$, namely $2^n = \sum_{k=0}^n \binom{n}{k}$, and for any choice of orthonormal basis $e_1, \ldots, e_n$ of V, the elements of the form $e_{i_1} \ldots e_{i_k}$ for $k \geqslant 0$ and $1 \leqslant i_1 < \ldots < i_k \leqslant n$ form a basis of $\mathrm{Cl}(V)$. (Note that the case $k = 0$ is included here: the product $e_{i_1} \ldots e_{i_k}$ is in this case understood to be $1 \in \mathrm{Cl}(V)$.)

When V is taken to be $\mathbb{R}^n$ with its standard inner product of signature (k, ℓ), we will denote

$$\mathrm{Cl}(k, \ell) := \mathrm{Cl}(\mathbb{R}^n), \qquad \text{and in particular} \qquad \mathrm{Cl}(n) := \mathrm{Cl}(n, 0).$$

EXAMPLE 50.9. For $\mathbb{R}$ with its Euclidean inner product, we can pick a basis $e \in \mathbb{R}$ with $|e| = 1$, and the elements $1, e \in \mathrm{Cl}(1)$ then form a basis of $\mathrm{Cl}(1)$. Since $e^2 = -\langle e, e\rangle = -1$, the algebra $\mathrm{Cl}(1)$ is isomorphic to the complex numbers.

EXAMPLE 50.10. For $\mathbb{R}^2$ with its Euclidean inner product, a choice of orthonormal basis $e_1, e_2 \in \mathbb{R}^2$ gives rise to a basis of $\mathrm{Cl}(2)$ in the form $1, e_1, e_2, e_3 := e_1 e_2$. The Clifford algebra relations then imply

$$e_3^2 = (e_1 e_2)^2 = e_1 e_2 e_1 e_2 = -e_1^2 e_2^2 = -1,$$

and one can similarly show that $e_2 e_3 = e_1$ and $e_3 e_1 = e_2$. It follows that the algebra $\mathrm{Cl}(2)$ is isomorphic to the quaternions.

REMARK 50.11. The pattern seen in Examples 50.9 and 50.10 clearly cannot be continued further since we have already exhausted the available associative division algebras. Most Clifford algebras indeed do not have the property that all nonzero elements admit multiplicative inverses. The reader might enjoy working out the case $\mathrm{Cl}(1, 1)$ to see what I mean.

Every Clifford algebra contains an open subset

$$\mathrm{Cl}^\times(V) := \left\{ x \in \mathrm{Cl}(V) \;\middle|\; \text{there exists } y \in \mathrm{Cl}(V) \text{ such that } xy = yx = 1 \right\}.$$

To see that it is open, there is a trick that works for any finite-dimensional algebra: choose a norm $\|\cdot\|$ on $\mathrm{Cl}(V)$ and scale it so that $\|xy\| \leqslant \|x\| \cdot \|y\|$ for all $x, y \in \mathrm{Cl}(V)$, making $\mathrm{Cl}(V)$ into a so-called *Banach algebra*. The series $1 - x + x^2 - x^3 + \ldots$ is then guaranteed to converge whenever $\|x\| < 1$, and it defines a multiplicative inverse for $1 + x$. One can use this to derive a series formula for the inverse of $x + y$ whenever x is invertible and $\|y\|$ is sufficiently small. This series formula can also be used to show that the inversion map $\mathrm{Cl}^\times(V) \to \mathrm{Cl}^\times(V) : x \mapsto x^{-1}$ is smooth, and multiplication on $\mathrm{Cl}(V)$ is clearly also smooth, thus $\mathrm{Cl}^\times(V)$ is a Lie group with respect to multiplication. It contains, for instance, every $v \in V \subset \mathrm{Cl}(V)$ with $\langle v, v\rangle \neq 0$, since $-\frac{v}{\langle v,v\rangle}$ is then a multiplicative inverse for v. The subgroup $\mathrm{Spin}(V) \subset \mathrm{Cl}^\times(V)$ is defined as the group consisting of all products of evenly many elements $v \in V$ with $\langle v, v\rangle = \pm 1$, that is,

$$\mathrm{Spin}(V) := \left\{ v_1 \ldots v_{2N} \in \mathrm{Cl}(V) \;\middle|\; N \geqslant 1,\ v_j \in V \text{ such that } \langle v_j, v_j\rangle = \pm 1 \text{ for } j = 1, \ldots, 2N \right\}.$$

It is far from obvious from this definition whether $\mathrm{Spin}(V)$ is a submanifold of $\mathrm{Cl}^\times(V)$, but we will address this issue in Theorem 50.15 below. There is a natural representation of $\mathrm{Spin}(V)$ on $\mathrm{Cl}(V)$, called the **adjoint representation**

$$\mathrm{Ad} : \mathrm{Spin}(V) \to \mathrm{GL}(\mathrm{Cl}(V)) : x \mapsto \mathrm{Ad}_x, \qquad \mathrm{Ad}_x(y) := xyx^{-1}.$$

Observe that the transformations $\mathrm{Ad}_x : \mathrm{Cl}(V) \to \mathrm{Cl}(V)$ are not just linear, they are also algebra homomorphisms, so Ad sends $\mathrm{Spin}(V)$ to the group $\mathrm{Aut}(\mathrm{Cl}(V))$ of **automorphisms** of the Clifford algebra.

EXERCISE 50.12. For any codimension 1 subspace $H \subset V$ on which the restriction of $\langle\ ,\ \rangle$ is nondegenerate, one can define the **reflection about** H as the unique linear map $V \to V$ that fixes every point in H but sends $v \mapsto -v$ for all $v \in H^\perp$. (Note that this definition does not make sense if $\langle\ ,\ \rangle|_H$ is degenerate, because $H^\perp$ is then contained in H; see Lemma 24.7 from the first semester.)

(a) For $x \in V$ with $\langle x, x\rangle = \pm 1$, show that the reflection $V \to V$ about $x^\perp$ is given by $v \mapsto -xvx^{-1}$.
(b) Deduce that for each $x \in \mathrm{Spin}(V)$, the transformation $\mathrm{Ad}_x : \mathrm{Cl}(V) \to \mathrm{Cl}(V)$ preserves the subspace $V \subset \mathrm{Cl}(V)$ and acts on it by orientation-preserving orthogonal transformations.

The upshot of the preceding exercise is that the adjoint representation of $\mathrm{Spin}(V)$ on $\mathrm{Cl}(V)$ gives rise to a natural group homomorphism

$$\Phi : \mathrm{Spin}(V) \to \mathrm{SO}(V) : x \mapsto \mathrm{Ad}_x|_V.$$

We will see below that $\mathrm{Spin}(V)$ is naturally a Lie group and Φ is a smooth double cover, thus generalizing the cover $\mathrm{SU}(2) \to \mathrm{SO}(3)$. Proving this will require computing the kernel of Φ, for which the following exercise serves as preparation.

EXERCISE 50.13. An **anti-homomorphism** $\mathcal{A} \to \mathcal{B}$ between two associative algebras is by definition a linear map $\psi : \mathcal{A} \to \mathcal{B}$ such that $\psi(xy) = \psi(y)\psi(x)$ for all $x, y \in \mathcal{A}$. Prove:

(a) Every linear map $V \to V$ has a unique extension to an anti-homomorphism $T(V) \to T(V)$, and the extension is bijective if and only if the given map $V \to V$ is an isomorphism.
(b) For the identity map $V \to V$, the unique extension to an anti-homomorphism $T(V) \to T(V)$ descends to the quotient $\mathrm{Cl}(V) = T(V)/I$, thus defining an anti-homomorphism

$$\psi : \mathrm{Cl}(V) \to \mathrm{Cl}(V)$$

that is the identity map on V.
(c) For $x \in \mathrm{Spin}(V)$, the anti-homomorphism $\psi : \mathrm{Cl}(V) \to \mathrm{Cl}(V)$ satisfies $\psi(x) = x^{-1}$.

The next exercise gives further evidence for a close relationship between $\mathrm{Spin}(V)$ and $\mathrm{SO}(V)$ by showing that the algebraic properties of the product on $\mathrm{Cl}(V)$ secretly contain the Lie algebra structure of $\mathfrak{so}(V)$.

EXERCISE 50.14. Given an orthonormal basis $e_1, \dots, e_n \in V$, let $\mathfrak{spin}(V) \subset \mathrm{Cl}(V)$ denote the vector space spanned by all products of the form $e_i e_j$ for $i \neq j$. Prove:

(a) $\mathfrak{spin}(V) \subset \mathrm{Cl}(V)$ does not depend on the choice of orthonormal basis $e_1, \dots, e_n \in V$.
(b) $\mathfrak{spin}(V)$ is a Lie algebra with respect to the commutator bracket $[x, y] := xy - yx$.
(c) For any $v, w \in V$ satisfying $\langle v, v\rangle = \pm 1$, $\langle w, w\rangle = \pm 1$ and $\langle v, w\rangle = 0$, we have $vw \in \mathfrak{spin}(V)$ and $e^{\frac{1}{2}tvw} \in \mathrm{Spin}(V)$ for all $t \in \mathbb{R}$, where for $x \in \mathrm{Cl}(V)$, we define $e^x := \sum_{k=0}^{\infty} \frac{x^k}{k!} \in \mathrm{Cl}(V)$.
(d) Under the assumptions of part (c), can you give a geometric interpretation to the family of transformations $\Phi(e^{\frac{1}{2}tvw}) \in \mathrm{SO}(V)$?
Hint: Evaluate $\Phi(e^{\frac{1}{2}vw})$ on v and w and on an arbitrary vector orthogonal to both.
(e) Construct a smooth map $\varphi : \mathfrak{spin}(V) \to \mathrm{Cl}(V)$ whose derivative at $0 \in \mathfrak{spin}(V)$ is the inclusion $\mathfrak{spin}(V) \hookrightarrow \mathrm{Cl}(V)$, such that the image of φ is in $\mathrm{Spin}(V)$ and the derivative of $\Phi \circ \varphi : \mathfrak{spin}(V) \to \mathrm{SO}(V)$ at 0 is a Lie algebra isomorphism $\mathfrak{spin}(V) \to \mathfrak{so}(V)$.
Hint: Using the orthonormal basis $e_1, \dots, e_n \in V$, first define $\varphi(te_ie_j)$ for each $t \in \mathbb{R}$ and $i \neq j$, then extend it to the rest of $\mathfrak{spin}(V)$ in whatever way is convenient.

THEOREM 50.15. *The group $\mathrm{Spin}(V)$ is a smooth submanifold of $\mathrm{Cl}(V)$ and is thus a Lie group. Moreover, the group homomorphism $\Phi : \mathrm{Spin}(V) \to \mathrm{SO}(V)$ defined by restricting the adjoint representation to $V \subset \mathrm{Cl}(V)$ is a covering map of degree 2.*

SKETCH OF THE PROOF. By Exercise 50.12, the image of the map $\Phi : \mathrm{Spin}(V) \to \mathrm{SO}(V)$ consists of all products of evenly many reflections. By a classical result known as the Cartan-Dieudonné theorem (see e.g. [**Gar11**, §4.8]), all of $\mathrm{O}(V)$ is generated by reflections, and since each individual reflection is orientation reversing, it follows that the image of Φ is $\mathrm{SO}(V)$. We claim that the kernel of Φ contains only the two elements ± 1. Indeed, if $x \in \mathrm{Spin}(V)$ satisfies $xvx^{-1} = v$ for all $v \in V$, then x commutes with all of $\mathrm{Cl}(V)$, which implies $x \in \mathbb{R} \subset \mathrm{Cl}(V)$, so the claim reduces to showing that ± 1 are the only two scalars in $\mathrm{Spin}(V)$. One sees this from Exercise 50.13: the unique anti-homomorphism $\psi : \mathrm{Cl}(V) \to \mathrm{Cl}(V)$ that satisfies $\psi(v) = v$ for all $v \in V$ also satisfies $\psi(x) = x^{-1}$ for all $x \in \mathrm{Spin}(V)$, and if $x \in \mathrm{Spin}(V) \cap \mathbb{R}$, it follows that $x = x^{-1}$, and thus $x = \pm 1$.

Having established that $\Phi : \mathrm{Spin}(V) \to \mathrm{SO}(V)$ is a surjective group homomorphism with kernel $\{\pm 1\}$, the smoothness of $\mathrm{Spin}(V)$ follows easily in light of Exercise 50.14: in particular, the inverse function theorem provides a neighborhood $\mathcal{U}$ of 0 in the vector space $\mathfrak{spin}(V) \subset \mathrm{Cl}(V)$ such that the map $\mathcal{U} \to \mathrm{SO}(V) : X \mapsto \Phi(\varphi(X))$ is a diffeomorphism onto a neighborhood of $\mathbb{1}$ in $\mathrm{SO}(V)$. This is enough information to conclude that $\mathcal{U} \to \mathrm{Spin}(V) : X \mapsto \varphi(X)$ gives a parametrization of $\mathrm{Spin}(V)$ on a neighborhood of 1, from which one deduces that $\mathrm{Spin}(V)$ is a Lie group. □

From now on we will use the following notation: for $V = \mathbb{R}^n$ with the standard indefinite inner product of signature (k, ℓ), we write

$$\mathrm{Spin}(k, \ell) := \mathrm{Spin}(V) \subset \mathrm{Cl}(k, \ell),$$

and for positive inner products in particular,

$$\mathrm{Spin}(n) := \mathrm{Spin}(n, 0) \subset \mathrm{Cl}(n).$$

By definition, every spin group comes equipped with a two-to-one Lie group homomorphism to the corresponding special orthogonal group,

$$\Phi : \mathrm{Spin}(k, \ell) \to \mathrm{SO}(k, \ell).$$

EXAMPLE 50.16. Since $\mathrm{SO}(3) \cong \mathbb{RP}^3$ has a double cover as its universal cover, standard covering space theory implies that *every* nontrivial covering space of SO(3) is simply connected, hence Spin(3) is also the universal cover of SO(3). It then follows via the uniqueness of the universal cover (cf. Corollary 39.24) that Spin(3) is isomorphic to SU(2).

EXERCISE 50.17. What is Spin(2)?
Hint: This is a rare example of a spin group that is not simply connected.

50.5. Dirac operators on manifolds. Now that the groups $\mathrm{Spin}(k,\ell)$ are in place, you could presumably write down the general definition of a spin structure yourself.

DEFINITION 50.18. Suppose $E \to M$ is an oriented real vector bundle of rank m with a bundle metric $\langle\ ,\ \rangle$ of signature (k,ℓ). A **spin structure** on E is an equivalence class of $\mathrm{SO}(k,\ell)$-bundle atlases on $E \to M$ with transition functions $g_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{SO}(k,\ell)$, together with a system of $\mathrm{Spin}(k,\ell)$-valued transition functions $h_{\beta\alpha} : \mathcal{U}_\alpha \cap \mathcal{U}_\beta \to \mathrm{Spin}(k,\ell)$ such that $g_{\beta\alpha} = \Phi \circ h_{\beta\alpha}$ for the double cover $\Phi : \mathrm{Spin}(k,\ell) \to \mathrm{SO}(k,\ell)$. Equivalently, a spin structure on E is a principal $\mathrm{Spin}(k,\ell)$-bundle $P \to M$ together with a fiber-preserving smooth map $\Psi : P \to F^{\mathrm{SO}}(E) := F^{\mathrm{SO}(k,\ell)}(E)$ that sends P_p to $F^{\mathrm{SO}}(E_p)$ for each $p \in M$ and satisfies $\Psi(\phi g) = \Psi(\phi)\Phi(g)$ for each $\phi \in P$ and $g \in \mathrm{Spin}(k,\ell)$.

The case most frequently of interest is where $E \to M$ is the tangent bundle of an oriented pseudo-Riemannian manifold (M,g), and a spin structure on TM is then called a **spin structure on** (M,g). While physicists are primarily interested in manifolds with Lorentzian signature, the theory of spin structures on Riemannian manifolds has nicer properties and has had a far greater impact on pure mathematics. One simple reason for this is that, just like the corresponding special orthogonal groups, $\mathrm{Spin}(n)$ is compact and connected, while $\mathrm{Spin}(k,\ell)$ for $k,\ell \geqslant 1$ is neither (prove it!). A spin structure on (M,g) also makes it possible to define a coordinate-invariant generalization of the Dirac equation for sections of certain vector bundles over M, and we will see that from an analytical perspective, this equation likewise has nicer properties in the Riemannian than in the indefinite case.

50.5.1. *Clifford and spinor bundles.* The definition of a Dirac operator requires two fundamental ingredients: one is a pseudo-Riemannian manifold with a spin structure, and the other is a representation of the corresponding Clifford algebra. Let's start with the latter. Let

$$\eta_{ij} = \eta^{ij} := \pm\delta_{ij}$$

denote the components of the standard flat metric on $\mathbb{R}^n$ with signature (k,ℓ), and choose a set of matrices $\gamma_1,\dots,\gamma_n \in \mathbb{C}^{N\times N}$ that satisfy the relations

$$\gamma_i\gamma_j + \gamma_j\gamma_i = -2\eta_{ij}\mathbb{1} \in \mathbb{C}^{N\times N} \tag{50.5}$$

for all $i,j \in \{1,\dots,n\}$. Any such choice determines a unique representation of $\mathrm{Cl}(k,\ell)$ via linear maps on $\mathbb{C}^N$, i.e. it makes $\mathbb{C}^N$ into a module over $\mathrm{Cl}(k,\ell)$, with the standard basis $e_1,\dots,e_n \in \mathbb{R}^n \subset \mathrm{Cl}(k,\ell)$ acting on $\mathbb{C}^N$ by

$$e_j v := \gamma_j v, \qquad v \in \mathbb{C}^N,\ j = 1,\dots,n.$$

For certain formulas it will be useful to pretend that the matrices $\gamma_1,\dots,\gamma_n$ are the local coordinate components of a 1-form that can be fed into a musical isomorphism to raise its index, thus defining new matrices $\gamma^1,\dots,\gamma^n$ by

$$\gamma^i := \eta^{ij}\gamma_j \in \mathbb{C}^{N\times N}.$$

These also satisfy the relations in (50.5), since $\gamma^i = \langle e_i, e_i\rangle\gamma_i = \pm\gamma_i$ for each i. The representations of $\mathrm{Cl}(k,\ell)$ can be classified using standard methods of representation theory, and a full account of it can be found in most books on spin geometry. We will skip that detail here, other than to observe that we've already seen a couple of examples: up to factors of i, the Dirac matrices in (50.3) or (50.4) define representations of $\mathrm{Cl}(1,3)$, and a representation of $\mathrm{Cl}(3)$ can be defined by putting factors of i in front of the Pauli matrices in §39.2.

Now suppose (M,g) is an oriented pseudo-Riemannian manifold of signature (k,ℓ) with a spin structure, represented by a principal $\mathrm{Spin}(k,\ell)$-bundle $P \to M$ that doubly covers $F^{\mathrm{SO}}(TM)$. The representation we've chosen for the algebra $\mathrm{Cl}(k,\ell)$ restricts to $\mathrm{Spin}(k,\ell) \subset \mathrm{Cl}(k,\ell)$ as a group representation

$$\sigma : \mathrm{Spin}(k,\ell) \to \mathrm{GL}(N,\mathbb{C}),$$

and thus enables us to build an associated vector bundle, the **spinor bundle**

$$E := P^{\sigma} = (P \times \mathbb{C}^N) \big/ \mathrm{Spin}(k,\ell) \to M.$$

This is a complex vector bundle over M that has the same $\mathrm{Spin}(k,\ell)$-valued transition functions as the spin structure on TM.

We next define another vector bundle, the so-called **Clifford bundle**

$$\mathrm{Cl}(TM) := \bigcup_{p \in M} \mathrm{Cl}(T_pM),$$

whose fiber over each point p is the Clifford algebra of the tangent space T_pM with its inner product g_p of signature (k,ℓ). To see that this set really is a smooth vector bundle in a natural way, we can define it as another associated bundle. Observe that the canonical action of $\mathrm{SO}(k,\ell)$ on $\mathbb{R}^n$ extends uniquely to an $\mathrm{SO}(k,\ell)$-action on $\mathrm{Cl}(k,\ell)$ via algebra homomorphisms; this action is well defined since $\mathrm{SO}(k,\ell)$ preserves the scalar product on $\mathbb{R}^n$ appearing in the Clifford algebra relation, so that the obvious action of $\mathrm{SO}(k,\ell)$ by algebra homomorphisms on the tensor algebra $T(\mathbb{R}^n)$ preserves the ideal $I \subset T(\mathbb{R}^n)$ generated by elements $\mathbf{v}^2 + \langle \mathbf{v}, \mathbf{v} \rangle$ for $\mathbf{v} \in \mathbb{R}^n$. Composing the representation $\mathrm{SO}(k,\ell) \to \mathrm{GL}(\mathrm{Cl}(k,\ell))$ with the double cover $\mathrm{Spin}(k,\ell) \to \mathrm{SO}(k,\ell)$ then gives a representation of $\mathrm{Spin}(k,\ell)$ on $\mathrm{Cl}(k,\ell)$, and it is one that we've seen before: it is the adjoint representation, since the restriction of the latter to $\mathbb{R}^n$ was used to define the action of $\mathrm{Spin}(k,\ell)$ on $\mathbb{R}^n$. The Clifford bundle can now be defined as the unique $\mathrm{SO}(k,\ell)$-bundle with the same transition functions as TM and standard fiber $\mathrm{Cl}(k,\ell)$; equivalently, it is the unique $\mathrm{Spin}(k,\ell)$-bundle with standard fiber $\mathrm{Cl}(k,\ell)$ such that the transition functions come from the spin structure of TM. In terms of principal bundles, we obtain two equivalent definitions of $\mathrm{Cl}(TM)$ as an associated bundle,

$$\mathrm{Cl}(TM) = \left(F^{\mathrm{SO}}(TM) \times \mathrm{Cl}(k,\ell)\right) \big/ \mathrm{SO}(k,\ell) = \left(P \times \mathrm{Cl}(k,\ell)\right) \big/ \mathrm{Spin}(k,\ell).$$

Since $\mathrm{SO}(k,\ell)$ and $\mathrm{Spin}(k,\ell)$ act on $\mathrm{Cl}(k,\ell)$ by algebra isomorphisms, each fiber of $\mathrm{Cl}(TM)$ now naturally inherits the structure of a Clifford algebra isomorphic to $\mathrm{Cl}(k,\ell)$.

We claim that there is a natural smooth linear bundle map

$$\mathrm{Cl} : \mathrm{Cl}(TM) \otimes E \to E : X \otimes \eta \mapsto X\eta, \tag{50.6}$$

which can be interpreted as a bilinear bundle map $\mathrm{Cl}(TM) \oplus E \to E$ and makes each fiber E_p into a module over the Clifford algebra $\mathrm{Cl}(T_pM)$. One sees this by viewing both E and $\mathrm{Cl}(TM)$ as associated bundles for the principal $\mathrm{Spin}(k,\ell)$-bundle $P \to M$, because the linear map

$$\mathrm{Cl}(k,\ell) \otimes \mathbb{C}^N \to \mathbb{C}^N$$

defined via the chosen representation of $\mathrm{Cl}(k,\ell)$ on $\mathbb{C}^N$ is $\mathrm{Spin}(k,\ell)$-equivariant. This follows easily from the fact that $\mathrm{Spin}(k,\ell)$ acts on $\mathrm{Cl}(k,\ell)$ via conjugation. The resulting map (50.6) is called **Clifford multiplication** on E.

Only one more ingredient is needed before we can write down a Dirac equation: if we want to differentiate sections of the spinor bundle E, we need a connection on E. As it turns out, there is a canonical choice: $TM \to M$ already has a canonical connection, the Levi-Cività connection, which determines a principal connection on $F^{\mathrm{SO}}(TM)$. Since $\Psi : P \to F^{\mathrm{SO}}(TM)$ is a covering map of degree 2, every parallel section of $F^{\mathrm{SO}}(TM)$ along a path in M has exactly two lifts to sections of P along the same path, and it follows that there is a uniquely determined parallel transport map on P that commutes with the covering map. The resulting principal connection on $P \to M$ and its associated connection on $E \to M$ are both called the **spin connection**.

EXERCISE 50.19. Show that the spin connection on $E \to M$ is compatible with the Clifford multiplication map (50.6) and the Levi-Cività connection on $TM \to M$ in the following sense: for all $X, Y \in \mathfrak{X}(M)$ and $\eta \in \Gamma(E)$,

$$\nabla_X(Y\eta) = (\nabla_X Y)\eta + Y\nabla_X \eta.$$

Another way of putting this is that Clifford multiplication defines a bundle map $TM \to \mathrm{End}(E)$ which is parallel with respect to these two connections on TM and E.

With this data in place, we can finally define the **Dirac operator**

$$D : \Gamma(E) \to \Gamma(E)$$

as follows. The covariant derivative ∇ takes a section $\eta \in \Gamma(E)$ to a section $\nabla\eta$ of $\mathrm{Hom}(TM, E)$, which is naturally isomorphic to $T^*M \otimes E$. The bundle metric on TM determines a musical isomorphism $\sharp : T^*M \to TM$ that is inverse to $\flat : TM \to T^*M : X \mapsto \langle X, \cdot\rangle$, thus giving a linear bundle map $\sharp \otimes \mathbb{1} : T^*M \otimes E \to TM \otimes E$ sending products $\lambda \otimes \eta \in T_p^*M \otimes E_p$ at each point $p \in M$ to $\lambda^\sharp \otimes \eta \in T_pM \otimes E_p$. Any section of $TM \otimes E$ can then be plugged into the Clifford multiplication $\mathrm{Cl} : TM \otimes E \to E$, producing a section in E. In total, D is the composition of the covariant derivative operator ∇ with two smooth linear bundle maps:

$$\begin{tikzcd} \Gamma(E) \arrow[r, "\nabla"] \arrow[rrr, bend left, "D"] & \Gamma(T^*M \otimes E) \arrow[r, "\sharp\otimes\mathbb{1}"] & \Gamma(TM \otimes E) \arrow[r, "\mathrm{Cl}"] & \Gamma(E). \end{tikzcd}$$

To see that the result of this rather abstract construction is actually something familiar, choose a local frame $X_1, \dots, X_n$ for TM over some region $\mathcal{U} \subset M$ that is orthonormal, meaning $\langle X_i, X_j\rangle = \eta_{ij}$. Any vector field Y over $\mathcal{U}$ can then be written as $Y = Y^i X_i$ for suitable component functions $Y^i : \mathcal{U} \to \mathbb{R}$, which satisfy

$$\langle X_i, Y\rangle = \langle X_i, Y^j X_j\rangle = \eta_{ij} Y^j =: Y_i, \qquad \text{thus} \qquad Y^i = \eta^{ij} Y_j = \eta^{ij}\langle X_j, Y\rangle.$$

For any $\eta \in \Gamma(E|_{\mathcal{U}})$, we thus have

$$\nabla\eta = \nabla_{(\cdot)}\eta = \nabla_{\eta^{ij}\langle X_j, \cdot\rangle X_i}\eta = \eta^{ij}\langle X_j, \cdot\rangle \otimes \nabla_{X_i}\eta,$$

and applying $\sharp \otimes \mathbb{1}$ turns this into $\eta^{ij} X_j \otimes \nabla_{X_i}\eta$. Feeding this into Clifford multiplication thus gives the local formula

$$D\eta = \eta^{ij} X_j \nabla_{X_i}\eta =: X^i \nabla_{X_i}\eta,$$

where as a matter of notational convenience, we are defining the Clifford multiplication of X^i (with raised index) on E to be the Clifford multiplication of the linear combination $\eta^{ij} X_j = \langle X_i, X_i\rangle X_i = \pm X_i$.

EXAMPLE 50.20. In the special case where M is $\mathbb{R}^n$ with a flat metric having components $\eta_{ij} = \langle \partial_i, \partial_j\rangle$, the bundle TM and its connection are both trivial, so the frame bundle $F^{\mathrm{SO}}(TM)$ and principal $\mathrm{Spin}(k, \ell)$-bundle P are also trivial, and therefore so is the spinor bundle, implying that we can describe sections $\eta \in \Gamma(E)$ as functions $\eta : \mathbb{R}^n \to \mathbb{C}^N$. We can also take the orthonormal frame in the calculation above to be $\partial_1, \dots, \partial_n$, and the action of these tangent vectors on $E = \mathbb{R}^n \times \mathbb{C}^N$

by Clifford multiplication is then given by the matrices $\gamma_1, \dots, \gamma_n$ that were chosen in (50.5) for defining the representation of $\mathrm{Cl}(k, \ell)$ on $\mathbb{C}^N$. The Dirac operator in this case thus becomes

$$D : C^\infty(\mathbb{R}^n, \mathbb{C}^N) \to C^\infty(\mathbb{R}^n, \mathbb{C}^N) : \eta \mapsto \gamma^j \partial_j \eta,$$

exactly as in the classical Dirac equation.

50.5.2. *Self-adjointness.* [110]

One of the fundamental properties of Dirac operators on Riemannian manifolds is that they are **formally self-adjoint**, meaning that for a natural choice of Hermitian bundle metric $\langle\ ,\ \rangle$ on the spinor bundle $E \to M$, they satisfy

$$\int_M \langle \xi, D\eta \rangle \, d\mathrm{vol} = \int_M \langle D\xi, \eta \rangle \, d\mathrm{vol}$$

for all smooth compactly supported sections $\xi, \eta \in \Gamma(E)$. This is one detail on which it makes a difference whether the metric on M is positive or indefinite.

To see why it works, we first need to be slightly more specific about our choice of representation for $\mathrm{Cl}(k, \ell)$ on a complex vector space $\mathbb{C}^N$. We claim that without changing the value of N, the matrices $\gamma_1, \dots, \gamma_n$ satisfying (50.5) can be chosen to be unitary. Indeed, the standard orthonormal basis $e_1, \dots, e_n \in \mathbb{R}^n$ generates a finite multiplicative subgroup

$$G := \left\{ \pm e_{i_1} \dots e_{i_m} \in \mathrm{Cl}^\times(k, \ell) \;\middle|\; m \geqslant 0,\ 1 \leqslant i_1 < \dots < i_m \leqslant n \right\} \subset \mathrm{Cl}^\times(k, \ell),$$

and any representation of $\mathrm{Cl}(k, \ell)$ on $\mathbb{C}^N$ restricts to G as a group representation $G \to \mathrm{GL}(N, \mathbb{C})$. (Conversely, it is not hard to show that any representation $\rho : G \to \mathrm{GL}(N, \mathbb{C})$ satisfying $\rho(-1) = -\mathbb{1}$ extends uniquely to a representation of $\mathrm{Cl}(k, \ell)$.) Now since G is finite, the usual averaging trick (cf. Theorem 38.13) can be used to construct a Hermitian inner product on $\mathbb{C}^N$ that is invariant under the action of G. If we now conjugate our chosen representation of $\mathrm{Cl}(k, \ell)$ by a transformation of $\mathbb{C}^N$ changing the standard basis to one that is orthonormal for the newly constructed inner product, the matrices representing $e_1, \dots, e_n$ and all their products become unitary.

Unitarity has a useful consequence when g is positive that does not hold more generally:

Lemma 50.21. *For any choice of matrices $\gamma_1, \dots, \gamma_n \in \mathbb{C}^{N \times N}$ satisfying (50.5), the induced representation $\sigma : \mathrm{Spin}(k, \ell) \to \mathrm{GL}(N, \mathbb{C})$ has image in $\mathrm{SL}_\pm(N, \mathbb{C}) := \left\{ \mathbf{A} \in \mathbb{C}^{N \times N} \;\middle|\; \det \mathbf{A} \in \{1, -1\} \right\}$, so in particular, in the Euclidean signature case $\eta_{ij} = \delta_{ij}$, we conclude $\sigma(\mathrm{Spin}(n)) \subset \mathrm{SL}(N, \mathbb{C})$. If we additionally require the matrices γ_i to be unitary, then the action of $\mathrm{Cl}(n)$ on $\mathbb{C}^N$ has the following properties:*

(1) $\mathbb{R}^n \subset \mathrm{Cl}(n)$ acts on $\mathbb{C}^N$ by anti-Hermitian transformations.
(2) The representation σ of $\mathrm{Spin}(n)$ on $\mathbb{C}^N$ is unitary.

Proof. For $i \neq j$, we have $\gamma_i \gamma_j = -\gamma_j \gamma_i$ and thus $\mathrm{tr}(\gamma_i \gamma_j) = \mathrm{tr}(\gamma_j \gamma_i) = -\mathrm{tr}(\gamma_j \gamma_i)$, implying that $\gamma_i \gamma_j$ belongs to the space of traceless matrices $\mathfrak{sl}(N, \mathbb{C})$. By Exercise 50.14, these products generate the image of the induced Lie algebra representation $\sigma_* : \mathfrak{spin}(k, \ell) \to \mathfrak{gl}(N, \mathbb{C})$, and it follows that σ maps the identity component of $\mathrm{Spin}(k, \ell)$ into $\mathrm{SL}(N, \mathbb{C})$. If k and ℓ are both positive, then $\mathrm{SO}(k, \ell)$ has exactly two connected components. (The second one contains transformations that simultaneously reverse the orientations of spacelike and timelike subspaces, e.g. $-\mathbb{1} \in \mathrm{SO}(1, 3)$).) It follows that $\mathrm{Spin}(k, \ell)$ in this case also has two components, and since $\det \circ \sigma \equiv 1$ on the identity component, we conclude that -1 is the only other possible value, and is excluded in the case

[110] The actual lecture ended with the definition of the Dirac operator and an extremely vague statement of the Atiyah-Singer index theorem, so everything else from this point onward in Lecture 50 is included purely for the sake of interest.

$(k, \ell) = (n, 0)$ where $\mathrm{SO}(n)$ and $\mathrm{Spin}(n)$ are connected. Continuing in the Euclidean signature case with $\gamma_i \in \mathrm{U}(N)$ for all i, the relation $\gamma_i\gamma_j + \gamma_j\gamma_i = -2\delta_{ij}$ implies $\gamma_i^{-1} = -\gamma_i$ and thus

$$\gamma_i^\dagger = \gamma_i^{-1} = -\gamma_i,$$

so that every vector in $\mathbb{R}^n$ acts on $\mathbb{C}^N$ by a linear combination of anti-Hermitian matrices, which is therefore also anti-Hermitian. Similarly, for $i \neq j$, we now find

$$(\gamma_i\gamma_j)^\dagger = \gamma_j^\dagger\gamma_i^\dagger = \gamma_j\gamma_i = -\gamma_i\gamma_j,$$

implying $\gamma_i\gamma_j \in \mathfrak{u}(N)$, and the image of $\sigma_* : \mathfrak{spin}(n) \to \mathfrak{gl}(N, \mathbb{C})$ is therefore also contained in $\mathfrak{u}(N)$. Since $\mathrm{Spin}(n)$ is connected, this implies $\sigma(\mathrm{Spin}(n)) \subset \mathrm{U}(N)$. □

Exercise 50.22. Find an explicit representation of $\mathrm{Cl}(1,1)$ on $\mathbb{C}^2$ for which the induced representation $\mathrm{Spin}(1,1) \to \mathrm{SL}_\pm(2,\mathbb{C})$ does not have image contained in $\mathrm{SL}(2,\mathbb{C})$.

Lemma 50.21 implies that if (M, g) is a Riemannian manifold with a spin structure, we are free to assume the associated spinor bundle $E \to M$ of rank N has its structure group contained in $\mathrm{U}(N)$, thus it has a natural Hermitian bundle metric $\langle\ ,\ \rangle$ that is compatible with the spin connection. Moreover, the map $TM \to \mathrm{End}(E)$ defined via Clifford multiplication has its image in the linear subbundle

$$\mathfrak{u}(E) \subset E,$$

whose fiber over each point $p \in M$ is the space of anti-Hermitian transformations $E_p \to E_p$. This plus the fact from Exercise 50.19 that Clifford multiplication is parallel with respect to the spin connection will be enough to conclude that the Dirac operator is formally self-adjoint.

Proofs of such statements are always based in some fashion on the combination of a Leibniz rule with the fact that $\int_M d\lambda = 0$ for every compactly supported exact $(n-1)$-form λ. In the present context, the latter is most conveniently rephrased as the identity

$$\int_M \mathrm{div}(X)\, d\mathrm{vol} = 0 \qquad \text{for all } X \in \mathfrak{X}(M) \text{ with compact support,} \tag{50.7}$$

where $d\mathrm{vol} \in \Omega^n(M)$ denotes the Riemannian volume form and the **divergence** of a vector field is defined as the unique function $\mathrm{div}(X) : M \to \mathbb{R}$ satisfying

$$\mathcal{L}_X d\mathrm{vol} = \mathrm{div}(X) \cdot d\mathrm{vol}.$$

The vanishing of the integral in (50.7) is an easy consequence of Stokes' theorem since, by Cartan's magic formula, $\mathcal{L}_X d\mathrm{vol} = d(\iota_X d\mathrm{vol})$, implying $\int_M \mathrm{div}(X)\, d\mathrm{vol} = \int_M d(\iota_X d\mathrm{vol})$. We computed a local coordinate formula for $\mathrm{div}(X)$ in Exercise 12.16 last semester: writing $X = X^j\partial_j$ and $d\mathrm{vol} = f\, dx^1 \wedge \ldots \wedge dx^n$ for a function f defined on the domain of the coordinates, one finds

$$\mathrm{div}(X) = \frac{1}{f}\partial_j(fX^j).$$

In the pseudo-Riemannian context, the Levi-Cività connection can be used to write down a more useful coordinate-invariant formula for $\mathrm{div}(X)$. Assume in particular that the coordinates $(x^1, \ldots, x^n)$ are Riemann normal coordinates about a point $p \in M$, so that the Christoffel symbols of ∇ vanish at p, and by Proposition 36.13, the function f satisfies $f(p) = 1$ and $\partial_1 f(p) = \ldots = \partial_n f(p) = 0$. At this one point, we therefore have

$$\mathrm{div}(X)(p) = \partial_j X^j(p) = (\nabla_j X)^j(p) = \mathrm{tr}(\nabla X(p)).$$

Since the same calculation can be carried out in Riemann normal coordinates about any point, and neither of the real-valued functions $\mathrm{div}(X)$ and $\mathrm{tr}(\nabla X)$ actually depends on a choice of coordinates, this proves the general formula

$$\mathrm{div}(X) = \mathrm{tr}(\nabla X), \tag{50.8}$$

which is valid for the Levi-Cività connection on any pseudo-Riemannian manifold (M, g), assuming the volume form $d$vol used in the definition of $\mathrm{div}(X)$ is the canonical one. If the metric is positive, one can express the trace in terms of any local orthonormal frame $e_1, \ldots, e_n$ for TM as

$$\mathrm{div}(X) = \sum_{j=1}^{n} \langle e_j, \nabla_{e_j} X \rangle. \tag{50.9}$$

THEOREM 50.23. *Assume (M, g) is a Riemannian manifold with a spin structure, and $E \to M$ is an associated spinor bundle with a Hermitian bundle metric $\langle\ ,\ \rangle$ such that the spin connection ∇ is compatible with $\langle\ ,\ \rangle$ and the map $TM \to \mathrm{End}(E)$ defined via Clifford multiplication takes values in $\mathfrak{u}(E) \subset \mathrm{End}(E)$. Then the Dirac operator $D : \Gamma(E) \to \Gamma(E)$ is formally self-adjoint.*

PROOF. By (50.7), it will suffice to prove that for any smooth compactly supported sections $\xi, \eta \in \Gamma(E)$, the complex-valued function $\langle \xi, D\eta \rangle - \langle D\xi, \eta \rangle$ is the divergence of a compactly-supported vector field. The vector field $X \in \mathfrak{X}(M)$ in question is uniquely determined by the condition that for every $Y \in \mathfrak{X}(M)$,

$$g(Y, X) = \langle \xi, Y\eta \rangle,$$

with Y on the right hand side acting on η via Clifford multiplication. Indeed, picking another vector field $Z \in \mathfrak{X}(M)$ and applying the operator ∇_Z to both sides of this relation gives

$$g(\nabla_Z Y, X) + g(Y, \nabla_Z X) = \langle \nabla_Z \xi, Y\eta \rangle + \langle \xi, Y \nabla_Z \eta \rangle + \langle \xi, (\nabla_Z Y)\eta \rangle,$$

in which we have used the compatibility of ∇ with the bundle metrics on both TM and E and also with Clifford multiplication. Since $\langle \xi, (\nabla_Z Y)\eta \rangle = g(\nabla_Z Y, X)$ and the action of Y on E is anti-Hermitian, this simplifies to

$$g(Y, \nabla_Z X) = \langle \nabla_Z \xi, Y\eta \rangle + \langle \xi, Y \nabla_Z \eta \rangle = -\langle Y \nabla_Z \xi, \eta \rangle + \langle \xi, Y \nabla_Z \eta \rangle.$$

Choosing a local orthonormal frame $e_1, \ldots, e_n$, plugging in $Y = Z = e_j$ and summing over $j = 1, \ldots, n$ now gives

$$\mathrm{div}(X) = -\langle D\xi, \eta \rangle + \langle \xi, D\eta \rangle$$

as claimed. □

The perspective taken in this lecture has been that certain geometric objects arise naturally out of a desire to formulate coordinate-invariant versions of notions that originate in theoretical physics. This is a true statement, but it also severely understates the impact that Dirac operators have subsequently had on pure mathematics. The most prominent example is probably the *Atiyah-Singer index theorem*, which gives a deep relationship between topology and analysis. This theorem makes crucial use of the fact that on any Riemannian manifold, Dirac operators are *elliptic*—as we will see when we study Hodge theory, ellipticity is a useful property of linear differential operators, guaranteeing in particular that if the underlying manifold M is compact, then the dimension of the solution space $\ker D \subset \Gamma(E)$ and of the cokernel $\mathrm{coker}\, D := \Gamma(E)/\mathrm{im}(D)$ are both finite. The **index** of an operator with this property is defined as the integer

$$\mathrm{ind}(D) := \dim \ker(D) - \mathrm{codim}\, \mathrm{im}(D) = \dim \ker(D) - \dim \mathrm{coker}(D) \in \mathbb{Z},$$

and the Atiyah-Singer theorem created a thriving industry of extracting topological invariants from the indices of elliptic operators on manifolds. For a Dirac operator on its own, the index is not very interesting because self-adjointness implies $\mathrm{codim}\, \mathrm{im}(D) = \dim \ker(D)$, so that the index is automatically 0. More can be said however if the underlying representation of $\mathrm{Cl}(n)$ is chosen to have a special property: it can typically be arranged that the matrices $\gamma_1, \ldots, \gamma_n \in \mathbb{C}^{N \times N}$ defining this representation all admit splittings into off-diagonal block form

$$\gamma_j = \begin{pmatrix} 0 & \alpha_j \\ \beta_j & 0 \end{pmatrix}$$

with blocks α_j, β_j of equal size. (This is for instance not true of the original set of Dirac matrices in (50.3), but becomes true if we replace γ_0 with the alternative in (50.4).) The geometric meaning of this condition is that $\mathbb{C}^N$ splits into two subspaces of equal dimension that are interchanged by the action of $\mathbb{R}^n \subset \mathrm{Cl}(n)$, but therefore also preserved by the action of $\mathrm{Spin}(n)$. As a result, the spinor bundle $E \to M$ likewise splits into two subbundles $E = E^+ \oplus E^-$ of equal rank such that the Dirac operator sends $\Gamma(E^\pm) \to \Gamma(E^\mp)$, producing a similar off-diagonal splitting

$$D = \begin{pmatrix} 0 & D_+ \\ D_- & 0. \end{pmatrix}$$

We call D an **odd Dirac operator** if it admits a splitting of this form. The formal self-adjointness of D can now be used to show that $\mathrm{coker}(D_+) \cong \ker(D_-)$ and $\ker(D_+) \cong \mathrm{coker}(D_-)$, hence $\mathrm{ind}(D) = \mathrm{ind}(D_+) + \mathrm{ind}(D_-) = 0$, but $\mathrm{ind}(D_\pm) \in \mathbb{Z}$ need not be zero.

THEOREM 50.24 (Atiyah-Singer index theorem, vague version). *For any odd Dirac operator on a spinor bundle $E \to M$ over a compact Riemannian manifold (M, g), the index $\mathrm{ind}(D_+) \in \mathbb{Z}$ can be expressed precisely in terms of topological invariants of M and E.*

This statement should remind you a little of the Gauss-Bonnet theorem, and indeed, Gauss-Bonnet can be derived from it as a special case via a clever choice of spinor bundle, as can many important topological results about smooth manifolds, such as the Hirzebruch signature theorem. For detailed accounts of the Atiyah-Singer theorem, see for instance [**Roe98**, **BB85**, **Boo77**, **LM89**].

Incidentally, none of this works when (M, g) has indefinite signature: Dirac operators in the indefinite case are *not* elliptic, including the notable example of Dirac's original version of the Dirac equation, which admits wave-like solutions (a property one does not typically expect from elliptic equations). The implications of the indefinite Dirac equation for pure mathematics are correspondingly mild in comparison with the version that was later adapted by mathematicians.

REMARK 50.25. If (M, g) has Lorentz signature $(1, n)$ for some $n \geqslant 1$, then the spinor bundle $E \to M$ is determined by a spin structure and a representation of $\mathrm{Cl}(1, n)$, where the latter means a set of matrices $\gamma_\mu \in \mathbb{C}^{N \times N}$ for $\mu = 0, \ldots, n$ that anticommute with each other and satisfy $\gamma_0^2 = -\mathbb{1}$ and $\gamma_j^2 = \mathbb{1}$ for $j = 1, \ldots, n$. We can easily arrange for these matrices to be unitary, but Lemma 50.21 fails for two reasons to produce from this an honest unitary representation of $\mathrm{Spin}(1, n)$. The first is that while γ_0 is still anti-Hermitian, the negativity of the norm squared in dimensions $1, \ldots, n$ makes $\gamma_1, \ldots, \gamma_n$ Hermitian, and the products $\gamma_0 \gamma_i$ corresponding to generators of $\mathfrak{spin}(1, n)$ are similarly Hermitian, and thus not in $\mathfrak{u}(N)$. The second problem is that $\mathrm{Spin}(1, n)$ is not connected, so proving something about the Lie algebra representation $\sigma_* : \mathfrak{spin}(1, n) \to \mathfrak{sl}(N, \mathbb{C})$ does not automatically imply a corresponding statement about $\sigma : \mathrm{Spin}(1, n) \to \mathrm{SL}_\pm(N, \mathbb{C})$. For the latter issue there is an easy fix that only requires imposing an extremely reasonable extra condition on (M, g): we require namely that the set of time-like vectors $\{X \in TM \mid \langle X, X \rangle > 0\}$ should have two separate connected components, labelled “forward” and “backward”. This is a type of orientation condition that makes it possible to distinguish the past from the future if (M, g) is interpreted as a spacetime manifold. If M is also oriented in the usual sense, so that its structure group is $\mathrm{SO}(1, n)$, then labelling the two directions of time makes it possible to reduce the structure group to its identity component $\mathrm{SO}^+(1, n) \subset \mathrm{SO}(1, n)$, the so-called **proper orthochronous** Lorentz group, consisting of all orientation-preserving Lorentz transformations that also preserve the orientation of time. With this reduction in place, a spin structure on (M, g) lifts its structure group from $\mathrm{SO}^+(1, n)$ to the identity component $\mathrm{Spin}^+(1, n)$ of $\mathrm{Spin}(1, n)$, so that the spinor bundle $E \to M$ now has a connected structure group.

To deal with the fact that $\gamma_1, \ldots, \gamma_n$ are not anti-Hermitian, physicists have a favorite trick: the matrix $i\gamma_0$ is invertible and Hermitian, so we can endow $\mathbb{C}^N$ with a nondegenerate sesquilinear form given by

$$(\xi, \eta) := \langle \xi, i\gamma_0 \eta \rangle.$$

This is not a Hermitian inner product: indeed, the relations $\gamma_0^2 = -\mathbb{1}$ and $\gamma_0\gamma_j = -\gamma_j\gamma_0$ imply that the spectrum of $i\gamma_0$ must always contain a mixture of 1 and -1; for both of the concrete examples introduced in (50.3) and (50.4) with $N = 4$, the eigenvalues include two of each. But if we interpret $(\ , \ )$ as an indefinite complex inner product on $\mathbb{C}^N$, then it is easy to check that all of the γ_μ and $\gamma_\mu\gamma_\nu$ for $\mu \neq \nu$ satisfy the relations

$$(\xi, \gamma_\mu \eta) + (\gamma_\mu \xi, \eta) = 0, \qquad \text{and} \qquad (\xi, \gamma_\mu\gamma_\nu \eta) + (\gamma_\mu\gamma_\nu \xi, \eta) = 0$$

for $\xi, \eta \in \mathbb{C}^N$, and it follows that the action of the connected group $\mathrm{Spin}^+(1, n)$ on $\mathbb{C}^N$ preserves this indefinite inner product, thus giving rise to an indefinite complex bundle metric on $E \to M$ that is compatible with the spin connection. For this bundle metric, the Lorentzian analogue of Theorem 50.23 is true, making the Dirac operator formally self-adjoint with respect to an indefinite analogue of the usual L^2-pairing on $\Gamma(E)$. In physics, the indefinite bundle metric can be used for giving coordinate-invariant definitions of various measurable quantities associated to Dirac spinors, notably a notion of probability density. Physicists prefer to write the indefinite pairing on $\mathbb{C}^N$ in terms of row and column vectors as

$$(\xi, \eta) = \xi^\dagger i\gamma_0 \eta =: \bar{\xi}\eta,$$

where the row vector $\bar{\xi} := \xi^\dagger i\gamma_0$ is called the **Dirac adjoint** of $\xi \in \mathbb{C}^N$. (The factor of i does not appear in physics textbooks, but we need to include it because physicists' version of the standard Clifford algebra relation sets $\gamma_\mu\gamma_\nu + \gamma_\nu\gamma_\mu$ equal to $2\eta_{\mu\nu}$ instead of $-2\eta_{\mu\nu}$.)

50.6. A taste of gauge theory. Some content that was not covered in the lecture will appear here someday.

51. Manifolds with constant curvature

For the next two lectures we'll be studying Riemannian manifolds that have more uniformity or symmetry than in the generic case. Two motivations for this come to mind: one is that examples of such manifolds arise naturally in various contexts and are interesting in their own right. The other concerns the hardest question that one can ask about Riemannian manifolds: can we classify them up to isometry? This is possible in dimension 1, where there is no notion of curvature and thus all Riemannian manifolds are locally isometric, but from dimension two upwards, it would be virtually hopeless to achieve a complete classification. What one can reasonably try instead is to classify all Riemannian manifolds with some special property up to isometry, e.g. the connected and complete Riemannian manifolds with constant sectional curvature. We will see that for this very special class, the classification problem is tractable.

51.1. A criterion for local isometries. Recall that as a corollary of the fundamental theorem that a connection on a vector bundle is flat if and only if its Riemann tensor vanishes (see §46.4), a Riemannian manifold is locally isometric to Euclidean space if and only if its sectional curvature along all 2-dimensional subspaces vanishes. Our first aim is to generalize this to a result about manifolds with constant but possibly nonzero sectional curvature. The question can be posed as follows: given two Riemannian n-manifolds (M, g), (M', g'), points $p \in M$, $p' \in M'$ and an orthogonal linear transformation $\Phi : T_pM \to T_{p'}M'$, do there exist neighborhoods $p \in \mathcal{U} \subset M$ and $p' \in \mathcal{U}' \subset M$ admitting an isometry $\varphi : (\mathcal{U}, g) \to (\mathcal{U}', g')$ such that $\varphi(p) = p'$ and $T_p\varphi = \Phi$?

The answer is clearly no in general, but we'd like to find a minimal extra condition that makes it yes. Observe that since isometries map geodesics to geodesics, an explicit formula for φ can easily be written down if it exists: it must take the form

$$\varphi(\exp_p(X)) = \exp_{p'}(\Phi(X)) \tag{51.1}$$

for all X in some neighborhood $\mathcal{O} \subset T_pM$ of the origin. Let us fix such a neighborhood $\mathcal{O}$ that is small enough for $\mathcal{O} \xrightarrow{\exp_p} M$ to be an embedding onto an open set

$$\mathcal{U} := \exp_p(\mathcal{O}) \subset M,$$

and also such that $\exp_{p'}$ is well defined on

$$\mathcal{O}' := \Phi(\mathcal{O}) \subset T_{p'}M'.$$

Note that we are not assuming $\exp_{p'}$ is an embedding on $\mathcal{O}'$, and there will be situations in which we prefer to do without that assumption, though it is certainly satisfied if $\mathcal{O} \subset T_pM$ is chosen small enough. Under these assumptions, the map $\varphi : \mathcal{U} \to M$ is uniquely defined by (51.1), but it might not be an isometry, nor even a local isometry. In order to state a sufficient condition for this, observe that the orthogonal transformation $\Phi : T_pM \to T_{p'}M'$ extends to a smooth family of orthogonal transformations

$$\Phi_q : T_qM \to T_{\varphi(q)}M', \qquad q \in \mathcal{U},$$

defined for each $q \in \mathcal{U}$ by

$$\Phi_q := P^1_{\gamma'} \circ \Phi \circ (P^1_\gamma)^{-1},$$

where $\gamma : [0, 1] \to M$ is the geodesic segment $\gamma(t) := \exp_p(tX)$ with $X \in \mathcal{O}$ such that $\gamma(1) = q$, $\gamma'(t) := \exp_{p'}(t\Phi(X))$ is its image under φ, and $P^1_\gamma : T_pM \to T_qM$ and $P'^1_\gamma : T_{p'}M' \to T_{\varphi(q)}M'$ are the parallel transport maps along these geodesics using the Levi-Cività connections of (M, g) and (M', g'). That Φ_q is orthogonal follows from the fact that Φ is, together with the compatibility of both Levi-Cività connections with their respective metrics.

THEOREM 51.1 (Cartan criterion). *Suppose the covariant Riemann tensors of (M, g) and (M', g') are related by*

$$\mathrm{Riem}_q(V, X, Y, Z) = \mathrm{Riem}_{\varphi(q)}(\Phi_q(V), \Phi_q(X), \Phi_q(Y), \Phi_q(Z))$$

for all $q \in \mathcal{U}$ and $V, X, Y, Z \in T_qM$. Then the map $\varphi : (\mathcal{U}, g) \to (M', g')$ defined via (51.1) is a local isometry.

REMARK 51.2. Recall that a map $\psi : (M, g) \to (M', g')$ is called a *local isometry* if for every point $p \in M$, there are neighborhoods $\mathcal{U}_p \subset M$ of p and $\mathcal{U}_{\psi(p)} \subset M'$ of $\psi(p)$ such that ψ maps $\mathcal{U}_p$ diffeomorphically onto $\mathcal{U}_{\psi(p)}$ with $\psi^*g' = g$. By the inverse function theorem, this is equivalent to the condition that $T_p\psi : T_pM \to T_{\psi(p)}M'$ is an orthogonal transformation for all $p \in M$. In Theorem 51.1, it is not claimed that the map $\varphi : \mathcal{U} \to M'$ is a diffeomorphism; it might in fact be non-injective, depending on the behavior of $\exp_{p'} : \mathcal{O}' \to M'$, which was not assumed to be an embedding.

PROOF OF THEOREM 51.1. We need to show that for every $q \in \mathcal{U}$, the linear map $T_q\varphi : T_qM \to T_{\varphi(q)}M'$ is orthogonal. We claim that, in fact, this map is Φ_q, which is orthogonal by construction. Let us denote $q = \exp_p(X)$ for $X \in \mathcal{O}$, and $q' := \varphi(q) = \exp_{p'}(\Phi(X))$, and present both of these as the end points of the geodesic segments

$$\gamma(t) := \exp_p(tX), \qquad \gamma'(t) := \varphi \circ \gamma(t) = \exp_{p'}(t\Phi(X)), \qquad t \in [0, 1].$$

By definition, φ satisfies the relation

$$\varphi \circ \exp_p = \exp_{p'} \circ \Phi : \mathcal{O} \to M',$$

so for $Z \in T_X \mathcal{O} = T_p M$, the chain rule implies

$$T_q \varphi \circ T_X(\exp_p) Z = T_{\Phi(X)}(\exp_{p'}) \circ \Phi(Z) \in T_{q'} M'.$$

The derivative of exp in vertical directions was computed in Proposition 36.10: $T_X(\exp_p)Z = \eta(1)$, where $\eta \in \Gamma(\gamma^* TM)$ is the unique Jacobi vector field along γ satisfying $\eta(0) = 0$ and $\nabla_t \eta(0) = Z$. Since $\exp_p$ was assumed to be an embedding on $\mathcal{O}$, the points p and q cannot be conjugate along γ, and it therefore follows from Exercise 36.21 that for any given $Y \in T_q M$, there exists a unique Jacobi vector field η along γ satisfying $\eta(0) = 0$ and $\eta(1) = Y$. Setting $Z := \nabla_t \eta(0)$, we then have $T_X(\exp_p)Z = Y$ and can thus write

$$T_q \varphi(Y) = \eta'(1),$$

where η' is the unique Jacobi vector field along γ' with $\eta'(0) = 0$ and $\nabla_t \eta'(0) = \Phi(Z)$.

Now comes the crucial step: we claim that the two Jacobi vector fields η and η' along γ and γ' respectively are related by

$$\eta'(t) = \Phi_{\gamma(t)} \eta(t),$$

in which case the computation above implies $T_q \varphi(Y) = \Phi_q(Y)$ and thus finishes the proof. To establish the claim, choose a parallel orthonormal frame $e_1(t), \dots, e_n(t) \in T_{\gamma(t)} M$ along γ, and observe that

$$e'_j(t) := \Phi_{\gamma(t)} e_j(t) \in T_{\gamma'(t)} M', \qquad j = 1, \dots, n$$

then likewise defines a parallel orthonormal frame for TM' along γ'. We can now write $\eta(t) = \eta^i(t) e_i(t)$ for unique component functions $\eta^i : [0,1] \to \mathbb{R}$, and since the e_i are parallel, taking the inner product of the Jacobi equation $\nabla_t^2 \eta(t) + R(\eta(t), \dot\gamma(t))\dot\gamma(t) = 0$ with $e_i(t)$ turns it into the system of n equations

$$\ddot\eta^i(t) + \operatorname{Riem}_{\gamma(t)}(e_i(t), \eta(t), \dot\gamma(t), \dot\gamma(t)) = 0, \qquad i = 1, \dots, n.$$

Along γ', we likewise have $\xi(t) := \Phi_{\gamma(t)} \eta(t) = \eta^i(t) e'_i(t)$, and this satisfies the Jacobi equation if and only if

$$\ddot\eta^i(t) + \operatorname{Riem}_{\gamma'(t)}(e'_i(t), \xi(t), \dot\gamma'(t), \dot\gamma'(t)) = 0, \qquad i = 1, \dots, n.$$

But by definition $\gamma'(t) = \varphi(\gamma(t))$, $e'_i(t) = \Phi_{\gamma(t)} e_i(t)$ and $\xi(t) = \Phi_{\gamma(t)} \eta(t)$, and since the velocity of each geodesic is parallel, $\dot\gamma'(t) = \Phi_{\gamma(t)} \dot\gamma(t)$. The stated assumptions on the Riemann tensors thus imply that ξ also satisfies the Jacobi equation, and since $\xi(0) = 0$ and $\nabla_t \xi(0) = \dot\eta^i(0) e'_i(0) = \Phi(\nabla_t \eta(0)) = \Phi(Z)$, we conclude $\xi = \eta'$. $\square$

We will see several applications of Cartan's criterion. The first gives a complete solution to the local isometry problem for manifolds with constant sectional curvature.

THEOREM 51.3. *If (M, g) and (M', g') are two Riemannian manifolds with the same constant sectional curvature, then for every $p \in M$ and $p' \in M'$ and every orthogonal transformation $\Phi : T_p M \to T_{p'} M'$, there exist neighborhoods $p \in \mathcal{U} \subset M$ and $p' \in \mathcal{U}' \subset M'$ such that the map φ in Theorem 51.1 is an isometry*

$$(\mathcal{U}, g) \xrightarrow{\varphi} (\mathcal{U}', g')$$

with $\varphi(p) = p'$ and $T_p \varphi = \Phi$.

PROOF. The result will follow if we can establish the hypothesis on the Riemann tensor in Theorem 51.1. The latter follows in this situation from the fact that the sectional curvature determines the Riemann tensor (Theorem 35.13). Indeed, using the family of orthogonal transformations $\Phi_q : T_q M \to T_{\varphi(q)} M'$ in the setting of Theorem 51.1, define the tensor $\Phi^*\mathrm{Riem} \in \Gamma(T^0_4 \mathcal{U})$ by

$$(\Phi^*\mathrm{Riem})_q(V, X, Y, Z) := \mathrm{Riem}_{\varphi(q)}(\Phi_q(V), \Phi_q(X), \Phi_q(Y), \Phi_q(Z))$$

for $q \in \mathcal{U}$ and $V, X, Y, Z \in T_q M$. Given any two linearly-independent vectors $X, Y \in T_q M$, let $P \subset T_q M$ be the plane spanned by X, Y and $P' := \Phi_q(P) \subset T_{q'} M'$, where $q' := \varphi(q)$. We then have

$$\mathrm{Riem}_q(X, X, Y, Y) = K_S(P) \cdot \left(\langle X, X\rangle\langle Y, Y\rangle - \langle X, Y\rangle^2\right)$$

and

$$\begin{aligned}(\Phi^*\mathrm{Riem})_q(X, X, Y, Y) &= K_S(P') \cdot \left(\langle \Phi_q(X), \Phi_q(X)\rangle\langle \Phi_q(Y), \Phi_q(Y)\rangle - \langle \Phi_q(X), \Phi_q(Y)\rangle^2\right) \\ &= \mathrm{Riem}_q(X, X, Y, Y),\end{aligned}$$

due to the orthogonality of Φ_q and the assumption $K_S(P) = K_S(P')$. Since both Riem and $\Phi^*\mathrm{Riem}$ satisfy the symmetry relations listed in Theorem 35.7, it now follows from the proof of Theorem 35.13 that they are identical on $\mathcal{U}$, so the hypothesis of Theorem 51.1 is satisfied. □

51.2. Locally symmetric spaces. We'll have more to say about constant curvature, but first, here is another local uniformity condition that is less restrictive.

DEFINITION 51.4. A Riemannian manifold (M, g) is **locally symmetric** if every point $p \in M$ has a neighborhood $p \in \mathcal{U} \subset M$ admitting an isometry $(\mathcal{U}, g) \to (\mathcal{U}, g)$ of the form $\exp_p(X) \mapsto \exp_p(-X)$.

In other words, (M, g) is locally symmetric if the natural "antipodal map" about every point is an isometry, at least on a small neighborhood of that point. As usual, the local isometry $\varphi : \mathcal{U} \to \mathcal{U}$ fixing $p \in \mathcal{U}$ is unique if it exists, since its first derivative at p is required to be the antipodal map $T_p\varphi = -\mathbb{1}$. One can extend Definition 51.4 to a global condition and call (M, g) a **Riemannian symmetric space** if for every $p \in M$, there exists a global isometry $\varphi \in \mathrm{Isom}(M, g)$ satisfying $\varphi(p) = p$ and $T_p\varphi = -\mathbb{1}$. This is a much more restrictive condition, and there is a correspondingly large literature on the classification of Riemannian symmetric spaces up to isometry (see e.g. [**Ber03**]). We will not get into that subject here, but merely note one advantage of the less restrictive local condition: it admits an easy characterization in terms of curvature.

THEOREM 51.5. *A Riemannian manifold is locally symmetric if and only if its Riemann tensor is parallel.*

PROOF. We first assume (M, g) is locally symmetric and try to prove $\nabla R \equiv 0$, which is equivalent to the condition $\nabla\mathrm{Riem} \equiv 0$. The covariant derivative of Riem is a tensor field of type $(0, 5)$ and satisfies $\nabla\mathrm{Riem} = \varphi^*(\nabla\mathrm{Riem})$ whenever φ is an isometry. Given $p \in M$, the existence of an isometry φ on a neighborhood of p with $\varphi(p) = p$ and $T_p\varphi = -\mathbb{1}$ thus implies

$$\nabla\mathrm{Riem}(V, W, X, Y, Z) = \nabla\mathrm{Riem}(-V, -W, -X, -Y, -Z) = -\nabla\mathrm{Riem}(V, W, X, Y, Z)$$

for all $V, W, X, Y, Z \in T_p M$, and thus $\nabla\mathrm{Riem} = -\nabla\mathrm{Riem} = 0$.

Conversely, suppose $\nabla\mathrm{Riem} = 0$. Given $p \in M$, we apply Theorem 51.1 in the case $(M', g') := (M, g)$, $p' := p$ and $\Phi := -\mathbb{1} : T_p M \to T_p M$, so that the geodesic segments $\gamma : [0, 1] \to M$ and $\gamma' = \varphi \circ \gamma : [0, 1] \to M$ form two halves of the same geodesic from $q = \gamma(1) = \exp_p(X)$ through p to $q' := \varphi(q) = \gamma'(1) = \exp_p(-X)$. The orthogonal transformation $\Phi_q : T_q M \to T_{q'} M$ is then formed by composing the parallel transport along γ from q to p with the antipodal map on T_pM followed by another parallel transport along γ' from p to q', and in total, this is the same thing as inserting a minus sign in front of the parallel transport map along the geodesic from q to q'. Since Riem is parallel and takes evenly many arguments, it follows that $\mathrm{Riem}_q(V,X,Y,Z) = \mathrm{Riem}_{q'}(\Phi_q(V),\Phi_q(X),\Phi_q(Y),\Phi_q(Z))$ for all $V,X,Y,Z \in T_qM$, so Theorem 51.1 implies that φ is a local isometry. □

51.3. Constant sectional curvature. We are now in a position to attack a global classification question: what are all the isometry classes of complete Riemannian manifolds with constant sectional curvature?

Let us briefly review the three standard examples, which were introduced in §24.4 of last semester's notes.

Example 51.6. For $R > 0$, let $S^n_R \subset \mathbb{R}^{n+1}$ denote the sphere of radius R in $\mathbb{R}^{n+1}$ with the standard Euclidean metric. Its sectional curvature is

$$K_S \equiv \frac{1}{R^2} > 0.$$

To prove this, observe first that K_S must be constant because S^n_R has many isometries: the canonical action of $\mathrm{O}(n+1)$ on Euclidean $(n+1)$-space restricts to an action on S^n_R that can be used to send any 2-dimensional subspace $P \subset TS^n_R$ to any other one. It thus suffices to compute $K_S(P)$ for a particular subspace at a particular point, and this is easy because the geodesics on S^n_R are quite simple: we can find a point $p \in S^n_R$ and 2-dimensional subspace $P \subset T_pS^n_R$ such that the submanifold formed by following geodesics from p in directions tangent to P is the intersection of S^n_R with a 3-dimensional subspace of $\mathbb{R}^{n+1}$. That intersection is a copy of S^2_R, whose Gaussian curvature can be computed via an explicit computation of the Gauss map (see §27.3), and the answer is $1/R^2$.

Example 51.7. Euclidean space $(\mathbb{R}^n, g_{\mathrm{E}})$ has constant sectional curvature $K_S \equiv 0$. I think you can carry out this computation without hints.

Example 51.8. The **hyperbolic n-space** is defined analogously to the sphere, but lives in Minkowski space $(\mathbb{R}^{n+1}, g_M := -(dx^0)^2 + (dx^1)^2 + \ldots + (dx^n)^2)$ instead of Euclidean space. Concretely, for $R > 0$, we consider the Riemannian submanifold $H^n_R \subset (\mathbb{R}^{n+1}, g_M)$ defined by

$$H^n_R := \left\{ \mathbf{x} = (x^0, \ldots, x^n) \in \mathbb{R}^{n+1} \;\middle|\; g_M(\mathbf{x},\mathbf{x}) = -R^2 \text{ and } x^0 > 0 \right\}.$$

Writing $\mathbf{x} = (t, \mathbf{v}) \in \mathbb{R} \times \mathbb{R}^n$, the condition $g_M(\mathbf{x},\mathbf{x}) = -R^2$ cuts out the two-sheeted hyperboloid $t^2 - |\mathbf{v}|^2 = R^2$, and the condition $x^0 = t > 0$ is then imposed in order to single out a connected component, so that H^n_R is diffeomorphic to $\mathbb{R}^n$. The action on $(\mathbb{R}^{n+1}, g_M)$ by the orthochronous Lorentz group $\mathrm{O}^+(1,n)$ then preserves H^n_R and can be used to identify any 2-dimensional subspace $P \subset H^n_R$ with any other, implying that the sectional curvature is constant. As with the sphere, one can then make convenient choices of $p \in H^n_R$ and $P \subset T_pH^n_R$ in order to reduce the computation of $K_S(P)$ to the 2-dimensional case, which follows again by computing a version of the Gauss map for a Riemannian 2-manifold in Minkowski $\mathbb{R}^3$; the answer is

$$K_S \equiv -\frac{1}{R^2} < 0.$$

Theorem 51.9. *Every complete and simply connected Riemannian n-manifold (M,g) with constant sectional curvature is isometric to one of the examples S^n_R, $(\mathbb{R}^n, g_{\mathrm{E}})$ or H^n_R discussed above.*

Proof. Let (N,h) denote the unique one of the examples S^n_R, $(\mathbb{R}^n, g_{\mathrm{E}})$ or H^n_R whose sectional curvature matches that of (M,g). Pick two points $p \in N$ and $p' \in M$, and an orthogonal transformation $\Phi : T_pN \to T_{p'}M$.

The case $K_S \leqslant 0$ is slightly easier, because by the Cartan-Hadamard theorem (Theorem 36.17), we know that $\exp_p : T_pN \to N$ and $\exp_{p'} : T_{p'}M \to M$ are both covering maps, which implies since N and M are both simply connected that both are diffeomorphisms. We therefore obtain a diffeomorphism $\varphi : N \to M$ in the form

$$\varphi = \exp_{p'} \circ \Phi \circ (\exp_p)^{-1} : N \to M,$$

and applying the Cartan criterion as in Theorem 51.3 shows that φ is an isometry $(N, h) \to (M, g)$.

In the case $K_S > 0$ and $(N, h) = S^n_R$, an extra argument is required since $\exp_p : T_pN \to N$ is not a diffeomorphism to the sphere. We know however precisely why $\exp_p$ fails to be a diffeomorphism, since we know the geodesics on S^n_R: they all are periodic with length $2\pi R$, thus $\exp_p$ is an embedding on the open ball $B_{\pi R}(0) \subset T_pN$ of radius πR about the origin, in fact it defines a diffeomorphism

$$T_pN \supset \mathcal{O} := B_{\pi R}(0) \xrightarrow{\exp_p} \mathcal{U} := S^n_R \backslash \{-p\}.$$

Since (M, g) is complete, we also know that $\exp_{p'}$ is well defined on $\mathcal{O}' := \Phi(\mathcal{O})$, though we do not know as yet whether it's an embedding. The Cartan criterion and the argument of Theorem 51.3 using constant sectional curvature in any case give a local isometry

$$\varphi = \exp_{p'} \circ \Phi \circ (\exp_p)^{-1} : S^n_R \backslash \{-p\} \to M.$$

In order to extend this map over $-p$, we can pick a point $q \in S^n_R \backslash \{p, -p\}$ and perform the same trick to construct a local isometry

$$\varphi' : S^n_R \backslash \{-q\} \to M,$$

which can be chosen to have any desired value and orthogonal first derivative at one point, thus we can arrange it to satisfy $\varphi'(q) = \varphi(q)$ and $T_q\varphi' = T_q\varphi$, implying that φ and φ' are identically equal on their common domain $S^n_R \backslash \{-p, -q\}$. The union of their two domains covers S^n_R, so we conclude that both are restrictions of a globally-defined local isometry

$$\psi : S^n_R \to M.$$

By Lemma 36.19, ψ is a covering map, and since M is simply connected, it follows that ψ is a diffeomorphism, and therefore an isometry. □

Recall that by standard covering space theory, every manifold M can be presented as the quotient of its own universal cover $\widetilde{M}$ by the group of deck transformations, a discrete group Γ isomorphic to the fundamental group of M that acts on $\widetilde{M}$ freely and properly discontinuously (cf. Exercise 40.24). When (M, g) is a complete Riemannian manifold, the cover $\widetilde{M}$ inherits from this a complete metric $\widetilde{g}$ for which the covering projection $\widetilde{M} \to M$ is a local isometry, and Γ therefore acts on $\widetilde{M}$ by isometries.

COROLLARY 51.10. *Every complete Riemannian manifold with constant sectional curvature is isometric to $(N, h)/\Gamma$, where (N, h) is one of the examples S^n_R, $(\mathbb{R}^n, g_{\mathrm{E}})$ or H^n_R, and Γ is a countable discrete group that acts on (N, h) freely and properly discontinuously by isometries.* □

In this way, the classification of complete Riemannian manifolds with constant sectional curvature has been reduced to the study of discrete subgroups of the three possible isometry groups $\mathrm{Isom}(S^n_R) = \mathrm{O}(n + 1)$, $\mathrm{Isom}(\mathbb{R}^n, g_{\mathrm{E}})$ and $\mathrm{Isom}(H^n_R) = \mathrm{O}^+(n + 1)$.

52. Manifolds with many isometries

52.1. A hierarchy of symmetry conditions. Instead of imposing conditions directly on curvature, in this lecture we consider Riemannian manifolds whose isometry groups are assumed to be larger than in the generic case. Recall from Lecture 49 that $\mathrm{Isom}(M, g)$ is in general a Lie group that acts smoothly not just on M, but also on the orthonormal frame bundle $F^{\mathrm{O}}(TM)$, and picking any reference frame $\phi \in F^{\mathrm{O}}(TM)$ determines an embedding

$$\mathrm{Isom}(M,g) \hookrightarrow F^{\mathrm{O}}(TM) : \psi \hookrightarrow \psi_*\phi$$

that identifies it with a closed subset and smooth submanifold of $F^{\mathrm{O}}(TM)$. It follows that

$$\dim \mathrm{Isom}(M,g) \leqslant \dim F^{\mathrm{O}}(TM) = \frac{n(n+1)}{2}, \qquad \text{assuming } \dim M = n,$$

and if M is compact, then $\mathrm{Isom}(M,g)$ is also compact. (Note that the latter is false in general if (M,g) has a metric of indefinite signature, since the orthonormal frame bundle in that case is never compact—Exercise 49.27 exhibits an actual counterexample.)

One can also consider the action of $\mathrm{Isom}(M,g)$ on individual tangent vectors of unit length, defined by

$$\mathrm{Isom}(M,g) \times STM \to STM : (\psi, X) \mapsto T\psi(X),$$

where $STM \subset TM$ denotes the unit sphere bundle in TM, also known as the **unit tangent bundle**. All of these actions can be turned into symmetry conditions on a Riemannian manifold (M,g). Let us arrange them in order of increasing strictness.

Definition 52.1. A connected Riemannian n-manifold (M,g) is called

(1) **homogeneous** if the action of $\mathrm{Isom}(M,g)$ on M is transitive;
(2) **homogeneous and isotropic** if the action of $\mathrm{Isom}(M,g)$ on STM is transitive;
(3) **weakly frame homogeneous** or **maximally symmetric** if every orbit of the action of $\mathrm{Isom}(M,g)$ on $F^{\mathrm{O}}(TM)$ is a union of connected components;
(4) **frame homogeneous** if the action of $\mathrm{Isom}(M,g)$ on $F^{\mathrm{O}}(TM)$ is transitive.

What follows is a sequence of observations about the various conditions in Definition 52.1 and the relations between them, leading up to a complete classification of manifolds that satisfy conditions (3) or (4).

Remark 52.2. Every manifold (M,g) satisfying any of the conditions in Definition 52.1 is complete. In particular, homogeneity implies that the injectivity radius is a constant function on M, and since this function is also positive, it follows that there is a constant $\epsilon > 0$ such that every geodesic with unit initial velocity is guaranteed to exist for at least time ϵ. If that is true everywhere, then no geodesic can ever cease to exist in finite time.

Remark 52.3. The results of §41.2 determine the possible diffeomorphism classes of a homogeneous Riemannian manifold: since $\mathrm{Isom}(M,g)$ is a Lie group, transitivity of its action on M implies that there is a diffeomorphism

$$\mathrm{Isom}(M,g)/G_p \xrightarrow{\cong} M : [\psi] \mapsto \psi(p)$$

for any point $p \in M$, where $G_p \subset \mathrm{Isom}(M,g)$ is the stabilizer of p. Every homogeneous Riemannian manifold (M,g) is thus diffeomorphic to a quotient of a Lie group G by a closed subgroup—what we have previously called a *homogeneous space*—and M is compact if and only if G is compact.

Exercise 52.4. Show that connected Riemannian symmetric spaces (see the discussion following Definition 51.4) are homogeneous, though locally symmetric spaces need not be.

Exercise 52.5. Given a point $p \in M$, one calls (M,g) **isotropic at** p if the stabilizer $G_p \subset \mathrm{Isom}(M,g)$ of p under its action on M acts transitively on the unit sphere in T_pM. In other words, this means that any tangent vector at p can be mapped to any other tangent vector of the same length at p by an isometry. Show that if (M,g) is connected and isotropic at every point p, then it is homogeneous and isotropic. (The converse is also trivially true.)

Remark 52.6. Motivated in part by Exercise 52.5, some authors refer to the condition we are calling "homogeneous and isotropic" simply as "isotropic".

Remark 52.7. Since M in Definition 52.1 was assumed connected, its orthonormal frame bundle $F^{\mathrm{O}}(TM)$ has either one or two connected components, the latter if and only if M is orientable. Conditions (3) and (4) in the definition are thus equivalent if M is not orientable, and if M is orientable, then it is frame homogeneous if and only if it is weakly frame homogeneous and also admits an orientation-reversing isometry. We will see below that this distinction is not actually meaningful in practice, because the list of possible Riemannian manifolds satisfying condition (3) is so short that all of them also satisfy the stronger condition (4).

Exercise 52.8. Show that an n-dimensional connected Riemannian manifold (M, g) is weakly frame homogeneous if and only if it is complete and $\dim \mathfrak{isom}(M, g) = \frac{n(n+1)}{2}$.

Remark 52.9. The term "weakly frame homogeneous" is not standard, and has been introduced here only for convenience; we will dispense with it after proving that there are no actual examples satisfying this condition that are not also frame homogeneous (without the "weakly"). The term "maximally symmetric" is used mostly by physicists, and is usually understood to mean that the dimension of the space of Killing vector fields takes its maximum possible value $\frac{n(n+1)}{2}$; recall from Corollary 49.24 that if (M, g) is complete, then every Killing vector field has a global flow, so this assumption means that $\dim \mathrm{Isom}(M, g)$ also takes its largest possible value, and it is equivalent by Exercise 52.8 to what we have called weak frame homogeneity. Beware: you will sometimes also hear physicists claiming that conditions (2) and (3) are equivalent, but Exercise 52.11 and Remark 52.12 below show that this is false.

Proposition 52.10. *If (M, g) is weakly frame homogeneous then it has constant sectional curvature.*

Proof. Given $p, q \in M$ and 2-dimensional subspaces $P \subset T_pM$ and $Q \subset T_qM$, one can choose orthonormal bases $X_1, \ldots, X_n \in T_pM$ and $Y_1, \ldots, Y_n \in T_qM$ such that P is spanned by X_1, X_2 and Q is spanned by Y_1, Y_2. After adjusting the orientations of these bases if necessary, weak frame homogeneity guarantees in light of Remark 52.7 the existence of an isometry $\psi \in \mathrm{Isom}(M, g)$ satisfying $T\psi(X_j) = Y_j$ for $j = 1, \ldots, n$, thus $T\psi(P) = Q$ and therefore $K_S(P) = K_S(Q)$. $\square$

Exercise 52.11. Show that for every $n \in \mathbb{N}$, $\mathbb{CP}^n$ admits a Riemannian metric that is homogeneous and isotropic.
Hint: Arrange it so that the natural action of $\mathrm{U}(n+1)$ on the unit sphere in $\mathbb{C}^{n+1}$ descends to the quotient $S^{2n+1}/S^1 = \mathbb{CP}^n$ as an action by isometries.

Remark 52.12. You may recall from the atlas constructed for $\mathbb{CP}^n$ in Exercise 32.11 that $\mathbb{CP}^n$ contains submanifolds diffeomorphic to $\mathbb{CP}^{n-1}$ such that $\mathbb{C}^n \backslash \mathbb{CP}^{n-1}$ is diffeomorphic to $\mathbb{C}^n$. One can use this decomposition of $\mathbb{CP}^n$ to prove via the Seifert-van Kampen theorem that it is simply connected for all $n \in \mathbb{N}$; alternatively, general perturbation and transversality results in differential topology (cf. [**Hir94**]) imply that any continuous loop in $\mathbb{CP}^n$ admits a small perturbation that does not intersect the submanifold $\mathbb{CP}^{n-1} \subset \mathbb{CP}^n$ and is thus contained in $\mathbb{C}^n$, implying that it is contractible. It follows via Theorem 51.9 that if $\mathbb{CP}^n$ admits a complete metric with constant sectional curvature, then it must be diffeomorphic to either S^{2n} or $\mathbb{R}^{2n}$. For $n = 1$ this is true, but not for $n \geqslant 2$, which shows that from dimension four upwards, there exist Riemannian manifolds that are homogeneous and isotropic but not weakly frame homogeneous. For $\mathbb{CP}^n$ there is a clearly discernible reason why this can happen: isotropy means that the tangent spaces of $\mathbb{CP}^n$ do not have any preferred directions, but they do have preferred 2-dimensional subspaces, because $\mathbb{CP}^n$ is naturally a complex manifold, and the isometries of the metric constructed in Exercise 52.11

are holomorphic. This means in particular that the actions of these isometries on tangent spaces always send complex subspaces to complex subspaces, and one finds if one computes the sectional curvature of $\mathbb{CP}^n$ with such a metric that it has different values along complex lines than it does along other real 2-dimensional subspaces.

EXAMPLE 52.13. The examples $(\mathbb{R}^n, g_{\mathrm{E}})$, S^n_R and H^n_R for $R > 0$, which are complete and simply connected with constant sectional curvature, are all frame homogeneous. This is easy to check because their isometry groups can be described explicitly: $\mathrm{Isom}(\mathbb{R}^n, g_{\mathrm{E}})$ is the semidirect product of $\mathrm{O}(n)$ and $\mathbb{R}^n$, consisting of all affine transformations of the form $\mathbf{x} \mapsto \mathbf{A}\mathbf{x} + \mathbf{b}$ with $\mathbf{A} \in \mathrm{O}(n)$ and $\mathbf{b} \in \mathbb{R}^n$, while

$$\mathrm{Isom}(S^n_R) \cong \mathrm{O}(n+1), \qquad H^n_R \cong \mathrm{O}^+(1, n),$$

acting by isometries on Euclidean and Minkowski $\mathbb{R}^{n+1}$ respectively.

The conditions in Definition 52.1 come up often in cosmology, which attempts to identify the global structure of the universe throughout time, understood (at least in classical general relativity without quantum mechanics) as a pseudo-Riemannian 4-manifold with Lorentz signature. One step in this effort is to produce global models of the 3-dimensional universe at a "fixed" time, which would appear as Riemannian hypersurfaces in the 4-dimensional spacetime manifold. Following observational evidence that, on the cosmological scale, there do not appear to be any preferred locations, preferred directions or preferred reference frames in the universe, cosmologists typically postulate that the global structure of 3-dimensional space should be described by a Riemannian manifold that is either homogeneous and isotropic or maximally symmetric (the physicists' term for weakly frame homogeneous). Unfortunately, physicists have a tendency to learn differential geometry from other physicists, who do not always realize that they have not thought through all the details: as a result, certain false claims get repeated frequently without question, and one of them—especially common on internet forums frequented by physics graduate students—is that every homogeneous and isotropic Riemannian manifold is also maximally symmetric. We saw above that $\mathbb{CP}^n$ for $n \geqslant 2$ provides a counterexample to this claim, but I can conjecture an explanation for physicists' failure to notice that it is false: the case cosmologists actually care about is dimension three, and in that case it's true!

THEOREM 52.14. *A connected Riemannian manifold of dimension $n \leqslant 3$ is homogeneous and isotropic if and only if it is weakly frame homogeneous.*

PROOF. The case $n = 1$ is trivial, and the proof for $n = 2$ is just an easier variation on the case $n = 3$, so let's assume $n = 3$. The main idea is to exploit the fact from Exercise 41.8 that the only Lie subgroup of $\mathrm{SO}(3)$ acting transitively on S^2 is $\mathrm{SO}(3)$ itself. This extends easily to the statement that if $H \subset \mathrm{O}(3)$ is a closed subgroup acting transitively on S^2, then H is either $\mathrm{SO}(3)$ or $\mathrm{O}(3)$. The easiest way to prove both of these statements is by looking at the Lie algebra $\mathfrak{so}(3)$, whose nontrivial Lie subalgebras other than $\mathfrak{so}(3)$ itself are all 1-dimensional. The connected subgroups generated by these Lie subalgebras are all families of rotations about a fixed axis, and are thus isomorphic to $\mathrm{SO}(2) \cong S^1$. If $H \subset \mathrm{O}(3)$ is a closed subgroup other than $\mathrm{SO}(3)$ or $\mathrm{O}(3)$, it follows that H is either discrete or is diffeomorphic to a countable disjoint union of circles. Since $\mathrm{O}(3)$ is compact, H is therefore either finite or a finite disjoint union of circles, and in either case, the map $H \to S^2 : h \mapsto hp$ defined by choosing any point $p \in S^2$ certainly cannot be surjective.

Now, if (M, g) is 3-dimensional and isotropic at some point $p \in M$, consider the stabilizer subgroup $G_p \subset \mathrm{Isom}(M, g)$ for p. As a closed subgroup of $\mathrm{Isom}(M, g)$, G_p is a Lie group, and it is also compact since for any fixed frame $\phi \in F^{\mathrm{O}}(T_pM)$, the map $G_p \to F^{\mathrm{O}}(T_pM) : \psi \mapsto \psi_*\phi$ identifies it with a closed subset of the compact manifold $F^{\mathrm{O}}(T_pM) \cong \mathrm{O}(3)$. If we pick an orthonormal basis so as to identify T_pM with $\mathbb{R}^3$, the action of G_p on $T_pM \cong \mathbb{R}^3$ defines a Lie group homomorphism

$$\Phi : G_p \to \mathrm{O}(3),$$

which is injective since any isometry that fixes both p and the directions of all tangent vectors at p must be the identity. It follows that Φ identifies G_p with a Lie subgroup of $\mathrm{O}(3)$, which we shall continue to denote by G_p. Isotropy at p implies that this subgroup acts on the unit sphere $S^2 \subset \mathbb{R}^3$ transitively, so it follows that G_p is either $\mathrm{O}(3)$ or $\mathrm{SO}(3)$, and in particular, it acts transitively either on $F^{\mathrm{O}}(T_pM)$ or on each of its two connected components.

Finally, if (M, g) is both homogeneous and isotropic, it now follows that the orbit of any given frame $\phi \in F^{\mathrm{O}}(T_pM)$ under $\mathrm{Isom}(M, g)$ contains either one or both connected components of $F^{\mathrm{O}}(T_qM)$ for every point $q \in M$, which proves that (M, g) is weakly frame homogeneous. $\square$

Next, we show that the list of weakly frame homogeneous manifolds in Example 52.13 is almost already complete.

THEOREM 52.15. *Every weakly frame homogeneous Riemannian n-manifold is isometric to either $(\mathbb{R}^n, g_{\mathrm{E}})$, the sphere S^n_R, hyperbolic space H^n_R or real projective space*

$$\mathbb{RP}^n_R := S^n_R / \mathbb{Z}_2$$

for some $R > 0$, where $\mathbb{RP}^n_R$ carries the metric it inherits as a quotient of S^n_R by a group of isometries.

PROOF. If (M, g) is weakly frame homogeneous, then it is complete by Remark 52.2 and has constant sectional curvature by Proposition 52.10, so by Corollary 51.10, it is isometric to

$$(\widetilde{M}, \widetilde{g})/\Gamma$$

where $(\widetilde{M}, \widetilde{g})$ is either $(\mathbb{R}^n, g_{\mathrm{E}})$, S^n_R or H^n_R for some $R > 0$, and $\Gamma \subset \mathrm{Isom}(\widetilde{M}, \widetilde{g})$ is a discrete subgroup acting freely and properly on $\widetilde{M}$. By Exercise 52.8, $\dim \mathfrak{isom}(M, g) = \frac{n(n+1)}{2}$, and we observe: if $\pi : \widetilde{M} \to M$ denotes the quotient projection, then every nontrivial Killing vector field $X \in \mathfrak{isom}(M, g)$ gives rise to a nontrivial vector field

$$\widetilde{X} := X \circ \pi \in \mathfrak{X}(\widetilde{M})$$

that also satisfies the Killing equation since $\pi : (\widetilde{M}, \widetilde{g}) \to (M, g)$ is a local isometry, thus defining an injective map $\pi^* : \mathfrak{isom}(M, g) \hookrightarrow \mathfrak{isom}(\widetilde{M}, \widetilde{g})$. Since $\dim \mathfrak{isom}(\widetilde{M}, \widetilde{g})$ is also $\frac{n(n+1)}{2}$ and, in particular, cannot be larger, this proves that every Killing vector field on $(\widetilde{M}, \widetilde{g})$ is of the form $X \circ \pi$ for some $X \in \mathfrak{isom}(M, g)$, and is therefore invariant under the action of the discrete group Γ. It follows that the flows generated by these vector fields likewise commute with the isometries in Γ, and since these flows generate the identity component $\mathrm{Isom}_0(\widetilde{M}, \widetilde{g}) \subset \mathrm{Isom}(\widetilde{M}, \widetilde{g})$ of the isometry group, we conclude

$$\psi \circ \varphi = \varphi \circ \psi \qquad \text{for all } \psi \circ \Gamma \text{ and } \varphi \in \mathrm{Isom}_0(\widetilde{M}, \widetilde{g}).$$

Now consider the three cases separately:

- If (M, g) has positive curvature, then $(\widetilde{M}, \widetilde{g})$ is S^n_R for some $R > 0$, and $\mathrm{Isom}_0(\widetilde{M}, \widetilde{g})$ is therefore $\mathrm{SO}(n+1)$. The only nontrivial subgroup of $\mathrm{Isom}(\widetilde{M}, \widetilde{g}) = \mathrm{O}(n+1)$ that commutes with everything in $\mathrm{SO}(n+1)$ is $\mathbb{Z}_2 \cong \{\mathbb{1}, -\mathbb{1}\} \subset \mathrm{O}(n+1)$, so Γ is either that or the trivial group, giving rise to the two possibilities $\mathbb{RP}^n_R$ or S^n_R for (M, g).

- If (M, g) has zero curvature, then $(\widetilde{M}, \widetilde{g})$ is $(\mathbb{R}^n, g_{\mathrm{E}})$ and $\mathrm{Isom}_0(\widetilde{M}, \widetilde{g})$ is thus generated by the rotations $\mathrm{SO}(n)$ and translations of $\mathbb{R}^n$, while $\mathrm{Isom}(\widetilde{M}, \widetilde{g})$ is generated by the translations and $\mathrm{O}(n)$. In this case, no nontrivial element of $\mathrm{Isom}(\widetilde{M}, \widetilde{g})$ commutes with all of $\mathrm{Isom}_0(\widetilde{M}, \widetilde{g})$, so Γ must be the trivial group and thus $(M, g) \cong (\mathbb{R}^n, g_{\mathrm{E}})$.
- If (M, g) has negative curvature, then $(\widetilde{M}, \widetilde{g})$ is H^n_R for some $R > 0$, $\mathrm{Isom}(\widetilde{M}, \widetilde{g})$ is the orthochronous Lorentz group $\mathrm{O}^+(1, n)$ and $\mathrm{Isom}_0(\widetilde{M}, \widetilde{g}) = \mathrm{SO}^+(1, n)$. The requirement that elements of $\mathrm{O}^+(1, n)$ preserve the direction of time (and thus preserve H^n_R as a component of a two-sheeted hyperboloid) prevents it from containing $-\mathbb{1}$, so in this case again, $\mathrm{O}^+(1, n)$ contains no nontrivial elements that commute with all of $\mathrm{SO}^+(1, n)$, leading to the conclusion $\Gamma = \{\mathbb{1}\}$ and $(M, g) \cong H^n_R$.

□

52.2. Einstein metrics. In the effort to impose conditions that make classification up to isometry a tractable problem, it is natural to ask: of all the metrics one could define on a given manifold, does there exist one that is the "best"? Clearly the exact meaning of the word "best" is open to debate, but for instance, in dimension two, a reasonable goal is to look for metrics that have constant Gaussian curvature, and it turns out that they always exist, and are even unique if one fixes the conformal class of the metric in advance. The obvious generalization of this condition to higher dimensions would be constant sectional curvature, but we have already seen that that condition is too strong: most higher-dimensional manifolds have universal covers that are not diffeomorphic to S^n or $\mathbb{R}^n$, and thus do not have any metrics of constant sectional curvature. The following is a more reasonable condition to impose in higher dimensions.

Definition 52.16. A Riemannian metric g on an n-manifold M is called an **Einstein metric** if its Ricci tensor $\mathrm{Ric} \in \Gamma(T^0_2 M)$ is a scalar multiple of g at every point, i.e. $\mathrm{Ric} = f \cdot g$ for some function $f : M \to \mathbb{R}$.

One of the motivations for studying Einstein metrics comes from general relativity, because in the Lorentzian setting, the condition in Definition 52.16 is closely related to the *Einstein equation*, which determines the time-evolution of the Lorentzian metric (interpreted as a gravitational field) on a spacetime manifold containing matter. A purely Riemannian motivation to study Einstein metrics comes from the question of finding a "best" metric. One way to approach this question is by maximizing or minimizing certain real-valued functionals that can be defined on the set of Riemannian metrics, and one of the simplest is the so-called **Einstein-Hilbert action**, defined as an integral of the scalar curvature. If we assume M is compact and write the scalar curvature and Riemannian volume form as Scal_g and $d\mathrm{vol}_g$ respectively in order to emphasize their dependence on the metric g, the Einstein-Hilbert action is given by

$$F(g) := \int_M \mathrm{Scal}_g \, d\mathrm{vol}_g \in \mathbb{R}.$$

To define an interesting variational problem for this functional, we fix a constant $C > 0$ and take the domain of F to be the set of all metrics that have volume C,

$$\mathcal{G} := \left\{ \text{Riemannian metrics } g \;\middle|\; \int_M d\mathrm{vol}_g = C \right\}.$$

The proof of the following proposition is a somewhat messy but standard computation in the calculus of variations; we are only including the statement here for the sake of context, so we will omit the proof.

PROPOSITION 52.17. *A Riemannian metric $g \in \mathcal{G}$ of volume C on a compact manifold M is an Einstein metric if and only if it is a critical point of the Einstein-Hilbert action on $\mathcal{G}$, i.e. for every smooth 1-parameter family $\{g_t \in \mathcal{G}\}_{t \in (-\epsilon,\epsilon)}$ with $g_0 = g$,*

$$\left.\frac{d}{dt} F(g_t)\right|_{t=0} = 0.$$

□

Here are a few easy observations about Einstein metrics.

REMARK 52.18. Recall from Lecture 36 that the scalar curvature $\mathrm{Scal} : M \to \mathbb{R}$ is defined at each point as the trace of the mixed version $\mathrm{Ric}^\sharp \in \Gamma(T^1_1 M)$ of the Ricci tensor, defined via the condition

$$\mathrm{Ric}(X,Y) = g(X, \mathrm{Ric}^\sharp(Y))$$

for all $X, Y \in \mathfrak{X}(M)$. The Einstein condition $\mathrm{Ric} = f \cdot g$ is equivalent to

$$\mathrm{Ric}^\sharp = f \cdot \mathbb{1} \in \Gamma(T^1_1 M),$$

so taking the trace of both sides at each point gives a formula for $f : M \to \mathbb{R}$ in terms of the scalar curvature, namely $f = \frac{\mathrm{Scal}}{n}$. The condition for g to be an Einstein metric can thus be rephrased as

$$\mathrm{Ric} = \frac{\mathrm{Scal}}{n} g.$$

EXERCISE 52.19. Show that if (M, g) is homogeneous and isotropic, then g is an Einstein metric.
Hint: At each point $p \in M$, the linear map $R^\sharp_p : T_pM \to T_pM$ is symmetric with respect to the inner product g_p. What can you say about the eigenspaces of $R^\sharp_p$ and the action of the stabilizer $G_p := \{\psi \in \mathrm{Isom}(M,g) \mid \varphi(p) = p\}$ on T_pM?

REMARK 52.20. In the case $\dim M = 2$, Exercise 36.8 gives the relation $\mathrm{Ric} = K_G \cdot g$, implying that *every* Riemannian metric in dimension 2 is Einstein. You can see an explanation for this if you look again at the Einstein-Hilbert action and remember the Gauss-Bonnet formula, which implies in light of the relation $\mathrm{Scal} = 2K_G$ that

$$F(g) = \int_M \mathrm{Scal}_g \, d\mathrm{vol}_g = 2 \int_M K_G \, d\mathrm{vol} = 4\pi\chi(M)$$

for every closed oriented Riemannian 2-manifold (M, g). In other words, the Einstein-Hilbert action is independent of the metric in dimension 2, implying that every metric is a critical point, and therefore an Einstein metric.

In dimensions three and upward, the study of Einstein metrics is a major industry within Riemannian geometry. We do not have space to say anything about it here beyond the most fundamental observation that, in contrast to the 2-dimensional case, most metrics on higher-dimensional manifolds are indeed *not* Einstein metrics, because most of them do not have constant scalar curvature:

THEOREM 52.21. *If $\dim M \geqslant 3$, then every Einstein metric on M has constant scalar curvature.*

The main tool we need for the proof of this theorem is a "contracted" version of the second Bianchi identity. Recall from Lecture 45: the second Bianchi identity is the relation

$$dF_A = [F_A, A]$$

for the equivariant curvature 2-form $F_A \in \Omega^2(E, \mathfrak{g})$ of a principal connection 1-form $A \in \Omega^1(E, \mathfrak{g})$ on a principal G-bundle $E \to M$. It is equivalent to the statement $d_A F_A = 0$ for the operator $d_A := d + A \wedge (\cdot)$ on the space of Ad-equivariant forms $\Omega^*_{\mathrm{Ad}}(E, \mathfrak{g})$, so under the natural isomorphism between $\Omega^*_{\mathrm{Ad}}(E, \mathfrak{g})$ and $\Omega^*(M, \mathrm{Ad}(E))$, it translates into the identity

$$d_\nabla \Omega_A = 0$$

for the bundle-valued curvature 2-form $\Omega_A \in \Omega^2(M, \mathrm{Ad}(E))$, where d_∇ is the covariant exterior derivative operator on $\Omega^*(M, \mathrm{Ad}(E))$ induced by the principal connection on E. To bring this into the context of Riemannian geometry, we recall from §46.4 that if E is the frame bundle $F(TM)$ and A defines the principal connection on $F(TM)$ corresponding to some chosen affine connection ∇ on M, then the Riemann tensor $R \in \Gamma(T^1_3 M)$ of ∇ can be derived from the bundle-valued curvature 2-form $\Omega_A \in \Omega^2(M, \mathrm{Ad}(E))$ by writing

$$R(\cdot,\cdot)V = \Omega_A \wedge V \in \Omega^2(M, TM) \tag{52.1}$$

for any $V \in \mathfrak{X}(M)$. Here E is a principal $\mathrm{GL}(n,\mathbb{R})$-bundle and TM is identified with the associated bundle $(E \times \mathbb{R}^n)/\,\mathrm{GL}(n,\mathbb{R})$; the adjoint bundle $\mathrm{Ad}(E)$ meanwhile has standard fiber $\mathfrak{gl}(n,\mathbb{R}) = \mathbb{R}^{n\times n}$, whose natural action on $\mathbb{R}^n$ is used to define a parallel bundle map $\mathrm{Ad}(E) \otimes TM \to TM$ and thus a wedge product pairing of $\Omega^*(M, \mathrm{Ad}(E))$ with $\Omega^*(M, TM)$. Applying d_∇ to the right hand side of (52.1) produces $\Omega_A \wedge \nabla V \in \Omega^3(M, TM)$ as a result of the Leibniz rule and the second Bianchi identity. If one now assumes ∇ is symmetric and uses (45.3) to compute the covariant exterior derivative of the left hand side, one ends up with the relation

$$(\nabla_X R)(Y, Z) + (\nabla_Y R)(Z, X) + (\nabla_Z R)(X, Y) = 0 \qquad \text{for all } X, Y, Z \in \mathfrak{X}(M), \tag{52.2}$$

which is known as the Riemannian variant of the **second Bianchi identity**, also sometimes called the **differential Bianchi identity**. To clarify the notation: for $X \in \mathfrak{X}(M)$ we are viewing $\nabla_X R \in \Gamma(T^1_3 R)$ as a multilinear bundle map $TM \oplus TM \oplus TM \to TM : (Y, Z, V) \mapsto (\nabla_X R)(Y, Z)V$, just as we do with the Riemann tensor itself. Notice that the differential Bianchi identity is not just a result about Riemannian or pseudo-Riemannian manifolds—the formula is valid for the Riemann tensor of any symmetric affine connection.

EXERCISE 52.22. Work out the details of the proof of the differential Bianchi identity (52.2). *Advice: When ω is a bundle-valued 2-form, the formula (45.3) for $d_\nabla \omega$ can be written in the form*

$$d_\nabla \omega(X, Y, Z) = \nabla_X \big(\omega(Y, Z)\big) - \omega\big([X, Y], Z\big) + \text{cyclic},$$

where the word "cyclic" means that additional (in this case four) terms appear, obtained from the written terms via all possible cyclic permutations of the triple (X, Y, Z). Similarly, for a bilinear bundle map $\mu : E_1 \oplus E_2 \to F$ and forms $\omega \in \Omega^2(M, E_1), \lambda \in \Omega^1(M, E_2)$, the formula (45.1) for $\mu(\omega, \lambda) \in \Omega^3(M, F)$ becomes

$$\mu(\omega, \lambda)(X, Y, Z) = \mu(\omega(X, Y), \lambda(Z)) + \text{cyclic}.$$

One last piece of advice: when Lie brackets appear, use the torsion tensor to get rid of them.

THEOREM 52.23 (twice contracted Bianchi identity). *On any pseudo-Riemannian manifold (M, g) with Levi-Cività connection ∇, the mixed Ricci tensor $\mathrm{Ric}^\sharp \in \Gamma(T^1_1 M)$ and scalar curvature $\mathrm{Scal} : M \to \mathbb{R}$ are related by the formula*

$$d(\mathrm{Scal})(X) = 2\,\mathrm{tr}\big(Y \mapsto \nabla_Y \mathrm{Ric}^\sharp(X)\big),$$

or in abbreviated form,

$$d(\mathrm{Scal}) = 2\,\mathrm{tr}(\mathrm{Ric}^\sharp).$$

Proof. In local coordinates, the stated relation says

$$\partial_j \mathrm{Scal} = 2\nabla_i R^i{}_j, \tag{52.3}$$

where the components of the tensors $R \in \Gamma(T^1_3 M)$, $\mathrm{Ric} \in \Gamma(T^0_2 M)$, $\mathrm{Ric}^\sharp \in \Gamma(T^1_1 M)$ and $\nabla \mathrm{Ric}^\sharp \in \Gamma(T^1_2 M)$ are written as

$$R^i{}_{jk\ell} := dx^i\big(R(\partial_j, \partial_k)\partial_\ell\big), \qquad R_{ij} := \mathrm{Ric}(\partial_i, \partial_j),$$
$$R^i{}_j := dx^i\big(\mathrm{Ric}^\sharp(\partial_j)\big), \qquad \nabla_i R^j{}_k := dx^j\big((\nabla_{\partial_i}\mathrm{Ric}^\sharp)(\partial_k)\big),$$

and by definition,

$$R_{ij} = R^k{}_{kij}, \quad R^i{}_j = g^{ik} R_{kj}, \quad R_{ij} = g_{ik} R^k{}_j, \quad \text{and} \quad \mathrm{Scal} = R^i{}_i.$$

Since $d(\mathrm{Scal})$ and $\mathrm{tr}(\mathrm{Ric}^\sharp)$ are both well-defined tensor fields, it will suffice to prove that (52.3) holds in some particular choice of local coordinates, which we may as well take to be Riemann normal coordinates about some point $p \in M$. At that one point, the Christoffel symbols vanish, so the operator ∇_i matches the ordinary partial derivative operator ∂_i and (52.3) thus becomes

$$\partial_j R^i{}_i = 2\partial_i R^i{}_j. \tag{52.4}$$

We will perform the entire calculation at this one point, and also use the fact that for this choice of coordinates, the first partial derivatives of g_{ij} and g^{ij} all vanish. The differential Bianchi identity in these coordinates says

$$\partial_i R^\ell{}_{jkm} + \partial_j R^\ell{}_{kim} + \partial_k R^\ell{}_{ijm} = 0$$

at p, so unpacking $\partial_j R^i{}_i$ at this point in terms of the Riemann tensor gives

$$\partial_j R^i{}_i = \partial_j\left(g^{ik} R^\ell{}_{\ell ik}\right) = g^{ik}\partial_j R^\ell{}_{\ell ik} = -g^{ik}\partial_\ell R^\ell{}_{ijk} - g^{ik}\partial_i R^\ell{}_{j\ell k}. \tag{52.5}$$

The last expression can be simplified using the symmetries of the Riemann tensor if we first rewrite it in terms of the covariant Riemann tensor with components $R_{ijk\ell} = \mathrm{Riem}(\partial_i, \partial_j, \partial_k, \partial_\ell) = g_{im} R^m{}_{jk\ell}$, hence $R^i{}_{jk\ell} = g^{im} R_{mjk\ell}$. Since the derivatives of g^{im} vanish, (52.5) becomes

$$\begin{aligned}
\partial_j R^i{}_i &= -g^{ik} g^{\ell m}\partial_\ell R_{mijk} - g^{ik} g^{\ell m}\partial_i R_{mj\ell k} \\
&= g^{ik} g^{\ell m}\partial_\ell R_{kijm} + g^{ik} g^{\ell m}\partial_i R_{m\ell jk} \\
&= g^{\ell m}\partial_\ell R^i{}_{ijm} + g^{ik}\partial_i R^\ell{}_{\ell jk} = \partial_\ell\big(g^{\ell m} R_{jm}\big) + \partial_i\big(g^{ik} R_{jk}\big) \\
&= \partial_\ell R^\ell{}_j + \partial_i R^i{}_j = 2\partial_i R^i{}_j,
\end{aligned} \tag{52.6}$$

where in the second line we've used the relations $R_{ijk\ell} = -R_{\ell jki} = -R_{ikj\ell}$, and in the last line we also used the symmetry of the Ricci tensor $R_{ij} = R_{ji}$. □

Proof of Theorem 52.21. Abbreviate $S := \mathrm{Scal}$. If g is an Einstein metric, then by Remark 52.18, $\mathrm{Ric}^\sharp = \frac{1}{n} S \cdot \mathbb{1}$ for $n := \dim M$. You will find it easy to check that the tensor field $\mathbb{1} \in \Gamma(T^1_1 M)$ is parallel, so applying the covariant derivative to both sides of this relation and applying the Leibniz rule on the right hand side gives

$$\nabla \mathrm{Ric}^\sharp = \frac{1}{n} dS \cdot \mathbb{1} \in \Gamma(T^1_2 M),$$

which in local coordinates says

$$\nabla_k R^i{}_j = \frac{1}{n}\partial_k S \cdot \delta^i_j.$$

Contracting the indices k and i in this expression gives

$$\nabla_i R^i{}_j = \frac{1}{n}\partial_i S \cdot \delta^i_j = \frac{1}{n}\partial_j S,$$

but according to the contracted Bianchi identity (52.3), the latter is also $\frac{2}{n}\nabla_i R^i{}_j$, giving

$$\nabla_i R^i{}_j = \frac{2}{n}\nabla_i R^i{}_j.$$

If $n \geqslant 3$, the only way for this equality to hold is if $\nabla_i R^i{}_j$ vanishes identically, which implies the same for $\partial_j S$. $\square$

Exercise 52.24. Suppose (M, g) is a connected Riemannian manifold of dimension $n \geqslant 3$ and $f : M \to \mathbb{R}$ is a smooth function such that the sectional curvature satisfies $K_S(P) = f(p)$ for all $P \subset T_pM$, $p \in M$. Prove that K_S is then constant. (Is this true for $n = 2$?)
Hint: Prove that g is an Einstein metric. You might find Equation 36.1 helpful.

53. Introduction to Hodge theory

Our goal for the last four lectures in this course is to prove the Hodge decomposition theorem, which identifies the de Rham cohomology $H^*_{\mathrm{dR}}(M)$ of a closed n-manifold with a special finite-dimensional subspace of $\Omega^*(M)$, the space of so-called *harmonic* forms. Applications of this theorem include the Poincaré duality isomorphism $H^k_{\mathrm{dR}}(M) \cong H^{n-k}_{\mathrm{dR}}(M)$ for closed oriented manifolds, plus a multitude of results in which—much like the Gauss-Bonnet formula—the properties of a chosen Riemannian on M and its curvature constrain the topology of M, or vice versa.

For real-valued functions f of n variables, the word "harmonic" refers to solutions of the **Laplace equation**

$$\sum_{j=1}^{n} \partial_j^2 f = 0, \tag{53.1}$$

which is usually regarded as the simplest interesting partial differential equation. You have likely seen it before, either in complex analysis (because the real and imaginary parts of holomorphic functions on $\mathbb{C} = \mathbb{R}^2$ are harmonic), or in physics (the electrostatic potential in a vacuum is harmonic, and the Laplace operator also shows up naturally as a component in standard wave equations, including the Schrödinger equation of quantum mechanics). In the context of a smooth manifold, we will have to work a bit before deciding how to define harmonic functions $f : M \to \mathbb{R}$; a first guess might be to require the equation (53.1) to be satisfied in local coordinates, but this definition would then depend on the choice of coordinates, so we need something better. Once a coordinate-invariant version of the Laplace equation for differential forms has been defined, we will see that the deep results of Hodge theory arise mainly from a particular property that this equation has: it is *elliptic*. Linear PDEs of the form $Df = 0$ for an elliptic operator D have a number of wondrous properties when considered specifically on *closed* manifolds: in particular, their solution spaces are finite dimensional, and the corresponding inhomogeneous equation $Df = g$ can also be solved for all g on a space of finite codimension. The most difficult part of the proof of the Hodge decomposition theorem will follow from these general properties of elliptic operators, so for most of the next few lectures, we will consider arbitrary elliptic operators rather than the Laplace operator specifically. We will need to use some tools from functional analysis: notably the Fourier transform and the basic properties of Sobolev spaces, some of which we will state as black boxes since the proofs would require too much of a digression. Similarly, giving complete proofs of the fundamental results on elliptic regularity would require at least a few extra lectures dominated by long and intricate strings of inequalities, which would change this geometry course into something of an altogether different character, so I will skip some of those details. Wherever possible, I will endeavor at least to make each stated result seem believable and communicate the basic principles behind it.

More comprehensive proofs of the Hodge decomposition theorem can be found in various standard textbooks such as [War83, dR84, Jos17]. Some of those treatments make the task slightly simpler by focusing specifically on the Laplace operator instead of general elliptic operators. In my opinion, however, sacrificing generality in this way does not make things *that* much easier, and at the same time, it obscures the interesting role played by more general phenomena that are relevant both here and in other subbranches of differential geometry where elliptic operators arise. I prefer where possible to state each result at its natural level of generality, especially when doing so helps elucidate the essential reasons why it is true. A good source for more detailed proofs of the results we will need on general elliptic operators is [Ebe].

53.1. Harmonic functions on a Riemannian manifold. Our first task is to write down a meaningful version of the Laplace equation (53.1) for functions $f : M \to \mathbb{R}$ on an n-manifold M so that the equation does not depend on any choice of coordinates. We dealt with a similar challenge when we discussed the Dirac equation in Lecture 50, and the solution here is similar: we must first endow M with more structure. For the Dirac equation, the structure we needed included an orientation, a pseudo-Riemannian metric and a spin structure on top of these, plus a choice of representation for the relevant Clifford algebra in order to define the spinor bundle on which the Dirac operator acts. For the Laplacian, a metric will suffice: we assume $(M, g = \langle\ ,\ \rangle)$ is a pseudo-Riemannian manifold, and we will later want to insist that g is also positive, but let's hold off on that assumption until it's really needed. For convenience we shall also usually assume that M is orientable, though this assumption is not essential. The metric and a choice of orientation determine a volume form $d\mathrm{vol} \in \Omega^n(M)$ and a Levi-Cività connection ∇, and the **divergence** of a vector field $X \in \mathfrak{X}(M)$ is then the unique function $\operatorname{div}(X) : M \to \mathbb{R}$ satisfying

$$\mathcal{L}_X(d\mathrm{vol}) = \operatorname{div}(X) \cdot d\mathrm{vol}, \qquad \text{or equivalently} \qquad \operatorname{div}(X) = \operatorname{tr}(\nabla X).$$

The equivalence of these two ways of defining $\operatorname{div}(X)$ was proved in §50.5.2. Recall that the **gradient** of a function $f : M \to \mathbb{R}$ is obtained by plugging the differential $df \in \Omega^1(M)$ into the musical isomorphism $\sharp : T^*M \to TM$ that inverts $\flat : TM \to T^*M : X \mapsto X_\flat := \langle X, \cdot\rangle$, so

$$\nabla f := (df)^\sharp \in \mathfrak{X}(M), \qquad \text{meaning} \qquad df = \langle \nabla f, \cdot \rangle.$$

The following now defines a natural generalization of the coordinate-based Laplace operator to pseudo-Riemannian manifolds:

$$\Delta f := -\operatorname{div}(\nabla f) = -\operatorname{tr}(\nabla^2 f),$$

where we write $\nabla^2 f := \nabla(\nabla f) \in \Omega^1(M, TM) = \Gamma(\operatorname{End}(TM))$.[111] Indeed, one checks easily that if $M = \mathbb{R}^n$ is endowed with a flat metric $g = g_{ij}\, dx^i\, dx^j$ with constant components so that ∇ is the trivial connection, then Δ is given by

$$\Delta f = -g^{ij} \partial_i \partial_j f, \tag{53.2}$$

which in the case of the standard Euclidean metric $g_{ij} = \delta_{ij}$ is just $-\sum_{j=1}^n \partial_j^2$. Not every author includes the minus sign, but there are good reasons to do so and they will be clarified below (see Remark 53.3). We call a function $f : M \to \mathbb{R}$ **harmonic** if it satisfies $\Delta f = 0$.

EXERCISE 53.1. Show that the divergence operator $\operatorname{div} : \mathfrak{X}(M) \to C^\infty(M)$ satisfies the Leibniz rule

$$\operatorname{div}(\varphi X) = \varphi \cdot \operatorname{div}(X) + d\varphi(X) \qquad \text{for all } \varphi \in C^\infty(M),\ X \in \mathfrak{X}(M).$$

[111] The reader should be aware that the symbol "∇^2" is also the standard physicists' notation for the Laplace operator, but mathematicians typically use it with a different meaning.

Exercise 53.1 gives rise to an integration-by-parts formula if we combine it with the observation via Cartan's magic formula that $\operatorname{div}(X)\cdot d\mathrm{vol} = \mathcal{L}_X(d\mathrm{vol}) = d(\iota_X d\mathrm{vol})$ is an exact n-form for every $X \in \mathfrak{X}(M)$. Indeed, we have

$$\int_M \varphi \cdot \operatorname{div}(X)\, d\mathrm{vol} = -\int_M d\varphi(X)\, d\mathrm{vol} \tag{53.3}$$

for all $\varphi \in C^\infty(M)$ and $X \in \mathfrak{X}(M)$ with compact support in $M\backslash\partial M$. This follows because the difference between the two sides of the equation is $\int_M \operatorname{div}(\varphi X)\, d\mathrm{vol} = \int_M d(\iota_{\varphi X} d\mathrm{vol})$, which vanishes by Stokes' theorem if $\iota_{\varphi X} d\mathrm{vol}$ has compact support in $M\backslash\partial M$. For the Laplace operator, this gives rise to the relation

$$\int_M \varphi\, \Delta f\, d\mathrm{vol} = -\int_M \varphi\, \operatorname{div}(\nabla f)\, d\mathrm{vol} = \int_M d\varphi(\nabla f)\, d\mathrm{vol} = \int_M \langle \nabla\varphi, \nabla f\rangle\, d\mathrm{vol} \tag{53.4}$$

for any two smooth functions $\varphi, f : M \to \mathbb{R}$ with compact support in $M\backslash\partial M$.

PROPOSITION 53.2. *On any connected and oriented Riemannian manifold (M, g), all solutions $f \in C^\infty(M)$ to the equation $\Delta f = 0$ with compact support in $M\backslash\partial M$ are constant. In particular, all solutions are constant if M is closed.*

PROOF. Assuming $\Delta f = 0$ and $\operatorname{supp}(f) \subset M\backslash\partial M$ is compact, (53.4) gives

$$0 = \int_M f\, \Delta f\, d\mathrm{vol} = \int_M \langle \nabla f, \nabla f\rangle\, d\mathrm{vol},$$

which is only possible if $\nabla f \equiv 0$. □

REMARK 53.3. The main property of Δ we used in the proof of Proposition 53.2 is that for a closed Riemannian manifold (M, g), the quadratic form defined on $C^\infty(M)$ by

$$Q(f) := \int_M f\, \Delta f\, d\mathrm{vol}$$

is nonnegative, and vanishes only on constant functions. This is a reason to include the minus sign in the definition of $\Delta = -\operatorname{tr}(\nabla^2)$; maybe it's discriminatory, but most people prefer quadratic forms to be nonnegative rather than nonpositive.

Let us take note of which hypotheses in Proposition 53.2 can or cannot be relaxed.

REMARK 53.4. If $f : M \to \mathbb{R}$ does not have compact support or is allowed to be nonzero on ∂M, then the use of (53.4) becomes invalid, because either $\int_M d(\iota_{\varphi X} d\mathrm{vol}) = \int_{\partial M} \iota_{\varphi X} d\mathrm{vol}$ can be nonzero or Stokes' theorem cannot be applied due to noncompact supports. It is easy to find examples of open Riemannian manifolds that admit infinite-dimensional spaces of nonconstant harmonic functions that either have noncompact support or are nonzero on the boundary, e.g. on the compact 2-disk $\mathbb{D} \subset \mathbb{C}$ or the complex plane $\mathbb{C} = \mathbb{R}^2$ with the standard Euclidean metric, the real part of every holomorphic function is harmonic.

REMARK 53.5. If (M, g) is a pseudo-Riemannian manifold with indefinite signature, then the computation in the proof of Proposition 53.2 is valid, but the conclusion is not: the relation $\int_M \langle \nabla f, \nabla f\rangle\, d\mathrm{vol} = 0$ does not imply that ∇f vanishes, since $\langle \nabla f, \nabla f\rangle$ can take any sign at each point. For example, on $\mathbb{T}^{n+1}$ with periodic coordinates $(t, \mathbf{x}) := (x^0, x^1, \ldots, x^n) \in S^1 \times \mathbb{T}^n$ and the Minkowski metric $g = dt^2 - (dx^1)^2 - \ldots - (dx^n)^2$, (53.2) becomes the wave operator

$$\Delta f = -\partial_t^2 f + \sum_{j=1}^n \partial_j^2 f.$$

Physicists prefer to the equation $\Delta f = 0$ in this case as $\Box f = 0$ in order to emphasize that it has a quite different character from the usual Laplace equation, e.g. it has an infinite-dimensional space of nonconstant solutions $f : \mathbb{T}^{n+1} \to \mathbb{R}$, containing the traveling waves

$$f(t, \mathbf{x}) = \cos\big(2\pi(\omega t - \langle \mathbf{k}, \mathbf{x} \rangle)\big), \qquad \text{for every } \mathbf{k} \in \mathbb{Z}^n \text{ and } \omega := \pm|\mathbf{k}|.$$

REMARK 53.6. Our proof of Proposition 53.2 used the orientation of M since it involved a global integral and a volume form $d\mathrm{vol}$, but the result is true without assuming orientability. Indeed, the operator $\Delta = -\operatorname{tr}(\nabla^2)$ depends on the metric g but not on the orientation; even the defining relation $\mathcal{L}_X(d\mathrm{vol}) = \operatorname{div}(X) \cdot d\mathrm{vol}$ for the divergence does not really depend on the orientation since replacing $d\mathrm{vol}$ with $-d\mathrm{vol}$ inserts a sign on both sides and thus leaves $\operatorname{div}(X)$ unchanged. One way to prove Proposition 53.2 for M non-orientable is to use the fact that M then has an orientable connected 2-fold covering space $\pi : \widetilde{M} \to M$; indeed, $\widetilde{M}$ in this situation can be defined as the subset of $\Lambda^n T^*M$ consisting of all alternating n-forms that evaluate to ± 1 on orthonormal bases. Pulling back g defines a metric $\tilde{g}$ such that the covering map $(\widetilde{M}, \tilde{g}) \to (M, g)$ is a local isometry, and it follows that $f : M \to \mathbb{R}$ is harmonic if and only if $f \circ \pi : \widetilde{M} \to \mathbb{R}$ is harmonic. The oriented case of Proposition 53.2 then implies that $f \circ \pi$ is constant, and therefore so is f. Alternatively, the proof of Proposition 53.2 still works as written in the non-orientable case if one interprets $d\mathrm{vol}$ as the canonical volume *element* determined by g instead of a volume form (see §11.4 in last semester's notes). The crucial ingredient needed is the formula $\int_M \operatorname{div}(X)\, d\mathrm{vol} = 0$ for all $X \in \mathfrak{X}(M)$ with compact support, and this can be derived from the oriented case using the covering trick described above.

To summarize: if we want the space of solutions to the equation $\Delta f = 0$ to be finite dimensional, then we need to work on a manifold that is both *closed* and *Riemannian*, i.e. no pseudo.[112] For convenience we will continue to assume that M is also oriented, because it will often be useful to have globally-defined volume forms, but in fact, it is also possible to develop all of Hodge theory without this assumption; the details are carried out in [**dR84**].

The following consequence of Proposition 53.2 hints at the right direction in which it should be generalized:

COROLLARY 53.7. *For any closed oriented Riemannian manifold (M, g), the kernel of the operator $\Delta : C^\infty(M) \to C^\infty(M)$ is identical to the kernel of the operator $d : C^\infty(M) \to \Omega^1(M)$, and there is thus a natural isomorphism* $\ker \Delta \to H^0_{\mathrm{dR}}(M) : f \mapsto [f]$. $\square$

53.2. Harmonic differential forms. In order to write down a Laplace-type operator on differential forms, we need a few algebraic preliminaries.

Suppose V is a real n-dimensional vector space equipped with a nondegenerate symmetric bilinear form $\langle\ ,\ \rangle$. In the cases we mainly care about, $\langle\ ,\ \rangle$ will be a positive inner product, but we will not assume positivity until it is really needed. As usual with scalar products that are allowed to be indefinite, a basis $e_1, \ldots, e_n \in V$ is called **orthonormal** if $\langle e_i, e_j \rangle = \pm\delta_{ij}$.

LEMMA 53.8. *The pairing $\langle\ ,\ \rangle$ on V determines a unique nondegenerate symmetric bilinear form $\langle\ ,\ \rangle$ on the exterior algebra $\Lambda^* V$ with the following properties:*

(1) $\langle a, b \rangle = ab$ for $a, b \in \Lambda^0 V = \mathbb{R}$;
(2) $\langle\ ,\ \rangle$ matches the given bilinear form on $V \subset \Lambda^1 V$;
(3) $\Lambda^k V$ and $\Lambda^\ell V$ are orthogonal subspaces of $\Lambda^* V$ whenever $k \neq \ell$;

[112]This is not to say that there exists no reasonable theory for the Laplace equation on compact manifolds with boundary, or noncompact manifolds—such theories exist, but they require imposing suitable boundary or asymptotic conditions on the solutions. We will not consider such generalizations here, but you will find some discussion of them e.g. in [**dR84**].

(4) *For any orthonormal basis* $e_1, \dots, e_n \in V$ *and any two sets* $1 \leqslant i_1 < \dots < i_k \leqslant n$ *and* $1 \leqslant j_1 < \dots < j_k \leqslant n$,

$$\langle e_{i_1} \wedge \dots \wedge e_{i_k}, e_{j_1} \wedge \dots \wedge e_{j_k} \rangle = \langle e_{i_1}, e_{j_1} \rangle \cdot \ldots \cdot \langle e_{i_k}, e_{j_k} \rangle.$$

In particular, the set

$$\left\{ e_{i_1} \wedge \dots \wedge e_{i_k} \in \Lambda^* V \;\middle|\; k \geqslant 0,\ 1 \leqslant i_1 < \dots < i_k \leqslant n \right\}$$

(including $1 \in \mathbb{R} = \Lambda^0 V$ *for the case* $k = 0$*) is then an orthonormal basis of* $\Lambda^* V$*, and* $\langle\ ,\ \rangle$ *is a positive inner product on* $\Lambda^* V$ *if its restriction to* V *is positive.*

Proof. Once $\langle\ ,\ \rangle$ has been chosen on V, the stated conditions clearly determine its extension to $\Lambda^* V$ uniquely since it is determined on all elements of a basis. To prove that such an extension exists independently of the choice of basis, it will suffice to define it on $\Lambda^k V$ for each $k = 0, \dots, n$ and then require $\Lambda^k V \perp \Lambda^\ell V$ whenever $k \neq \ell$. To define it on $\Lambda^k V$, we start by defining the pairing $\langle\ ,\ \rangle_\otimes$ on $V^{\otimes k}$ by

$$\langle v_1 \otimes \dots \otimes v_k, w_1 \otimes \dots \otimes w_k \rangle_\otimes := \langle v_1, w_1 \rangle \cdot \ldots \cdot \langle v_k, w_k \rangle.$$

You can see that this is well defined if you regard it as a $2k$-fold multilinear map $V \times \dots \times V \to \mathbb{R}$, which therefore corresponds to a unique linear map $V^{\otimes 2k} = V^{\otimes k} \otimes V^{\otimes k} \to \mathbb{R}$ and thus a bilinear form $V^{\otimes k} \times V^{\otimes k} \to \mathbb{R}$; it is manifestly also symmetric. The restriction of this bilinear form to $\Lambda^k V \subset V^{\otimes k}$ is not quite what we are looking for, but almost: we claim that the desired bilinear form on $\Lambda^k V$ is given by

$$\langle\ ,\ \rangle := \frac{1}{k!} \langle\ ,\ \rangle_\otimes \qquad \text{on } \Lambda^k V.$$

To see this, one can choose an orthonormal basis $e_1, \dots, e_n \in V$ and use the formula

$$v_1 \wedge \dots \wedge v_k = \sum_{\sigma \in S_k} (-1)^{|\sigma|} v_{\sigma(1)} \otimes \dots \otimes v_{\sigma(k)}$$

to compute $\langle e_{i_1} \wedge \dots \wedge e_{i_k}, e_{j_1} \wedge \dots \wedge e_{j_k} \rangle$ for any $1 \leqslant i_1 < \dots < i_k \leqslant n$ and $1 \leqslant j_1 < \dots < j_k \leqslant n$. The product

$$\begin{aligned}
\langle e_{i_1} \wedge \dots \wedge e_{i_k}, e_{j_1} \wedge \dots \wedge e_{j_k} \rangle_\otimes &= \sum_{\sigma, \sigma' \in S_k} (-1)^{|\sigma| + |\sigma'|} \langle e_{i_{\sigma(1)}} \otimes \dots \otimes e_{i_{\sigma(k)}}, e_{j_{\sigma'(1)}} \otimes \dots \otimes e_{j_{\sigma'(k)}} \rangle_\otimes \\
&= \sum_{\sigma, \sigma' \in S_k} (-1)^{|\sigma| + |\sigma'|} \langle e_{i_{\sigma(1)}}, e_{j_{\sigma'(1)}} \rangle \cdot \ldots \cdot \langle e_{i_{\sigma(k)}}, e_{j_{\sigma'(k)}} \rangle
\end{aligned}$$

will vanish unless $\{i_1, \dots, i_k\}$ and $\{j_1, \dots, j_k\}$ are the same set, and in the latter case, nontrivial contributions only come from summands in which the two permutations σ, σ' are the same. The sign $(-1)^{|\sigma| + |\sigma'|}$ is then 1 and the order of factors in the product does not matter, so the result is $k! \langle e_{i_1}, e_{i_1} \rangle \cdot \ldots \cdot \langle e_{i_k}, e_{i_k} \rangle$. □

The lemma allows us to assign to every pseudo-Riemannian manifold (M, g) a canonical choice of bundle metric $\langle\ ,\ \rangle$ on $\Lambda^* T^* M$ whose restriction to $T^* M$ corresponds to g under the musical isomorphism $T^* M \cong TM$. This bundle metric on $\Lambda^* T^* M$ will then be positive if and only if g is positive, and at each point, the two possible choices of volume form $d\mathrm{vol}$ that evaluate to ± 1 on orthonormal bases are then determined by the condition

$$\langle d\mathrm{vol}, d\mathrm{vol} \rangle = \pm 1.$$

Here the sign is negative if (M, g) has signature (k, ℓ) with ℓ odd, and is otherwise positive.

LEMMA 53.9. *Given a pairing* $\langle\ ,\ \rangle$ *on* $\Lambda^* V$ *as in Lemma 53.8 and a choice of top-dimensional form* $\mu \in \Lambda^n V$ *with* $\langle \mu, \mu \rangle = \pm 1$, *there exists a unique linear isomorphism* $* : \Lambda^* V \to \Lambda^* V$ *such that*

$$\langle \alpha, \beta \rangle \mu = \alpha \wedge *\beta \qquad \textit{for all } \alpha, \beta \in \Lambda^* V.$$

Moreover, for each $k = 0, \ldots, n$, $*$ *maps* $\Lambda^k V$ *to* $\Lambda^{n-k} V$ *as an orthogonal transformation*

$$(\Lambda^k V, \langle\ ,\ \rangle) \xrightarrow{*} (\Lambda^{n-k} V, \pm\langle\ ,\ \rangle),$$

where the sign $\pm$ *on the right hand side is* $\langle \mu, \mu \rangle$.

PROOF. Fix an orthonormal basis $e_1, \ldots, e_n \in V$ such that $\mu = e_1 \wedge \ldots \wedge e_n$. For each $1 \leqslant i_1 < \ldots < i_k \leqslant n$, writing $\beta := e_{i_1} \wedge \ldots \wedge e_{i_k}$, there is now a unique way to define $*\beta \in \Lambda^{n-k} V$ such that the condition $\langle \alpha, \beta \rangle e_1 \wedge \ldots \wedge e_n = \alpha \wedge *\beta$ is satisfied for all $\alpha \in \Lambda^k V$. Indeed, considering α of the form $e_{j_1} \wedge \ldots \wedge e_{j_k}$ for $1 \leqslant j_1 < \ldots < j_k \leqslant n$ reveals that $*\beta$ must be

$$*(e_{i_1} \wedge \ldots \wedge e_{i_k}) = (-1)^{|\sigma|} e_{j_1} \wedge \ldots \wedge e_{j_{n-k}} \quad \text{where} \quad \sigma : (1, \ldots, n) \mapsto (i_1, \ldots, i_k, j_1, \ldots, j_{n-k}),$$

i.e. the indices $j_1, \ldots, j_{n-k}$ are chosen to include all elements of $\{1, \ldots, n\}$ that are not in $\{i_1, \ldots, i_k\}$, and the permutation $\sigma \in S_n$ is then uniquely determined. This establishes the relation $\langle \alpha, \beta \rangle \mu = \alpha \wedge *\beta$ whenever α and β are wedge products of basis elements, and by multilinearity, it is therefore always satisfied. Since this condition uniquely determines $*$, we conclude that the definition is independent of the choice of basis. The orthogonality of $*$ follows from the fact that it evidently maps an orthonormal basis to an orthonormal basis, with the following caveat: for $\beta = e_{i_1} \wedge \ldots \wedge e_{i_k}$, the signs of $\langle \beta, \beta \rangle$ and $\langle *\beta, *\beta \rangle$ will match if and only if both $\{i_1, \ldots, i_k\}$ and its complement in $\{1, \ldots, n\}$ contain either an even or an odd number of indices j such that $\langle e_j, e_j \rangle = -1$. This holds if and only if the total number of indices in $\{1, \ldots, n\}$ with this property is even, which is equivalent to $\langle \mu, \mu \rangle$ being $+1$ instead of -1. □

For the rest of this section, assume (M, g) is a pseudo-Riemannian manifold with

$$\dim M = n.$$

DEFINITION 53.10. If (M, g) has an orientation and $d\mathrm{vol} \in \Omega^n(M)$ denotes the resulting volume form, the **Hodge star operator** is the smooth linear bundle isomorphism $* : \Lambda^* T^* M \to \Lambda^* T^* M$ that is (in light of Lemma 53.9) uniquely determined by the relation

$$\langle \alpha, \beta \rangle \, d\mathrm{vol} = \alpha \wedge *\beta,$$

where $\langle\ ,\ \rangle$ is the bundle metric on $\Lambda^* T^* M$ determined by g via Lemma 53.8.

The Hodge star operator defines for each $k = 0, \ldots, n$ a bundle isomorphism $\Lambda^k T^* M \to \Lambda^{n-k} T^* M$, and therefore also an isomorphism $\Omega^k(M) \to \Omega^{n-k}(M)$. Formally, there is no problem in saying this also for $k < 0$ or $k > n$, since in these cases both $\Lambda^k T^* M$ and $\Lambda^{n-k} T^* M$ are trivial—the former by definition, the latter because all alternating m-forms on a vector space of dimension less than m vanish.

EXERCISE 53.11. Show that for each $k \in \mathbb{Z}$, the Hodge star $* : \Lambda^k T^* M \to \Lambda^{n-k} T^* M$ satisfies $*^2 = (-1)^{k(n-k)}$, and thus $*^{-1} = (-1)^{k(n-k)} *$.

DEFINITION 53.12. For each $k \in \mathbb{Z}$, the operator $d^* : \Omega^k(M) \to \Omega^{k-1}(M)$ is defined by

$$d^* \omega = (-1)^k *^{-1} \big(d(*\omega) \big).$$

REMARK 53.13. Exercise 53.11 gives rise to the more direct (but harder to remember) formula

$$d^* = (-1)^{k + (k-1)(n-k+1)} * d * = (-1)^{n(k+1)-1} * d * : \Omega^k(M) \to \Omega^{k-1}(M).$$

For $k \leqslant 0$ or $k > n$, d^* should be understood as the trivial operator.

Proposition 53.14. *On any pseudo-Riemannian manifold (M, g), the operator $d^* : \Omega^*(M) \to \Omega^*(M)$ and the exterior derivative $d : \Omega^*(M) \to \Omega^*(M)$ are dual to each other in the sense that*

$$\int_M \langle \beta, d\alpha \rangle \, d\mathrm{vol} = \int_M \langle d^*\beta, \alpha \rangle \, d\mathrm{vol}$$

for all forms $\alpha, \beta \in \Omega^*(M)$ with compact support in $M \backslash \partial M$.

Proof. Both sides of the stated relation are trivial unless $\alpha \in \Omega^{k-1}(M)$ and $\beta \in \Omega^k(M)$ for some $k \in \{1, \ldots, n\}$, so assume this. By Stokes' theorem and the graded Leibniz rule for the wedge product, we have

$$\begin{aligned}
0 &= \int_M d(\alpha \wedge *\beta) = \int_M d\alpha \wedge *\beta + (-1)^{k-1} \int_M \alpha \wedge d(*\beta) \\
&= \int_M d\alpha \wedge *\beta - (-1)^k \int_M \alpha \wedge *\left(*^{-1} d(*\beta)\right) = \int_M d\alpha \wedge *\beta - \int_M \alpha \wedge *(d^*\beta) \\
&= \int_M \langle d\alpha, \beta \rangle \, d\mathrm{vol} - \int_M \langle \alpha, d^*\beta \rangle \, d\mathrm{vol}.
\end{aligned}$$

□

Remark 53.15. We did not include the assumption that M is orientable in the statement of Proposition 53.14, though the proof used that assumption in several essential ways, e.g. by using Stokes' theorem, and by referring to a globally defined volume form $d\mathrm{vol} \in \Omega^n(M)$. Our definition of the Hodge star operator also requires a volume form and thus an orientation. However, reversing the volume form changes $*$ by a sign, and since the star appears twice in the definition of $d^* : \Omega^k(M) \to \Omega^{k-1}(M)$, the latter also makes sense without an orientation. With this in mind, Proposition 53.14 also becomes true if $d\mathrm{vol}$ is understood as the canonical volume element on (M, g) instead of an n-form. One can deduce it from the oriented case using the double covering trick described in Remark 53.6. It is also possible to define the Hodge star operator in a more general way that works when M is not orientable, though in that case, it does not define a map $\Omega^k(M) \to \Omega^{n-k}(M)$, but something a bit more abstract instead. This approach is taken in [dR84].

Motivated by Proposition 53.14, we refer to $d^* : \Omega^*(M) \to \Omega^*(M)$ as the **formal adjoint** of the exterior derivative operator d. We can now define the operator that is the central object of study in Hodge theory.

Definition 53.16. On a pseudo-Riemannian manifold (M, g), the **Laplace-Beltrami operator** (also sometimes called the **Hodge Laplacian** or **Hodge-de Rham operator**) is defined by

$$\Delta := dd^* + d^*d : \Omega^*(M) \to \Omega^*(M).$$

Differential k-forms $\omega \in \Omega^k(M)$ satisfying $\Delta\omega = 0$ are called **harmonic k-forms**.

Note that since d and d^* increase and decrease respectively the degrees of the forms that they act on, Δ sends $\Omega^k(M)$ to itself for each $k \in \mathbb{Z}$. The next exercise is an easy calculation from the definitions; you just have to get the signs right.

Exercise 53.17. Show that on any oriented pseudo-Riemannian manifold (M, g), the Laplace-Beltrami and Hodge star operators commute:

$$\Delta(*\omega) = *(\Delta\omega) \qquad \text{for all } \omega \in \Omega^*(M).$$

Let us try to clarify why Δ is regarded as a type of *Laplace* operator on forms. One reason is that in the case $k = 0$, it agrees with the operator we considered in §53.1. To see this, suppose $\varphi, f \in C^\infty(M)$ are arbitrary functions with compact support in $M \backslash \partial M$, and note that $\Delta = d^*d$

on $\Omega^0(M) = C^\infty(M)$ since d^* vanishes on 0-forms by definition. Using Proposition 53.14, we then find

$$\int_M \varphi\, \Delta f \, d\mathrm{vol} = \int_M \langle \varphi, d^* df \rangle \, d\mathrm{vol} = \int_M \langle d\varphi, df \rangle \, d\mathrm{vol} = \int_M \langle \nabla\varphi, \nabla f \rangle \, d\mathrm{vol},$$

the same relation that was established for the operator $\Delta f := -\operatorname{div}(\nabla f)$ in (53.4). Since the function φ in this calculation was arbitrary, it follows that these two definitions of the operator $\Delta : C^\infty(M) \to C^\infty(M)$ are the same.

On k-forms with $k \geqslant 1$, some justification for applying the word "Laplacian" to Δ comes from writing it down in local coordinates $(x^1, \ldots, x^n)$: if $\omega \in \Omega^k(M)$ is written locally in terms of its components $\omega_{i_1 \ldots i_k}(\partial_{i_1}, \ldots, \partial_{i_k})$, one finds that the corresponding components of $\Delta\omega \in \Omega^k(M)$ are given by

$$(\Delta\omega)_{i_1 \ldots i_k} = -g^{j\ell} \partial_j \partial_\ell \omega_{i_1 \ldots i_k} + \text{terms of order} < 2, \tag{53.5}$$

where to clarify, the extra terms not appearing in this formula depend only on the values and first derivatives of the components of ω, but not on their second derivatives. The proof of this local formula will be an easy exercise once we have learned how to compute the principal symbol of Δ, which will do in the next lecture.

As a preview of what is to come, here is a statement of one of the main corollaries of the Hodge decomposition theorem, whose proof will occupy the next few lectures.

THEOREM 53.18. *On any closed Riemannian manifold (M, g), the space* $\ker \Delta \subset \Omega^*(M)$ *of harmonic forms is finite dimensional, all harmonic forms are closed, and the resulting map*

$$\ker \Delta \to H^*_{\mathrm{dR}}(M) : \omega \mapsto [\omega]$$

is an isomorphism.

If you are not already aware of de Rham's theorem (identifying $H^*_{\mathrm{dR}}(M)$ with the singular cohomology of M with real coefficients), or otherwise have no little or no knowledge of algebraic topology, you can now nonetheless deduce at least two results of fundamental importance about the topology of smooth manifolds:

COROLLARY 53.19. *For any closed manifold M, the de Rham cohomology $H^*_{\mathrm{dR}}(M)$ is finite dimensional.* □

COROLLARY 53.20 (Poincaré duality). *For any closed and oriented n-manifold M and each $k = 0, \ldots, n$, the map*

$$H^{n-k}_{\mathrm{dR}}(M) \xrightarrow{\mathrm{PD}} \left(H^k_{\mathrm{dR}}(M)\right)^*, \qquad \mathrm{PD}([\alpha])[\beta] := \int_M \alpha \wedge \beta$$

is an isomorphism, so in particular, $\dim H^k_{\mathrm{dR}}(M) = \dim H^{n-k}_{\mathrm{dR}}(M)$.

PROOF. Observe first that the map in question is well defined as a consequence of Stokes' theorem: if α and β are both closed, then by the graded Leibniz rule, $\alpha \wedge \beta$ will be exact whenever either α or β is exact, implying that the integral depends only on the cohomology classes of both. In other words, there is a well-defined bilinear pairing

$$H^k_{\mathrm{dR}}(M) \times H^{n-k}_{\mathrm{dR}}(M) \to \mathbb{R} : ([\alpha], [\beta]) \mapsto \int_M \alpha \wedge \beta,$$

which is antisymmetric if k and $n-k$ are both odd, and is otherwise symmetric. The main thing to prove is that this pairing is nondegenerate, i.e. for every $[\alpha] \neq 0 \in H^k_{\mathrm{dR}}(M)$, one can find some $[\beta] \in H^{n-k}_{\mathrm{dR}}(M)$ such that $\int_M \alpha \wedge \beta \neq 0$. This will imply for the following reasons that PD is an isomorphism: first, it is clearly injective, since for any $[\alpha] \neq 0 \in H^{n-k}_{\mathrm{dR}}(M)$, nondegeneracy provides a $[\beta] \in H^k_{\mathrm{dR}}(M)$ such that $\mathrm{PD}([\alpha])[\beta] \neq 0$. This implies $\dim H^{n-k}_{\mathrm{dR}}(M) \leqslant \dim H^k_{\mathrm{dR}}(M)$, but one can make the same argument with the roles of k and $n-k$ reversed in order to prove $\dim H^k_{\mathrm{dR}}(M) \leqslant \dim H^{n-k}_{\mathrm{dR}}(M)$, so their dimensions match, and PD is therefore an isomorphism.

The proof of nondegeneracy is where we can make use of Theorem 53.18. First choose a Riemannian metric g on M, which always exists by the usual partition of unity argument. This determines a positive bundle metric $\langle\ ,\ \rangle$ on Λ^*T^*M, along with a volume form $d\mathrm{vol} \in \Omega^n(M)$, Hodge star operator $* : \Lambda^k T^*M \to \Lambda^{n-k}T^*M$ and Laplace-Beltrami operator $\Delta : \Omega^*(M) \to \Omega^*(M)$. Given $[\alpha] \neq 0 \in H^k_{\mathrm{dR}}(M)$, we can now choose the representative $\alpha \in \Omega^k(M)$ to be harmonic, and it is necessarily nonzero. By Exercise 53.17, $*\alpha \in \Omega^{n-k}(M)$ is then also harmonic and therefore closed, and we have $\int_M \alpha \wedge *\alpha = \int_M \langle \alpha, \alpha \rangle\, d\mathrm{vol} > 0$ since $\alpha \neq 0$ and the bundle metric on Λ^*T^*M is positive. □

Exercise 53.21. Deduce from Corollary 53.20 that on any closed, connected and oriented n-manifold M, an n-form $\omega \in \Omega^n(M)$ is exact if and only if $\int_M \omega = 0$.

Remark 53.22. It is true for many of the results in our discussion of Hodge theory that our proofs make use of orientations for the sake of convenience, even though the results are true without orientability—Poincarè duality, however, is emphatically not an example of this. Our proof of Corollary 53.20 required the use of the Hodge star operator as an isomorphism from $\Omega^k(M)$ to $\Omega^{n-k}(M)$, as well as the ability to define a pairing of $H^k_{\mathrm{dR}}(M)$ with $H^{n-k}_{\mathrm{dR}}(M)$ by integrating n-forms over M, neither of which makes sense if M is not endowed with an orientation. And indeed, Corollary 53.20 is false for non-orientable manifolds; for example, if M is closed and connected but non-orientable, then $H^n_{\mathrm{dR}}(M) = 0 \not\cong H^0_{\mathrm{dR}}(M) = \mathbb{R}$. This is a standard fact from algebraic topology if you are willing to follow the de Rham isomorphism from $H^*_{\mathrm{dR}}(M)$ to singular cohomology, but one can also deduce it from Exercise 53.21 using the double covering trick of Remark 53.6. Let $\pi : \widetilde{M} \to M$ denote a double cover such that $\widetilde{M}$ is closed, connected and orientable, in which case the non-orientability of M implies that the unique nontrivial deck transformation $\psi : \widetilde{M} \to \widetilde{M}$ for this cover is orientation reversing. Given $\omega \in \Omega^n(M)$, the relation $\pi \circ \psi = \pi$ then implies $\psi^*\pi^*\omega = \pi^*\omega$ and thus

$$\int_{\widetilde{M}} \pi^*\omega = \int_{\widetilde{M}} \psi^*\pi^*\omega = -\int_{\psi(\widetilde{M})} \pi^*\omega = -\int_{\widetilde{M}} \pi^*\omega,$$

so that $\int_{\widetilde{M}} \pi^*\omega$ necessarily vanishes. (One can also see this by considering an arbitrary evenly covered subset $\mathcal{U} \subset M$: the preimage $\pi^{-1}(\mathcal{U}) \subset \widetilde{M}$ is then a disjoint union of two regions on which $\pi^*\omega$ looks identical but $\widetilde{M}$ carries opposite orientations, so they cancel each other in the integral.) It now follows from Exercise 53.21 that $\pi^*\omega = d\widetilde{\lambda}$ for some $\widetilde{\lambda} \in \Omega^{n-1}(\widetilde{M})$, and acting on both sides of this relation with ψ^* implies also

$$\psi^*\pi^*\omega = \pi^*\omega = \psi^* d\widetilde{\lambda} = d(\psi^*\widetilde{\lambda}),$$

so $\psi^*\widetilde{\lambda}$ is another primitive for $\pi^*\omega$, and therefore so is the average

$$\widetilde{\alpha} := \frac{1}{2}(\widetilde{\lambda} + \psi^*\widetilde{\lambda}).$$

The latter satisfies $\psi^*\widetilde{\alpha} = \widetilde{\alpha}$, implying that it is a pullback $\pi^*\alpha$ of some $\alpha \in \Omega^{n-1}(M)$, which then satisfies $\pi^*(d\alpha) = d(\pi^*\alpha) = \pi^*\omega$, and therefore $d\alpha = \omega$.

In algebraic topology, there does exist a version of Poincarè duality for non-orientable manifolds, but it requires homology and cohomology with $\mathbb{Z}_2$ coefficients, so it cannot be seen in a straightforward way via de Rham's theory.

53.3. Differential operators on manifolds. We now begin setting up the analysis required for the proof of Theorem 53.18. For the remainder of this lecture, we dispense with the pseudo-Riemannian metric g and assume only that M is a smooth n-manifold with a pair of smooth vector bundles

$$E, F \to M$$

of ranks k and ℓ respectively over $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$. We will need to make frequent use of multi-index notation: a **multi-index** is a tuple $\alpha = (\alpha_1, \ldots, \alpha_n)$ of nonnegative integers, which gives rise to the differential operator

$$\partial^\alpha := \partial_1^{\alpha_1} \ldots \partial_n^{\alpha_n}$$

for functions of n real variables, as well as the polynomial

$$\mathbf{z}^\alpha := z_1^{\alpha_1} \ldots z_n^{\alpha_n} \in \mathbb{C} \qquad \text{for} \qquad \mathbf{z} = (z_1, \ldots, z_n) \in \mathbb{C}^n.$$

The nonnegative integer $|\alpha| := \alpha_1 + \ldots + \alpha_m$ is called the **order** of the multi-index, and is equal to the order of the differential operator ∂^α and the degree of the polynomial $\mathbf{z}^\alpha$. For each integer $m \geqslant 0$ there exist only finitely many multi-indices α with $|\alpha| \leqslant m$, and summations of the form

$$\sum_{|\alpha| \leqslant m} \text{expression dependent on } \alpha,$$

should thus be understood to mean the sum over the finite set of all multi-indices with order at most m; the expression $\sum_{|\alpha|=m}$ has a similar interpretation, involving all multi-indices of order *equal* to m.

DEFINITION 53.23. Given the vector bundles $E, F \to M$ of ranks k and ℓ respectively over $\mathbb{F}$, a linear map $D : \Gamma(E) \to \Gamma(F)$ is called a (linear) **differential operator** from E to F if for every open subset $\mathcal{U} \subset M$ and section $\eta \in \Gamma(E)$, $D\eta|_{\mathcal{U}}$ depends only on $\eta|_{\mathcal{U}}$, and for every choice of chart $(x^1, \ldots, x^n)$ for M and local trivializations of E and F over $\mathcal{U}$, the resulting isomorphisms $\Gamma(E|_{\mathcal{U}}) \cong C^\infty(\mathcal{U}, \mathbb{F}^k)$ and $\Gamma(F|_{\mathcal{U}}) \cong C^\infty(\mathcal{U}, \mathbb{F}^\ell)$ identify $\Gamma(E|_{\mathcal{U}}) \xrightarrow{D} \Gamma(F|_{\mathcal{U}})$ with a map of the form

$$C^\infty(\mathcal{U}, \mathbb{F}^k) \to C^\infty(\mathcal{U}, \mathbb{F}^\ell) : f \mapsto \sum_{|\alpha| \leqslant m} c_\alpha \partial^\alpha f$$

for some integer $m \geqslant 0$ and smooth matrix-valued functions $c_\alpha : \mathcal{U} \to \operatorname{Hom}(\mathbb{F}^k, \mathbb{F}^\ell) = \mathbb{F}^{\ell \times k}$, called the **coefficients** of the operator for these choices of coordinates and trivializations. We say that D has **order** m if m is the largest integer for which the coefficients c_α for some multi-index α of order m on some region with an associated chart and trivializations are not identically zero.

In these notes, we will usually omit the word "linear" when referring to differential operators, since we have no plans to discuss nonlinear operators (much as that is an interesting topic). The word "partial" is occasionally also inserted before "differential operator" if the underlying manifold M has dimension greater than 1, to contrast with the 1-dimensional situation in which there is only one direction of differentiation, and the theory of *ordinary* differential equations therefore applies. The 1-dimensional setting can be viewed as an interesting test case in which the theory of elliptic operators becomes tractable without as much need for functional-analytic tools, and we will refer to this in a few exercises, but our main objective is to prove results that are interesting in dimensions two and higher.

EXAMPLE 53.24. The case $m = 0$ is allowed in Definition 53.23, but is not very interesting since requiring $|\alpha| \leqslant 0$ allows only for the trivial multi-index $\alpha = (0, \ldots, 0)$: a differential operator of order zero is equivalent in local trivializations to an operator of the form

$$C^\infty(\mathcal{U}, \mathbb{F}^k) \to C^\infty(\mathcal{U}, \mathbb{F}^\ell) : f \mapsto cf$$

for some smooth function $c : \mathcal{U} \to \mathrm{Hom}(\mathbb{F}^k, \mathbb{F}^\ell)$, and is thus the same thing as a smooth linear bundle map $E \to F$. Bundle maps are therefore sometimes called **zeroth-order** operators. Equivalently, we say that a linear map $D : \Gamma(E) \to \Gamma(F)$ is **tensorial** if it is a zeroth-order operator, and this is true if and only if it is C^∞-linear.

EXAMPLE 53.25. For each $k = 0, \dots, n-1$, the exterior derivative $d : \Omega^k(M) = \Gamma(\Lambda^k T^*M) \to \Gamma(\Lambda^{k+1}T^*M) = \Omega^{k+1}(M)$ is a differential operator of order 1 from $\Lambda^k T^*M$ to $\Lambda^{k+1}T^*M$. One sees this from the local coordinate formula proved in §8.2 of last semester's notes: choosing a chart $(x^1, \dots, x^n)$ on some region $\mathcal{U} \subset M$ and using the coordinate forms $dx^{i_1} \wedge \dots \wedge dx^{i_\ell}$ for $1 \leqslant i_1 < \dots < \dots < \dots i_\ell \leqslant n$ to trivialize $\Lambda^* T^*M$ over $\mathcal{U}$, each k-form ω is identified on $\mathcal{U}$ with a vector-valued function whose components are the real-valued functions $\omega_{i_1 \dots i_k} = \omega(\partial_{i_1}, \dots, \partial_{i_k})$. Proposition 8.7 then gives

$$(d\omega)_{i_1 \dots i_{k+1}} = (k+1)\partial_{[i_1}\omega_{i_2 \dots i_{k+1}]} := \frac{1}{k!} \sum_{\sigma \in S_{k+1}} (-1)^{|\sigma|} \partial_{i_{\sigma(1)}} \omega_{i_{\sigma(2)} \dots i_{\sigma(k+1)}},$$

which is a linear combination (with nonzero constant coefficients) of expressions of the form $\partial_\ell \omega_{j_1 \dots j_k}$.

EXAMPLE 53.26. On a pseudo-Riemannian manifold, the operator $d^* : \Omega^k(M) = \Gamma(\Lambda^k T^*M) \to \Gamma(\Lambda^{k-1}T^*M) = \Omega^{k-1}(M)$ for $k = 1, \dots, n$ is the composition of the first-order operator $d : \Gamma(\Lambda^{n-k}T^*M) \to \Gamma(\Lambda^{n-k+1}T^*M)$ with two zeroth-order operators (two copies of the Hodge star operator, multiplied by a sign), and is thus also a first-order differential operator.

EXAMPLE 53.27. For any vector bundle $E \to M$ with a connection, the covariant derivative defines a first-order differential operator $\nabla : \Gamma(E) \to \Gamma(\mathrm{Hom}(TM, E)) = \Omega^1(M, E)$, whose local expression in an arbitrary choice of chart and trivialization generally involves first partial derivatives in every direction and some zeroth-order terms (the Christoffel symbols).

EXAMPLE 53.28. On an oriented pseudo-Riemannian manifold (M, g) with a spin structure and spinor bundle $E \to M$, the Dirac operator $D : \Gamma(E) \to \Gamma(E)$ can be expressed locally as a linear combination of zeroth-order operators (defined via Clifford multiplication) composed with covariant derivatives, and is thus a first-order differential operator from E to itself.

EXAMPLE 53.29. On any pseudo-Riemannian n-manifold (M, g), the Laplace-Beltrami operator Δ is a second-order differential operator from $\Lambda^k T^*M$ to itself for any $k = 0, \dots, n$. The formula (53.5) shows indeed that there are always nonzero coefficients in front of second derivatives in any local expression for Δ, but no derivatives of higher order.

Definition 53.23 is very general, and knowing that arbitrary differential operators can be written locally in the form $\sum_{|\alpha| \leqslant m} c_\alpha \partial^\alpha$ will be useful for analytical purposes, but it is not a very practical way of describing most of the geometrically meaningful operators that arise on manifolds. As we will see starting in the next lecture, the important analytical properties of many differential operators are determined by their terms of highest order, i.e. the terms $\sum_{|\alpha|=m} c_\alpha \partial^\alpha$ in local coordinates. The following result gives a useful way of extracting the information contained in these terms without needing to choose coordinates and trivializations.

LEMMA 53.30. *For every differential operator $D : \Gamma(E) \to \Gamma(F)$ of order $m \geqslant 0$, there exists a unique smooth and fiber-preserving (but not necessarily linear) map*

$$\sigma_D : T^*M \to \mathrm{Hom}(E, F)$$

that is characterized by the following property: for any $p \in M$, $\lambda \in T_p^*M$, $v \in E_p$, $\eta \in \Gamma(E)$ with $\eta(p) = v$ and $f \in C^\infty(M)$ with $f(p) = 0$ and $d_p f = \lambda$,

$$\sigma_D(\lambda)v = \frac{1}{m!} D(f^m \eta)(p) \in F_p.$$

Moreover, for two operators $D, D' : \Gamma(E) \to \Gamma(F)$ of order $m \geqslant 1$, we have $\sigma_D = \sigma_{D'}$ if and only if $D - D' : \Gamma(E) \to \Gamma(F)$ is a differential operator of order strictly less than m.

The map $\sigma_D : T^*M \to \mathrm{Hom}(E, F)$ in Lemma 53.30 is called the **principal symbol** of the operator $D : \Gamma(E) \to \Gamma(F)$.

Proof of Lemma 53.30. The map σ_D is clearly unique if it is well defined; in order to prove the latter, we need to show that for any point $p \in M$, any section $\eta \in \Gamma(E)$ and any function $f : M \to \mathbb{R}$ with $f(p) = 0$, the value of the section $D(f^m \eta) \in \Gamma(F)$ at $p \in M$ depends on $d_p f \in T_p^*M$ and $\eta(p) \in E_p$ but not any further on the function f or section η. We can prove this by a direct computation in local coordinates, thus after choosing a chart and trivializations over some region $\mathcal{U} \subset M$, let us consider an operator of the form

$$D = \sum_{|\alpha| \leqslant m} c_\alpha \partial^\alpha : C^\infty(\mathcal{U}, \mathbb{F}^k) \to C^\infty(\mathcal{U}, \mathbb{F}^\ell)$$

defined via smooth matrix-valued functions $c_\alpha : \mathcal{U} \to \mathrm{Hom}(\mathbb{F}^k, \mathbb{F}^\ell)$. Applying any mth-order partial derivative operator $\partial_{j_1} \dots \partial_{j_m}$ to the function $f^m : \mathcal{U} \to \mathbb{R}$, we have

$$\partial_{j_1} \dots \partial_{j_m}(f^m) = m \partial_{j_1} \dots \partial_{j_{m-1}} \left(f^{m-1} \partial_{j_m} f \right) = m(m-1) \partial_{j_1} \dots \partial_{j_{m-2}} \left(f^{m-2} \cdot \partial_{j_{m-1}} f \cdot \partial_{j_m} f + \dots \right),$$

where "$\dots$" in the last expression is an abbreviation for terms that contain f to the power of at least $m-1$ and will therefore vanish if we differentiate them $m-2$ more times and evaluate at p, since $f(p) = 0$. Continuing in this way and discarding all terms that do not decrease the exponent on f fast enough to make a nontrivial contribution at p, we find

$$\partial_{j_1} \dots \partial_{j_m}(f^m)(p) = m! \partial_{j_1} f(p) \cdot \ldots \cdot \partial_{j_m} f(p),$$

which translates into the language of multi-indices as the formula

$$\partial^\alpha(f^m)(p) = m! \cdot \nabla f(p)^\alpha \in \mathbb{R} \qquad \text{for } |\alpha| = m,$$

with ∇f in this setting denoting the classical gradient vector

$$\nabla f(p) := (\partial_1 f(p), \dots, \partial_n f(p)) \in \mathbb{R}^n.$$

A similar computation gives $\partial^\alpha(f^m)(p) = 0$ whenever $|\alpha| < m$. It follows that for $\eta : \mathcal{U} \to \mathbb{F}^k$, $\partial^\alpha(f^m \eta)(p)$ can only be nonzero if $|\alpha| = m$, and the only nontrivial contributions resulting from the Leibniz rule in this case come from differentiating f^m but not η, so

$$D(f^m \eta)(p) = \sum_{|\alpha| \leqslant m} c_\alpha(p) \partial^\alpha(f^m \eta)(p) = \sum_{|\alpha| = m} c_\alpha(p) \partial^\alpha(f^m)(p) \cdot \eta(p) = m! \cdot \sum_{|\alpha| = m} c_\alpha(p) \nabla f(p)^\alpha \eta(p).$$

This proves the claim that $D(f^m \eta)(p)$ depends only on $d_p f$ and $\eta(p)$, and the resulting local formula for $T_p^*M \xrightarrow{\sigma_D} \mathrm{Hom}(E_p, F_p)$ after using the chosen trivializations to identify $E_p \cong \mathbb{F}^k$ and $F_p \cong \mathbb{F}^\ell$ is

$$\sigma_D(\lambda_i \, dx^i) = \sum_{|\alpha| = m} (\lambda^1, \dots, \lambda^n)^\alpha c_\alpha(p) \in \mathrm{Hom}(\mathbb{F}^k, \mathbb{F}^\ell) \cong \mathrm{Hom}(E_p, F_p). \tag{53.6}$$

In particular, σ_D is an mth-degree polynomial function of $\lambda \in T_p^*M$ whose coefficients are the mth-order coefficients of the operator D at p. This proves that for two operators D and D' of order m, $\sigma_D = \sigma_{D'}$ at p if and only if their mth-order coefficients at p are identical, which is true at all points $p \in M$ if and only if $D - D'$ is an operator of order strictly less than m. $\square$

The significance of the next definition is unlikely to seem obvious at this stage, but we will spend considerable effort unpacking it in the next lecture.

Definition 53.31. A differential operator $D : \Gamma(E) \to \Gamma(F)$ is called **elliptic** if its principal symbol $\sigma_D : T^*M \to \mathrm{Hom}(E, F)$ has the following property: for every $p \in M$,

$$\lambda \neq 0 \in T_p^*M \qquad \Longrightarrow \qquad \sigma_D(\lambda) : E_p \to F_p \text{ is invertible.}$$

The main justification for the condition in Definition 53.31 is that operators that satisfy it have amazing properties, and we will see in the next lecture that several operators that arise naturally in various geometric situations actually satisfy it. The bulk of the work behind the proof of the Hodge decomposition theorem will consist in using tools from functional analysis to prove the following result, which implies that the space of harmonic forms on a closed Riemannian manifold is finite dimensional.

Theorem 53.32. *If $D : \Gamma(E) \to \Gamma(F)$ is an elliptic differential operator between vector bundles over a closed manifold M, then the vector spaces*

$$\ker(D) := \{\eta \in \Gamma(E) \mid D\eta \equiv 0\} \qquad \textit{and} \qquad \mathrm{coker}(D) := \Gamma(F) \Big/ \{D\eta \mid \eta \in \Gamma(E)\}$$

are both finite dimensional.

Exercise 53.33. Suppose $E \to M$ is a vector bundle equipped with a connection $\nabla : \Gamma(E) \to \Gamma(F)$, where $F := \mathrm{Hom}(TM, E)$.

(a) Compute the principal symbol $\sigma_\nabla : T^*M \to \mathrm{Hom}(E, F)$, and show that ∇ is an elliptic operator if and only if $\dim M = 1$.

(b) Give a direct proof of Theorem 53.32 for $D = \nabla$ when $M = S^1 = \mathbb{R}/\mathbb{Z}$.
Hint 1: The image $\mathrm{im}(\nabla) \subset \Gamma(F)$ has the same codimension as $\mathrm{im}(\nabla) \times E_p \subset \Gamma(F) \times E_p$ for any point $p \in S^1$. Show that $\mathrm{im}(\nabla) \times E_p$ is the kernel of a linear map $\Gamma(F) \times E_p \to E_p$.
Hint 2: Given $\xi \in \Gamma(F)$, $t_0 \in \mathbb{R}$ and $v \in E_{[t_0]}$, the ODE $\nabla\eta(t) = \xi([t])$ has a unique solution $\eta(t) \in E_{[t]}$ defined for $t \in \mathbb{R}$ with initial condition $\eta(t_0) = v$, but this solution might not be periodic.

(c) Prove that $\ker(\nabla)$ is also finite dimensional when M is a compact manifold of dimension greater than 1, but show by example that $\mathrm{coker}(\nabla)$ need not be.

(d) Part (a) has a converse of sorts: if $\dim M = 1$ and $D : \Gamma(E) \to \Gamma(F)$ is any first-order elliptic operator between two bundles $E, F \to M$, then there exists a bundle isomorphism $\Phi : F \to \mathrm{Hom}(TM, E)$ such that $\Phi \circ D : \Gamma(E) \to \Gamma(\mathrm{Hom}(TM, E)) = \Omega^1(M, E)$ is a connection on E. Prove this.
Hint: For first-order operators, (53.6) shows that the symbol $\sigma_D : T^*M \to \mathrm{Hom}(E, F)$ is a fiberwise linear map, so you could also regard it as a smooth linear bundle map $T^*M \otimes E \to F$. Show that it's an isomorphism if D is elliptic and $\dim M = 1$.

54. Ellipticity

Having defined what it means for a differential operator to be elliptic, there are two immediate questions we should try to answer in this lecture: (1) Why is ellipticity a helpful condition? (2) Does any of the operators we already know satisfy it? We'll attack these two questions in reverse order.

54.1. The Hodge Laplacian is elliptic. To see whether the operator $\Delta = dd^* + d^*d : \Gamma(\Lambda^k T^*M) \to \Gamma(\Lambda^k T^*M)$ is elliptic, we need to compute its principal symbol

$$\sigma_\Delta : T^*M \to \mathrm{End}(\Lambda^k T^*M).$$

We will break the computation down into manageable pieces, starting with the principal symbol of $d : \Gamma(\Lambda^k T^*M) \to \Gamma(\Lambda^{k+1} T^*M)$. This is a first-order operator, and the answer follows directly from the Leibniz rule: given $\omega \in \Omega^k(M)$ and $f : M \to \mathbb{R}$ with $f(p) = 0$ at some point $p \in M$, we have

$$d(f\omega)_p = d_p f \wedge \omega_p + f(p) d\omega_p = d_p f \wedge \omega_p,$$

thus $\sigma_d : T^*M \to \mathrm{Hom}(\Lambda^k T^*M, \Lambda^{k+1}T^*M)$ takes the form

$$\sigma_d(\lambda)\omega = \lambda \wedge \omega.$$

We observe that for $\lambda \neq 0 \in T_p^*M$, the map $\Lambda^k T_p^*M \to \Lambda^{k+1}T_p^*M : \omega \mapsto \lambda \wedge \omega$ is typically not invertible; indeed, it is never injective for $k \geqslant 1$ since one can then take ω to be a wedge product of λ with something else, forcing $\lambda \wedge \omega$ to vanish. The only exception is $k = 0$, where $\omega \in \Lambda^0 T_p^*M = \mathbb{R}$ is just a real number $c \in \mathbb{R}$ and $\lambda \wedge \omega = c\lambda \in \Lambda^1 T_p^*M$, so in this case the map is injective, and it is then surjective if and only if $\dim \Lambda^1 T_p^*M = 1$, which holds only when $\dim M = 1$. We conclude from Definition 53.31: the differential

$$d : C^\infty(M) = \Gamma(\Lambda^0 T^*M) \to \Gamma(\Lambda^1 T^*M) = \Omega^1(M)$$

is an elliptic operator whenever M is a 1-manifold, but all other cases of the exterior derivative $d : \Omega^k(M) \to \Omega^{k+1}(M)$ are not elliptic. This is consistent with Theorem 53.32, since in most cases, the space $\ker d \subset \Omega^k(M)$ of closed k-forms is not a finite-dimensional space, the major exception being the case $k = 0$, since 0-forms are closed if and only if they are locally constant.

Exercise 54.1. Without assuming Theorem 53.32, show explicitly that the kernel and cokernel of $d : C^\infty(S^1) \to \Omega^1(S^1)$ are both finite dimensional, but the cokernel of $d : C^\infty(M) \to \Omega^1(M)$ is infinite dimensional whenever $\dim M \geqslant 2$.
Hint: The case $M = S^1$ follows from Exercise 53.33, but there is also an easier proof using the characterization of exact 1-forms in Lecture 13 from the first semester (see especially Exercise 13.16). For $\dim M \geqslant 2$, find an infinite-dimensional space of 1-forms supported in a small coordinate neighborhood that are not closed.

The next step is to compute the symbol of $d^* : \Omega^k(M) \to \Omega^{k-1}(M)$, and in this context, it will be useful to discuss the concept of formal adjoints in more general terms.

Definition 54.2. Suppose $E, F \to M$ are smooth vector bundles equipped with bundle metrics, both denoted by $\langle\ ,\ \rangle$, and M is equipped with a volume form $d\mathrm{vol} \in \Omega^n(M)$.[113] Two differential operators $D : \Gamma(E) \to \Gamma(F)$ and $D^* : \Gamma(F) \to \Gamma(E)$ are said to be each other's **formal adjoints** if the relation

$$\int_M \langle \xi, D\eta \rangle \, d\mathrm{vol} = \int_M \langle D^*\xi, \eta \rangle \, d\mathrm{vol}$$

holds for all smooth sections $\eta \in \Gamma(E)$ and $\xi \in \Gamma(F)$ with compact support in $M \backslash \partial M$.

We saw two examples of formal adjoints in the previous lecture: on a pseudo-Riemannian manifold (M, g), the operator $d^* : \Omega^{k+1}(M) \to \Omega^k(M)$ is the formal adjoint of $d : \Omega^k(M) \to \Omega^{k+1}(M)$ with respect to the bundle metrics on $\Lambda^* T^*M$ and volume element on M determined by g. It follows easily from this that the Laplace-Beltrami operator $\Delta = dd^* + d^*d$ is its own formal adjoint, so it is **formally self-adjoint**, meaning

$$\int_M \langle \alpha, \Delta\beta \rangle \, d\mathrm{vol} = \int_M \langle \Delta\alpha, \beta \rangle \, d\mathrm{vol}$$

holds for all $\alpha, \beta \in \Omega^*(M)$ with compact support in $M \backslash \partial M$.

[113] The use of a volume form tacitly assumes M is oriented, but the definition as stated also makes sense for a non-orientable manifold if $d\mathrm{vol}$ is taken to be a volume element as in §11.4.

Remark 54.3. Here is a frequently asked question: why do we include the word "formal" and not just call D^* the "adjoint" of D? If we were talking about linear operators between finite-dimensional spaces, this would be perfectly fine, but the issue is that in infinite dimensions, functional analysis gives the words "adjoint" and "self-adjoint" much more precise and technical definitions than we are using here—they require regarding D as an unbounded linear operator between two Hilbert spaces and specifying a dense subspace as its domain, such that the notion of the adjoint operator depends in an essential way on this choice of dense subspace (see e.g. [**RS80**]). Adding the word "formal" gives us permission to avoid worrying about any of these functional-analytic details unless and until we actually need them, which in our exposition, we won't.

Finding global coordinate-invariant formulas for formal adjoints is typically a tricky business, but one can always use local coordinates and a partition of unity to show that they exist:

Proposition 54.4. *Fix vector bundles $E, F \to M$ with bundle metrics, along with a volume form or volume element on M and a differential operator $D : \Gamma(E) \to \Gamma(F)$ of order $m \geqslant 0$. Then D has a unique formal adjoint $D^* : \Gamma(F) \to \Gamma(E)$, which is also a differential operator of order m, and depends in general on the bundle metrics and volume element in addition to D. However, the principal symbol of D^* is independent of the choice of volume form, and is given by*

$$\sigma_{D^*}(\lambda) = (-1)^m \left(\sigma_D(\lambda)\right)^* \qquad \text{for all } p \in M,\ \lambda \in T_p^*M,$$

where for a linear map $A : E_p \to F_p$ we denote by $A^* : F_p \to E_p$ its adjoint with respect to the bundle metrics, satisfying $\langle w, Av\rangle = \langle A^*w, v\rangle$ for all $v \in E_p$ and $w \in F_p$.

Proof. The uniqueness of D^* is easy to see: if there were two operators D_1^* and D_2^* both satisfying the conditions of a formal adjoint for D, then they would satisfy

$$\int_M \langle D_1^*\xi - D_2^*\xi, \eta\rangle \, d\mathrm{vol} = 0$$

for all $\xi \in \Gamma(F)$ and $\eta \in \Gamma(E)$ with compact support in $M\backslash\partial M$, and by the nondegeneracy of the bundle metric on E, this is only possible for all η if $D_1^*\xi - D_2^*\xi \equiv 0$.

It remains to prove existence. Let us say that an operator $D : \Gamma(E) \to \Gamma(F)$ has **support** in a subset $\mathcal{U} \subset M$ if $D\eta$ vanishes for all sections $\eta \in \Gamma(E)$ that are trivial in $\mathcal{U}$. Consider first the case where D has compact support in an open subset $\mathcal{U} \subset M$ on which there exists a chart $(x^1, \dots, x^n)$ and trivializations identifying $E|_{\mathcal{U}}$ with $\mathcal{U} \times \mathbb{F}^k$ and $F|_{\mathcal{U}}$ with $\mathcal{U} \times \mathbb{F}^\ell$ so that the bundle metrics become constant scalar products on $\mathbb{F}^k$ and $\mathbb{F}^\ell$. The operator D is then completely determined by its action on sections restricted to $\mathcal{U}$, which is given by a formula of the form $D = \sum_{|\alpha| \leqslant m} c_\alpha \partial^\alpha$ for a set of smooth compactly supported coefficient functions $c_\alpha : \mathcal{U} \to \mathrm{Hom}(\mathbb{F}^k, \mathbb{F}^\ell)$. The measure defined on $\mathcal{U}$ by the volume element $d\mathrm{vol}$ is the product of some smooth function $f : \mathcal{U} \to (0, \infty)$ with the Lebesgue measure, which we will denote by $dx^1 \dots dx^n$. Writing $\eta \in \Gamma(E)$ and $\xi \in \Gamma(F)$ over $\mathcal{U}$ as functions $\mathcal{U} \to \mathbb{F}^k$ and $\mathcal{U} \to \mathbb{F}^\ell$ respectively, we can apply classical integration by parts and exploit the compact support of c_α to avoid boundary terms, giving

$$\begin{aligned}
\int_M \langle \xi, D\eta\rangle \, d\mathrm{vol} &= \sum_{|\alpha| \leqslant m} \int_{\mathcal{U}} \langle \xi, c_\alpha \partial^\alpha \eta\rangle \, f \, dx^1 \dots dx^n = \sum_{|\alpha| \leqslant m} \int_{\mathcal{U}} \langle f c_\alpha^* \xi, \partial^\alpha \eta\rangle \, dx^1 \dots dx^n \\
&= \sum_{|\alpha| \leqslant m} (-1)^{|\alpha|} \int_{\mathcal{U}} \langle \partial^\alpha (f c_\alpha^* \xi), \eta\rangle \, dx^1 \dots dx^n \\
&= \int_{\mathcal{U}} \left\langle \sum_{|\alpha| \leqslant m} (-1)^{|\alpha|} \frac{1}{f} \partial^\alpha (f c_\alpha^* \xi), \eta \right\rangle f \, dx^1 \dots dx^n =: \int_M \langle D^*\xi, \eta\rangle \, d\mathrm{vol},
\end{aligned}$$

where we have defined an operator $D^* : \Gamma(F) \to \Gamma(E)$ of order m with compact support in $\mathcal{U}$ by the formula

$$(54.1) \qquad D^*\xi := \sum_{|\alpha| \leqslant m} (-1)^{|\alpha|} \frac{1}{f} \partial^\alpha (f c_\alpha^* \xi)$$

for sections $\xi \in \Gamma(F|_{\mathcal{U}})$, and extended it trivially to the rest of M. By construction, this operator satisfies $\int_M \langle \xi, D\eta \rangle \, d\text{vol} = \int_M \langle D^*\xi, \eta \rangle \, d\text{vol}$ for all $\eta \in \Gamma(E)$ and $\xi \in \Gamma(F)$, not just sections of compact support. Observe moreover that any compact subset $K \subset \mathcal{U}$ containing the support of D also contains the support of D^*. Its highest-order term D_m^* consists of all summands in (54.1) in which ξ is differentiated m times, which means that f and c_α^* cannot be differentiated at all, giving

$$D_m^*\xi = (-1)^m \sum_{|\alpha|=m} c_\alpha^* \partial^\alpha \xi.$$

This implies the stated formula for the principal symbol of D^*, and also that it does not depend on the choice of volume element, since the function f does not appear.

Now suppose $D : \Gamma(E) \to \Gamma(F)$ is an arbitrary differential operator with no assumptions about its support. Choose an open covering $\{\mathcal{U}_\beta\}_{\beta \in I}$ of M by subsets $\mathcal{U}_\beta$ with compact closure on which charts and trivializations as in the previous paragraph can be defined, choose a partition of unity $\{\varphi_\beta : M \to [0, 1]\}_{\beta \in I}$ subordinate to this covering, and define $D_\beta := \beta D$ for each $\beta \in I$. The construction in the previous paragraph then furnishes each of the compactly supported operators D_β with a formal adjoint D_β^*, and we define

$$D^* := \sum_{\beta \in I} D_\beta^* : \Gamma(F) \to \Gamma(E),$$

an expression that makes sense because the support of each operator D_β^* is contained in the support of the function φ_β, thus forming a locally finite covering. If $\eta \in \Gamma(E)$ and $\xi \in \Gamma(F)$ both have support inside some compact set $K \subset M$, then local finiteness also implies the existence of a finite subset $J \subset I$ such that $D = \sum_{\beta \in J} D_\beta$ and $D^* = \sum_{\beta \in J} D_\beta^*$ on K, thus

$$\int_M \langle \xi, D\eta \rangle \, d\text{vol} = \sum_{\beta \in J} \int_M \langle \xi, D_\beta \eta \rangle \, d\text{vol} = \sum_{\beta \in J} \int_M \langle D_\beta^* \xi, \eta \rangle \, d\text{vol} = \int_M \langle D^*\xi, \eta \rangle \, d\text{vol}.$$

This proves that D^* is indeed a formal adjoint for D, and the previous calculation in local coordinates can be reused in a neighborhood of any point to prove the stated formula for σ_{D^*}. □

Since a linear map between finite-dimensional spaces is invertible if and only if its adjoint is, Proposition 54.4 implies:

Corollary 54.5. *A differential operator D is elliptic if and only if its formal adjoint D^* is elliptic.* □

The next exercise is quite easy if you remember the local coordinate formula (53.6) for the principal symbol.

Exercise 54.6. Suppose $E, F, G \to M$ are vector bundles and we are given two differential operators $D_1 : \Gamma(E) \to \Gamma(F)$ and $D_2 : \Gamma(F) \to \Gamma(G)$ of orders $m_1, m_2 \geqslant 0$ respectively. Show that $D_2 \circ D_1 : \Gamma(E) \to \Gamma(G)$ is then a differential operator of order at most $m_1 + m_2$, and its order is exactly $m_1 + m_2$ if and only if the map $T^*M \to \text{Hom}(E, G)$ given by

$$\lambda \mapsto \sigma_{D_2}(\lambda) \circ \sigma_{D_1}(\lambda)$$

is nontrivial, in which case this is the principal symbol $\sigma_{D_2 \circ D_1}$.

The way to compute σ_Δ should now be clear; we just need a bit more algebra in order to identify the adjoint of the map $\sigma_d(\lambda)\omega = \lambda \wedge \omega$ and compute the resulting compositions. For the next two lemmas, fix an n-dimensional vector space V with a nondegenerate symmetric bilinear form $\langle\ ,\ \rangle$, transferred to V^* via the musical isomorphisms $V \to V^* : v \mapsto v_\flat := \langle v, \cdot\rangle$ and $\sharp = \flat^{-1} : V^* \to V$, and then extended to $\Lambda^* V^*$ via Lemma 53.8.

LEMMA 54.7. *For any $\lambda \in \Lambda^1 V^*$ and $\alpha, \beta \in \Lambda^* V^*$, $\langle \alpha, \lambda \wedge \beta\rangle = \langle \iota_{\lambda^\sharp}\alpha, \beta\rangle$.*

PROOF. The set of elements $\lambda \in V^* = \Lambda^1 V^*$ with $\langle\lambda,\lambda\rangle \neq 0$ is dense in V^*, so if we can prove the stated relation for all λ in this set, the general case will follow by continuity. Let us therefore assume $\langle\lambda,\lambda\rangle \neq 0$, and since both sides of the relation scale the same way under multiplication by positive numbers, rescale λ so that $\langle\lambda,\lambda\rangle = \pm 1$ without loss of generality. In this case there exists an orthonormal basis $e_1, \dots, e_n$ of V giving rise to dual vectors $e^i_\flat := \langle e_i, \cdot\rangle$ that form an orthonormal basis of V^* with $e^1_\flat = \lambda$. By bilinearity, it then suffices to verify that the relation holds whenever α and β are both products of the form $e^{i_1}_\flat \wedge \dots \wedge e^{i_k}_\flat$ with $k \geqslant 0$ and $1 \leqslant i_i < \dots < i_k \leqslant n$. The crucial ingredients are now the formulas

$$\begin{aligned} \iota_{e_1}(e^1_\flat \wedge e^{i_2}_\flat \wedge \dots \wedge e^{i_k}_\flat) &= e^1_\flat(e_1)\, e^{i_2}_\flat \wedge \dots \wedge e^{i_k}_\flat = \langle e_1, e_1\rangle e^{i_2}_\flat \wedge \dots \wedge e^{i_k}_\flat, \\ \iota_{e_1}(e^{i_1}_\flat \wedge \dots \wedge e^{i_k}_\flat) &= 0 \qquad \text{if } i_1 > 1. \end{aligned}$$

If the product forming β begins with $e^1_\flat = \lambda$, then $\lambda \wedge \beta = 0$ forces the left hand side to vanish, and the right hand side will also vanish because $\iota_{e_1}\alpha$ cannot contain a factor of $e^1_\flat$. Similarly, the left hand side vanishes if α does not contain a factor of $e^1_\flat$, and so does the right hand side since $\iota_{e_1}\alpha$ is then zero. The only way to get something nontrivial on either side is thus when β takes the form $e^{i_2}_\flat \wedge \dots \wedge e^{i_k}_\flat$ for some $i_2 > 1$ and $\alpha = e^1_\flat \wedge e^{i_2}_\flat \wedge \dots \wedge e^{i_k}_\flat$, in which case the relation says

$$\begin{aligned} \langle \alpha, \lambda \wedge \beta\rangle &= \langle e^1_\flat \wedge e^{i_2}_\flat \wedge \dots \wedge e^{i_k}_\flat, e^1_\flat \wedge e^{i_2}_\flat \wedge \dots \wedge e^{i_k}_\flat\rangle \\ &= \langle e_1, e_1\rangle \cdot \langle e^{i_2}_\flat \wedge \dots \wedge e^{i_k}_\flat, e^{i_2}_\flat \wedge \dots \wedge e^{i_k}_\flat\rangle \\ &= \big\langle \langle e_1, e_1\rangle e^{i_2}_\flat \wedge \dots \wedge e^{i_k}_\flat, e^{i_2}_\flat \wedge \dots \wedge e^{i_k}_\flat \big\rangle = \langle \iota_{\lambda^\sharp}\alpha, \beta\rangle. \end{aligned}$$

□

LEMMA 54.8. *For any $\lambda \in \Lambda^1 V^*$ and $\omega \in \Lambda^* V^*$, $\iota_{\lambda^\sharp}(\lambda \wedge \omega) + \lambda \wedge \iota_{\lambda^\sharp}\omega = \langle\lambda,\lambda\rangle\omega$.*

PROOF. This follows easily from the graded Leibniz rule for the interior product (cf. Exercise 14.7 from last semester), which in this situation gives

$$\iota_{\lambda^\sharp}(\lambda \wedge \omega) = (\iota_{\lambda^\sharp}\lambda) \wedge \omega - \lambda \wedge \iota_{\lambda^\sharp}\omega = \langle\lambda,\lambda\rangle\omega - \lambda \wedge \iota_{\lambda^\sharp}\omega.$$

□

By Proposition 54.4 and Exercise 54.7, the operator $d^* : \Omega^*(M) \to \Omega^*(M)$ has principal symbol $\sigma_{d^*} : T^*M \to \mathrm{End}(\Lambda^* T^*M)$ given by

$$\sigma_{d^*}(\lambda)\omega = -\iota_{\lambda^\sharp}\omega.$$

Applying Exercise 54.6 and Lemma 54.8 to the formula $\Delta = dd^* + d^*d$ now gives:

THEOREM 54.9. *On a pseudo-Riemannian manifold (M, g), the Laplace-Beltrami operator $\Delta : \Omega^k(M) \to \Omega^k(M)$ for each $k \in \mathbb{Z}$ has principal symbol $\sigma_\Delta : T^*M \to \mathrm{End}(\Lambda^k T^*M)$ given by*

$$\sigma_\Delta(\lambda)\omega = -\langle\lambda,\lambda\rangle\omega.$$

□

Corollary 54.10. *The operator* Δ *on* (M, g) *is elliptic if and only if the metric* g *is (positive or negative) definite.* $\square$

With Corollary 54.10 understood, we will not consider indefinite metrics any further in our discussion of Hodge theory. All metrics from now on will be Riemannian.

Exercise 54.11. Deduce from Theorem 54.9 the local coordinate formula for Δ advertised in (53.5), namely

$$(\Delta\omega)_{i_1\dots i_k} = -g^{j\ell}\partial_j\partial_\ell\omega_{i_1\dots i_k} + \text{terms of order} < 2.$$

Exercise 54.12. Write down the principal symbol of the Dirac operator $D : \Gamma(E) \to \Gamma(E)$ on a spinor bundle $E \to M$ over a pseudo-Riemannian manifold (M, g) with a spin structure. Under what conditions is D elliptic?

Exercise 54.13. Assume M is a complex n-manifold, so its tangent spaces are naturally complex vector spaces. We can associate to any complex vector bundle $E \to M$ another complex vector bundle $F := \overline{\mathrm{Hom}}(TM, E)$ whose fiber over a point $p \in M$ is the space of complex-antilinear maps $T_pM \to E_p$. A **Cauchy-Riemann type** operator on $E \to M$ is a first-order linear differential operator $D : \Gamma(E) \to \Gamma(F)$ that satisfies the Leibniz rule

$$D(f\eta) = f\,D\eta + \bar{\partial}f(\cdot)\eta \qquad \text{for all} \qquad \eta \in \Gamma(E),\ f \in C^\infty(M, \mathbb{C}),$$

where we define $\bar{\partial}f \in \Omega^1(M, \mathbb{C})$ by $\bar{\partial}f(X) := df(X) + i\,df(iX)$. Show that all Cauchy-Riemann type operators on $E \to M$ have the same principal symbol. What is it? In what situation are these operators elliptic?

54.2. Fourier transforms. For the rest of this lecture and most of the next one, this differential geometry course will feel more like an analysis course, so maybe I should first try to convey why this is unavoidable. As we saw in Lecture 53, one of the most important statements contained in the Hodge decomposition theorem is that the space of harmonic forms on a closed Riemannian manifold is finite dimensional, which then implies that $H^*_{\mathrm{dR}}(M)$ is finite dimensional for any closed manifold M. It is fairly easy to prove that all harmonic forms are also closed, and in fact we will see that $\ker(\Delta)$ is in general the intersection of the space of closed forms $\ker(d)$ with the space of so-called *co-closed* forms $\ker(d^*)$. But in most situations, $\ker(d)$ and $\ker(d^*)$ are both infinite-dimensional spaces, so it is quite surprising and non-obvious that their intersection should be finite dimensional.

Now, what methods do we have at our disposal for proving that a certain linear subspace

$$X \subset Y$$

of a manifestly infinite-dimensional vector space Y is finite dimensional? One example you have probably seen before is the space of solutions to a linear ODE, e.g. if $\mathbf{A}(t) \in \mathbb{R}^{n\times n}$ is a smooth family of matrices defined for $t \in \mathbb{R}$, then the subspace

$$X := \left\{\mathbf{x} \in C^\infty(\mathbb{R}, \mathbb{R}^n) \;\middle|\; \dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) \text{ for all } t \in \mathbb{R}\right\}$$

is finite dimensional because, by the uniqueness statement in the Picard-Lindelöf theorem, the map $X \to \mathbb{R}^n : \mathbf{x} \mapsto \mathbf{x}(0)$ is injective. This trick works for solutions of ODEs because they satisfy a very strong uniqueness result, but more general PDEs almost never have the property that their solutions are determined by their values at one point, nor even by the values of finitely many of their derivatives. So for more general PDEs, cleverer ideas are needed.

One such idea is to distinguish between finite and infinite dimensionality via the notion of compactness, as measured via the existence of convergent subsequences. Indeed, as you learned in first-year analysis, finite-dimensional vector spaces have the pleasant property that bounded sequences always have convergent subsequences, whereas arbitrary metric spaces do not have this property. In particular, a bounded sequence in an infinite-dimensional Hilbert space need not have a convergent subsequence, as shown by the example of a countably infinite orthonormal basis. So if we can arrange for the infinite-dimensional space Y to have a Hilbert space structure, then one way of proving that $X \subset Y$ is finite dimensional would be to prove that every sequence in X that is bounded with respect to the Hilbert space norm also has a convergent subsequence in that norm. This is the method we will use for proving that $\ker \Delta \subset \Omega^k(M)$ is finite dimensional, but doing so will require replacing $\Omega^k(M)$ with an actual Hilbert space, because the space of *smooth* k-forms is not complete with respect to any inner product that one can reasonably write down. The simplest and most useful Hilbert spaces are not spaces of smooth functions—they are typically spaces like $L^2(\mathbb{R}^n)$, whose elements are not always continuous, and are strictly speaking not even functions, but rather equivalence classes of functions defined almost everywhere. Having to consider such functions seems a bit strange in the context of differential geometry, where up until now we have assumed wherever possible that all maps are smooth—but in Hodge theory there is an enormous payoff for allowing non-smooth functions and forms, because doing so provides us with some analytical tools that are quite powerful.

One of those tools is the Fourier transform, of which we shall now give a brief overview without detailed proofs. Expositions of the main results with complete proofs can be found in various sources, for instance [**Wen20**, **LL01**, **RS80**].

For the rest of this lecture, we will be considering functions defined on $\mathbb{R}^n$ with values in a finite-dimensional complex inner product space $(V, \langle\ ,\ \rangle)$. I should clarify perhaps that $\langle\ ,\ \rangle$ is a *positive* Hermitian inner product—we do not want to allow indefinite products here because we want it to define an actual norm. Let us denote the Euclidean scalar product on $\mathbb{R}^n$ by

$$x \cdot y = \sum_{j=1}^n x^j y^j$$

for vectors $x = (x^1, \dots, x^n), y = (y^1, \dots, y^n) \in \mathbb{R}^n$. For integrable functions $f : \mathbb{R}^n \to V$ of the variables $x = (x^1, \dots, x^n) \in \mathbb{R}^n$, we will denote integration with respect to the Lebesgue measure by

$$\int_{\mathbb{R}^n} f(x)\, dx := \int_{\mathbb{R}^n} f(x^1, \dots, x^n)\, dx^1 \dots dx^n \in V,$$

and the L^2-inner product of two functions $f, g : \mathbb{R}^n \to V$ will be written as

$$\langle f, g \rangle_{L^2} := \int_{\mathbb{R}^n} \langle f(x), g(x) \rangle\, dx \in \mathbb{C}.$$

In addition to the usual Banach spaces $L^p(\mathbb{R}^n)$ with the norms

$$\|f\|_{L^p} := \left(\int_{\mathbb{R}^n} |f(x)|^p\, dx \right)^{1/p}$$

for $1 \leqslant p < \infty$, we will have occasion to consider the Banach spaces $C^k(\mathbb{R}^n)$ for integers $k \geqslant 0$, consisting of C^k-functions $f : \mathbb{R}^n \to V$ whose derivatives up to order k are all bounded, with the norm given by

$$\|f\|_{C^k} := \sum_{|\alpha| \leqslant k} \sup_{x \in \mathbb{R}^n} |\partial^\alpha f(x)|.$$

Integration is also used to define the **convolution**, an operation that produces out of a scalar-valued function $f : \mathbb{R}^n \to \mathbb{C}$ and a vector-valued function $g : \mathbb{R}^n \to V$ another vector-valued function $f * g : \mathbb{R}^n \to V$ given by

$$(f * g)(x) := \int_{\mathbb{R}^n} f(x - y) g(y)\, dy.$$

The **Fourier transform** is an operation $\mathscr{F}$ that, for a suitable class of functions $f : \mathbb{R}^n \to V$, changes f into a new function $\mathscr{F}(f) = \hat{f} : \mathbb{R}^n \to V$ given by

$$\mathscr{F}(f)(p) = \hat{f}(p) := \int_{\mathbb{R}^n} e^{-2\pi i p \cdot x} f(x)\, dx. \tag{54.2}$$

A similar operator $\mathscr{F}^*$ is defined by changing the sign on the complex exponential: for $g : \mathbb{R}^n \to V$,

$$\mathscr{F}^*(g)(x) = \check{g}(x) := \int_{\mathbb{R}^n} e^{2\pi i p \cdot x} g(p)\, dp. \tag{54.3}$$

These formulas do not actually give the most general possible definition of $\mathscr{F}$ and $\mathscr{F}^*$, as these operators can also be defined on certain classes of functions for which the integrals above are not defined, but the integrals are considered a starting point for the theory of Fourier transforms. In order to say something more precise, it is helpful to introduce the so-called **Schwartz space**, a space of smooth functions on $\mathbb{R}^n$ whose derivatives of all orders decay at infinity faster than any polynomial:

$$\mathscr{S}(\mathbb{R}^n) := \left\{ f \in C^\infty(\mathbb{R}^n, V) \;\middle|\; \lim_{|x|\to\infty} \left| x^\alpha \partial^\beta f(x) \right| = 0 \text{ for all multi-indices } \alpha, \beta \right\}.$$

While the conditions defining $\mathscr{S}(\mathbb{R}^n)$ are quite strict, it is clearly not a small space since it contains

$$C_0^\infty(\mathbb{R}^n) := \left\{ f \in C^\infty(\mathbb{R}^n, V) \;\middle|\; f \text{ has compact support} \right\}.$$

It also contains Gaussians such as $f(x) = e^{-|x|^2}$, and therefore also all their derivatives or products of their derivatives with polynomials. In particular, $\mathscr{S}(\mathbb{R}^n)$ is a large enough space to be dense in all of the spaces $L^p(\mathbb{R}^n)$ for $1 \leqslant p < \infty$, while also being small enough to be contained in all of those spaces, and sufficiently well-behaved so that all calculations one would like to perform involving integration by parts or differentiation under the integral sign actually work.

Here are the main facts about the Fourier transform that we will need to make use of.

THEOREM 54.14. *The operators $\mathscr{F}$ and $\mathscr{F}^*$ have the following properties:*

(1) They define bijective linear maps $\mathscr{S}(\mathbb{R}^n) \to \mathscr{S}(\mathbb{R}^n)$ with $\mathscr{F}^{-1} = \mathscr{F}^*$.
(2) They have unique extensions to unitary *isomorphisms $L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$.*
(3) They define continuous linear operators $L^1(\mathbb{R}^n) \to C^0(\mathbb{R}^n)$.
(4) On $\mathscr{S}(\mathbb{R}^n)$, they convert differential operators into multiplication by polynomials: specifically,

$$\begin{aligned} \widehat{\partial^\alpha f}(p) &= (2\pi i p)^\alpha \hat{f}(p), \qquad & \widecheck{\partial^\alpha g}(x) &= (-2\pi i x)^\alpha \check{g}(x), \\ \partial^\alpha \check{g}(x) &= (2\pi i)^{|\alpha|} \widecheck{g_\alpha}(x), \qquad & \partial^\alpha \hat{f}(p) &= (-2\pi i)^{|\alpha|} \widehat{f_\alpha}(p), \end{aligned} \tag{54.4}$$

for every multi-index α, where we denote $f_\alpha(x) := x^\alpha f(x)$, $g_\alpha(p) := p^\alpha g(p)$.
(5) It converts the convolution operation into multiplication: specifically, if $f \in L^1(\mathbb{R}^n)$ and $g \in L^2(\mathbb{R}^n)$ and at least one of the two functions is scalar-valued, one has

$$\mathscr{F}(f * g) = \hat{f}\hat{g}.$$

□

While we will not prove Theorem 54.14 here, it is worth making a few comments.

Point (2) in this theorem is what we meant when we said above that $\mathscr{F}$ and $\mathscr{F}^*$ cannot always be defined via the integrals (54.2) and (54.3), as these integrals are not well defined as Lebesgue integrals unless $f, g \in L^1(\mathbb{R}^n)$, which is not always true for L^2-functions on $\mathbb{R}^n$. The key word however is *unitary*, meaning that the Fourier transform satisfies the relation

$$\langle f, g \rangle_{L^2} = \langle \hat{f}, \hat{g} \rangle_{L^2},$$

a result known as *Plancherel's theorem.* For $f, g \in \mathscr{S}(\mathbb{R}^n)$, this relation can be proved directly from the integral formulas for $\mathscr{F}$ and $\mathscr{F}^*$, and it implies that the maps $\mathscr{F}, \mathscr{F}^* : \mathscr{S}(\mathbb{R}^n) \to \mathscr{S}(\mathbb{R}^n)$ are not only bijective but also continuous with respect to the L^2-norm. Since $\mathscr{S}(\mathbb{R}^n)$ is dense in $L^2(\mathbb{R}^n)$, it follows that both operators have unique extensions to continuous linear maps $L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ such that Plancherel's theorem is also valid for all $f, g \in L^2(\mathbb{R}^n)$. The general definition of $\widehat{f} \in L^2(\mathbb{R}^n)$ for $f \in L^2(\mathbb{R}^n)$ is therefore a bit cumbersome: instead of using (54.2) to compute it, one must in principle find a sequence $f_k \in \mathscr{S}(\mathbb{R}^n)$ that converges in L^2 to f, compute $\widehat{f}_k \in \mathscr{S}(\mathbb{R}^n)$ from the integral, and then define $\widehat{f}$ as the L^2-limit of $\widehat{f}_k$ (which is guaranteed to exist due to Plancherel's theorem).

On the other hand, one can plug any $f \in L^1(\mathbb{R}^n)$ into (54.2) and have a well-defined integral, and the result obviously satisfies

$$\|\widehat{f}\|_{L^\infty} \leqslant \|f\|_{L^1}.$$

If we apply this observation to a function $f \in \mathscr{S}(\mathbb{R}^n)$, then we also know from point (1) that $\widehat{f}$ is in $\mathscr{S}(\mathbb{R}^n)$ and therefore continuous, so it is sensible to write

$$\|\widehat{f}\|_{C^0} \leqslant \|f\|_{L^1}. \tag{54.5}$$

But now density comes through for us again: $\mathscr{S}(\mathbb{R}^n)$ is a dense subspace of $L^1(\mathbb{R}^n)$, so it follows that $\mathscr{F}$ extends uniquely to a continuous linear operator from $L^1(\mathbb{R}^n)$ to the L^∞-closure of $\mathscr{S}(\mathbb{R}^n)$, which happens to be $C^0(\mathbb{R}^n)$, so (54.5) is valid for all $f \in L^1(\mathbb{R}^n)$, and this proves point (3).

The relations in point (4) are easy to prove for Schwartz functions using a combination of differentiation under the integral sign and integration by parts. For any individual multi-index α, they are also valid more generally: in the case $|\alpha| = 1$ in particular, the proofs of the formulas for $\widehat{\partial_j f}$ and $\widecheck{\partial_j f}$ work for any $f \in L^1(\mathbb{R}^n)$ that has a continuous partial derivative $\partial_j f$ also in $L^1(\mathbb{R}^n)$ and that decays at infinity in the sense that

$$\lim_{R\to\infty} \sup_{x\in\mathbb{R}^n\setminus B_R^n(0)} |f(x)| = 0;$$

the latter condition eliminates the boundary term when integrating by parts. Similarly, the formulas for $\partial_j \widehat{f}$ and $\partial_j \check{f}$ can be proved via differentiation under the integral sign whenever both f and the function $x \mapsto x^j f(x)$ belong to $L^1(\mathbb{R}^n)$, and it follows from point (3) in this situation that $\partial_j \widehat{f}$ and $\partial_j \check{f}$ are both continuous and bounded. With these calculations as a foundation, one can also show that all four of the formulas in (54.4) are valid when f and g are taken to be *tempered distributions* instead of functions, which means in particular that they are valid for some generalized interpretation of the symbol "∂^α" acting on much larger classes of functions than just $\mathscr{S}(\mathbb{R}^n)$. This has to do with the notion of *weak derivatives* and *weak solutions*, which we'll get into a bit later; for now, the main message you should take from (54.4) is that the Fourier transform should certainly be a useful tool in the study of PDEs, because in sufficiently nice cases, it converts a PDE into an *algebraic* equation involving multiplication by a polynomial, which sounds much easier to solve than a PDE.

EXERCISE 54.15. Use the integral definition of $\mathscr{F}$ to prove the formula in Theorem 54.14(5) for the Fourier transform of a convolution. Assume whatever you want to assume about f and g in order to make the calculation work.

54.3. Sobolev spaces on $\mathbb{R}^n$. We can now introduce a class of Hilbert spaces that serves as the natural functional-analytic setting for many PDEs. Since our definition uses the Fourier transform, we will continue under the assumption that the functions we consider take values in a *complex* inner product space V. If one wishes instead to assume V is a real vector space, then the definitions can easily be adapted by regarding V as a subspace of its complexification.

The spaces we will define are motivated by two considerations that seem at first to conflict with each other:

(1) If we want to study differential operators of order $m \geqslant 1$, we need a space of functions that can be differentiated m times.
(2) The simplest Hilbert space we know of is $L^2(\mathbb{R}^n)$, so we want to define a space that is in some sense modelled on that one.

The naive definition one might at first produce from these considerations is

$$H^m(\mathbb{R}^n) := \left\{ f \in L^2(\mathbb{R}^n) \;\middle|\; \partial^\alpha f \in L^2(\mathbb{R}^n) \text{ for all } |\alpha| \leqslant m \right\},$$

with inner product and norm

$$\langle f, g \rangle_{H^m} := \sum_{|\alpha| \leqslant m} \langle \partial^\alpha f, \partial^\alpha g \rangle_{L^2}, \qquad \|f\|_{H^m} := \sqrt{\langle f, f \rangle_{H^m}},$$

which one might sometimes prefer to replace with the equivalent norm

$$\|f\|_{H^m} := \sum_{|\alpha| \leqslant m} \|\partial^\alpha f\|_{L^2}. \tag{54.6}$$

These definitions should for now by taken with a grain of salt, because there is a big technical problem: functions in $L^2(\mathbb{R}^n)$ are not typically continuous, much less m times differentiable, so expressions like $\partial^\alpha f$ cannot generally be defined, which would be a prerequisite for testing whether they belong to $L^2(\mathbb{R}^n)$. One could imagine perhaps considering only functions $f \in L^2(\mathbb{R}^n)$ that are also of class $C^m(\mathbb{R}^n)$, so that $\partial^\alpha f$ is a well-defined continuous function and we can require it to belong to $L^2(\mathbb{R}^n)$. But this does not produce a complete space in general, as one can see from the case $m = 0$: $H^0(\mathbb{R}^n)$ would by this definition be $L^2(\mathbb{R}^n) \cap C^0(\mathbb{R}^n)$ with the L^2-norm, but it is easy to find L^2-Cauchy sequences in that intersection that do not converge to continuous functions. In order to define an actual Hilbert space, we need a cleverer way of defining the expression $\partial^\alpha f$ for $f \in L^2(\mathbb{R}^n)$.

The Fourier transform provides the simplest of a few possible solutions to this problem, because for $f \in \mathscr{S}(\mathbb{R}^n)$, items (2) and (4) in Theorem 54.14 make it possible to rewrite the norm in (54.6) as

$$\|f\|_{H^m} = \sum_{|\alpha| \leqslant m} \|(2\pi i p)^\alpha \hat{f}\|_{L^2}, \tag{54.7}$$

i.e. for each multi-index α of order at most m, one needs to multiply the L^2-function $\hat{f}(p)$ on $\mathbb{R}^n$ by the polynomial $(2\pi i p)^\alpha$ and compute the L^2-norm of the product. The result might be finite or infinite, but it can be defined for *every* $f \in L^2(\mathbb{R}^n)$, with no need to talk about derivatives, using only the knowledge that every L^2-function has a Fourier transform. One then obtains a reasonable Hilbert space by taking the set of all $f \in L^2(\mathbb{R}^n)$ with $\|f\|_{H^m} < \infty$. It is conventional in practice to replace (54.7) with a different but equivalent norm that is easier to compute with: our official definition of the H^m-inner product and norm will thus be

$$\begin{aligned} \langle f, g \rangle_{H^m} &:= \int_{\mathbb{R}^n} (1 + |p|^2)^m \langle \hat{f}(p), \hat{g}(p) \rangle \, dp, \\ \|f\|_{H^m} &:= \sqrt{\langle f, f \rangle_{H^m}} = \left(\int_{\mathbb{R}^n} (1 + |p|^2)^m |\hat{f}(p)|^2 \, dp \right)^{1/2} = \|(1 + |p|^2)^{m/2} \hat{f}\|_{L^2}. \end{aligned} \tag{54.8}$$

You should take a moment to convince yourself that the norm defined in this way is equivalent to the one in (54.7), meaning each can be bounded by a constant times the other, so they define the same topology on the appropriate space of functions. This space is called a **Sobolev space**

$$H^m(\mathbb{R}^n) := \left\{ f \in L^2(\mathbb{R}^n) \;\middle|\; \|f\|_{H^m} < \infty \right\},$$

and the direct relationship between the H^m-norm in (54.8) and the L^2-norm makes it quite easy to show that $H^m(\mathbb{R}^n)$ is a Hilbert space, with the transformation

$$H^m(\mathbb{R}^n) \to L^2(\mathbb{R}^n) : f \mapsto \mathscr{F}^*\left((1+|p|^2)^{m/2}\hat{f}\right)$$

defining a unitary isomorphism. You should think of $H^m(\mathbb{R}^n)$ intuitively as the space of L^2-functions whose derivatives up to order m are also in L^2, even though this is not strictly true in the classical sense of differentiability. It is true for some weaker notion of differentiability, which for now you can understand to be defined in terms of multiplying polynomials by Fourier transforms.

Our standing assumption so far has been that $m \geqslant 0$ is an integer, but it is interesting to note that the norm defined in (54.8) makes sense for any nonnegative real number $m \geqslant 0$, thus defining a notion of "fractional differentiability". That notion has actual applications for PDEs in certain settings, though we will not make direct use of it in our exposition. One can also define $H^m(\mathbb{R}^n)$ for $m < 0$, but that would require a longer explanation, because that case of $H^m(\mathbb{R}^n)$ cannot be considered a subspace of $L^2(\mathbb{R}^n)$, and its elements are generally not even functions, nor equivalence classes of functions—they are *tempered distributions*. We will come back to this in Lecture 56 when it is needed for regularity results, but until then, you can always assume $m \geqslant 0$.

As mentioned above, the functions in $H^m(\mathbb{R}^n)$ are not generally m times differentiable in the classical sense, but it turns out that they are at least *somewhat* differentiable, to a degree that can be quantified based on the values of m and n. It will be important to be precise about this since, at the end of the day, after all necessary technical lemmas have been proved, the functions we are always most interested in are the smooth ones.

THEOREM 54.16 (Sobolev embedding theorem for $H^s(\mathbb{R}^n)$). *For any real number $s > n/2$ and every integer $k \geqslant 0$, there exists a continuous linear inclusion*

$$H^{s+k}(\mathbb{R}^n) \hookrightarrow C^k(\mathbb{R}^n).$$

A word about the meaning of the statement in Theorem 54.16: the elements of $C^k(\mathbb{R}^n)$ are functions $f : \mathbb{R}^n \to V$, while elements of $H^{s+k}(\mathbb{R}^n) \subset L^2(\mathbb{R}^n)$ strictly speaking are not functions, but equivalence classes of functions defined almost everywhere. It would therefore not be strictly correct to call $H^{s+k}(\mathbb{R}^n)$ a subset of $C^k(\mathbb{R}^n)$, but the words "there exists a continuous linear inclusion" should be taken to mean that for each of the equivalence classes $[f] \in H^{s+k}(\mathbb{R}^n)$, there is a (necessarily unique) representative f that is a function $\mathbb{R}^n \to V$ of class C^k, and the map $H^{s+k}(\mathbb{R}^n) \to C^k(\mathbb{R}^n) : [f] \mapsto f$ defined in this way is linear, injective, and continuous. Recall that for a linear map $A : X \to Y$ between two vector spaces with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$ respectively, continuity is equivalent to being *bounded*, meaning that A satisfies the estimate $\|Ax\|_Y \leqslant c\|x\|_X$ for some constant $c > 0$ independent of $x \in X$. The continuous inclusions $H^{s+k}(\mathbb{R}^n) \hookrightarrow C^k(\mathbb{R}^n)$ thus come with estimates of the form

$$\|f\|_{C^k} \leqslant c\|f\|_{H^{s+k}},$$

and proving such an estimate is equivalent to proving that the inclusion is continuous.

PROOF OF THEOREM 54.16. We first consider functions $f \in H^s(\mathbb{R}^n)$ with $2s > n$. The main step is to establish a bound on $\|\hat{f}\|_{L^1}$, as $f = \mathscr{F}^*(\hat{f})$ is then equal almost everywhere to a function in $C^0(\mathbb{R}^n)$ since the integral formula in (54.3) defines a continuous linear map $\mathscr{F}^* : L^1(\mathbb{R}^n) \to C^0(\mathbb{R}^n)$. We use the Cauchy-Schwarz inequality:

$$\begin{aligned}\|\hat{f}\|_{L^1} &= \int_{\mathbb{R}^n} \frac{1}{(1+|p|^2)^{s/2}} \cdot \left|(1+|p|^2)^{s/2}\hat{f}(p)\right| dp \leqslant \left\|\frac{1}{(1+|p|^2)^{s/2}}\right\|_{L^2} \cdot \left\|(1+|p|^2)^{s/2}\hat{f}\right\|_{L^2} \\ &= \left(\int_{\mathbb{R}^n} \frac{1}{(1+|p|^2)^s}\, dp\right)^{1/2} \cdot \|f\|_{H^s}\end{aligned}$$

Using n-dimensional polar coordinates, we see that the integral in the second line converges if and only if $\int_1^\infty \frac{r^{n-1}}{(1+r^2)^s}\,dr < \infty$. For large $r > 0$, the latter integrand behaves like $r^{n-1}/r^{2s} = r^{n-2s-1}$, so the integral converges if and only if $n - 2s < 0$, which is exactly the condition $2s > n$. This proves the continuous inclusion of $H^s(\mathbb{R}^n)$ into $C^0(\mathbb{R}^n)$.

If $f \in H^{s+k}(\mathbb{R}^n)$ with $k \in \mathbb{N}$, then the same argument bounds the L^1-norm of the function $p \mapsto p^\alpha \widehat{f}(p)$ for each multi-index α with $|\alpha| \leqslant k$ in terms of $\|f\|_{H^{s+k}}$. The same calculation that is behind the relations (54.4) implies in this situation that the partial derivatives $\partial^\alpha f$ up to order k exist and are continuous. Moreover, their C^0-norms are bounded in terms of the L^1-norm of $p^\alpha \widehat{f}$, which gives a bound for $\|f\|_{C^k}$ in terms of $\|f\|_{H^{s+k}}$. □

Corollary 54.17. *Any function belonging to $H^s(\mathbb{R}^n)$ for every $s \geqslant 0$ is smooth. In fact, a compactly supported function is smooth if and only if it belongs to $H^s(\mathbb{R}^n)$ for every $s \geqslant 0$.* □

Remark 54.18. Elsewhere in the literature, you will sometimes find the spaces $H^s(\mathbb{R}^n)$ denoted by $W^{s,2}(\mathbb{R}^n)$, $L^{s,2}(\mathbb{R}^n)$, $L^2_s(\mathbb{R}^n)$ or other variations. The use of the letter H (which stands for "Hilbert") is perhaps slightly unfortunate in our context, due to the danger of confusing it with a cohomology group (I have seen that particular misunderstanding arise among experts at conferences), but we will be able to avoid ever using both meanings of this symbol in the same context. As you might guess from the notation $W^{s,2}(\mathbb{R}^n)$, there also exist Sobolev spaces called $W^{k,p}(\mathbb{R}^n)$, as well as $W^{k,p}(\mathcal{U})$ for an open subset $\mathcal{U} \subset \mathbb{R}^n$, and for $p \neq 2$ these are all Banach spaces, but not Hilbert spaces. For $k \geqslant 0$ an integer and $1 \leqslant p < \infty$, $W^{k,p}(\mathcal{U})$ is most easily defined as the completion with respect to the norm

$$\|f\|_{W^{k,p}(\mathcal{U})} := \sum_{|\alpha| \leqslant k} \|\partial^\alpha f\|_{L^p(\mathcal{U})}$$

of the space of smooth functions f on $\mathcal{U}$ with $\|f\|_{W^{k,p}(\mathcal{U})} < \infty$. This completion forms a linear subspace of $L^p(\mathcal{U})$, and its elements are not generally k-times differentiable in the classical sense, but admit a notion of *weak* derivatives up to order k which are also L^p-functions. In this section we used the Fourier transform to define the required notion of weak differentiability, but that trick does not work well for L^p-functions with $p \neq 2$ or functions defined on a subset $\mathcal{U} \subset \mathbb{R}^n$ instead of the entirety of $\mathbb{R}^n$. The more generally applicable notion of weak differentiability is defined in terms of integration by parts—we will come back to this in Lecture 56 when we discuss weak regularity results for elliptic equations.

54.4. Regularity with constant coefficients. We are now in a position to demonstrate what the ellipticity condition is good for. We consider the following scenario: $D : \Gamma(E) \to \Gamma(F)$ is a differential operator of order $m \geqslant 1$ between two trivial vector bundles over $\mathbb{R}^n$ with *constant* coefficients, meaning D can be written in the form

$$D = \sum_{|\alpha| \leqslant m} c_\alpha \partial^\alpha : C^\infty(\mathbb{R}^n, \mathbb{F}^k) \to C^\infty(\mathbb{R}^n, \mathbb{F}^\ell) \tag{54.9}$$

for a set of constant matrices $c_\alpha \in \operatorname{Hom}(\mathbb{F}^k, \mathbb{F}^\ell)$. If we assume nothing further about D, then there is no reason to expect solutions of the equation $D\eta = 0$ to be smooth. The equation itself makes sense for any function $\eta : \mathbb{R}^n \to \mathbb{F}^k$ of class C^m, and there are simple examples of operators that admit solutions having m derivatives but no more, e.g. for $n = 2$, the wave equation

$$(\partial_1^2 - \partial_2^2)\eta = 0$$

has solutions of the form $\eta(x^1, x^2) = f(x^1 \pm x^2)$ for arbitrary C^2-functions $f : \mathbb{R} \to \mathbb{R}$. For elliptic operators, however, this cannot happen:

THEOREM 54.19. *If the operator in* (54.9) *is elliptic and* $\xi : \mathbb{R}^n \to \mathbb{F}^\ell$ *is a smooth function with compact support, then every solution* $\eta \in C^m(\mathbb{R}^n) \cap L^2(\mathbb{R}^n)$ *to the equation* $D\eta = \xi$ *is smooth.*

REMARK 54.20. The assumption that D is elliptic implies $k = \ell$, since otherwise no linear transformation $\mathbb{F}^k \to \mathbb{F}^\ell$ can be invertible, but we will see that the proof of Theorem 54.19 works under slightly more general assumptions that also permit $k < \ell$.

The phenemenon in Theorem 54.19 is known as **elliptic regularity**: solutions to elliptic PDEs are generally at least as "regular" as the functions that define the equation. A similar result in the more general context of an elliptic operator with nonconstant coefficients over a compact manifold will play an important role in the proof that such operators have finite-dimensional kernels and cokernels (Theorem 53.32), and it is also the reason why that result can be stated without mentioning any non-smooth objects.

Before proving Theorem 54.19, it is instructive to take a quick look at the one case in which the proof is easy: suppose $n = 1$. Our solutions in this case are functions $\eta(t)$ of a single variable, so the linear partial differential equation $D\eta = \xi$ with constant coefficients is actually an *ordinary* differential equation of the form

$$c_m \partial_t^m \eta + c_{m-1}\partial_t^{m-1}\eta + \ldots + c_1 \partial_t \eta + c_0 \eta = \xi$$

for constant matrices $c_0, \ldots, c_m \in \mathrm{Hom}(\mathbb{F}^k, \mathbb{F}^\ell)$. The operator D is then elliptic if and only if the matrix c_m is invertible, in which case the equation $D\eta = \xi$ is equivalent to

$$\partial_t^m \eta = c_m^{-1}\left(\xi - c_{m-1}\partial_t^{m-1}\eta - \ldots - c_1 \partial_t \eta - c_0 \eta\right). \tag{54.10}$$

If we now assume ξ is smooth and η is of class C^m, then the right hand side of (54.10) is of class C^1, implying the same for $\partial_t^m \eta$, which means η is actually of class C^{m+1}. Now it follows that the right hand side is of class C^2, thus so is $\partial_t^m \eta$, implying that η is in fact of class C^{m+2}. This process can be repeated indefinitely as long as ξ has more derivatives—if we did not assume ξ to be smooth, we would still deduce from this argument that η has at least m more continuous derivatives than ξ does.

The argument above will not generally work for $n \geqslant 2$, because *partial* differential operators typically do not give the kind of complete information about the highest-order derivatives that we see in (54.10). The inhomogeneous Laplace equation $-\sum_{j=1}^n \partial_j^2 \eta = \xi$, for example, does not provide a way to write any individual second partial derivative of η purely in terms of its lower-order derivatives. The remarkable fact about elliptic operators is that while the information they give about the highest-order derivatives of a solution is generally incomplete, it is still *enough* information to conclude results like Theorem 54.19.

The easiest way to see this is by performing a Fourier transform on both sides of the equation $D\eta = \xi$, and for this purpose, we will assume in much of the following discussion that

$$\mathbb{F} := \mathbb{C},$$

so that all functions can be multiplied by complex scalars and Fourier transforms are well defined. This is not actually a loss of generality: if D is given as a real-linear differential operator $C^\infty(\mathbb{R}^n, \mathbb{R}^k) \to C^\infty(\mathbb{R}^n, \mathbb{R}^\ell)$ with constant coefficients $c_\alpha \in \mathrm{Hom}(\mathbb{R}^k, \mathbb{R}^\ell)$, then it can always be extended to a complex-linear operator $C^\infty(\mathbb{R}^n, \mathbb{C}^k) \to C^\infty(\mathbb{R}^n, \mathbb{C}^\ell)$ by regarding the real matrices c_α as linear transformations $\mathbb{C}^k \to \mathbb{C}^\ell$. It will be clear that all results proved for this complexified operator hold for the original real-linear operator as well.

Since the matrices c_α are constant, (54.4) gives

$$\widehat{D\eta}(p) = \sum_{|\alpha| \leqslant m} c_\alpha \widehat{\partial^\alpha \eta}(p) = \sum_{|\alpha| \leqslant m} c_\alpha (2\pi i p)^\alpha \hat{\eta}(p) =: Q(p)\hat{\eta}(p),$$

where $Q : \mathbb{R}^n \to \mathrm{Hom}(\mathbb{C}^k, \mathbb{C}^\ell)$ is a matrix-valued polynomial whose top-degree term is

$$Q_m(p) := (2\pi i)^m \sum_{|\alpha|=m} p^\alpha c_\alpha.$$

Up to a constant, this is also a coordinate formula for the principal symbol of D at every point in $\mathbb{R}^n$; the symbol is the same at every point since the coefficients are constant. Ellipticity is thus equivalent to the assumption that $Q_m(p)$ is invertible for all $p \in \mathbb{R}^n \backslash \{0\}$. As mentioned in Remark 54.20 above, we will be able to get away with a slightly weaker assumption than this: let us suppose

$$Q_m(p) : \mathbb{C}^k \to \mathbb{C}^\ell \text{ is injective for all } p \in \mathbb{R}^n \backslash \{0\}.$$

This assumption gives rise to certain estimates for $Q(p)$ that can be translated into estimates of Sobolev norms for the functions η and ξ. To start with, any injective linear operator $A : \mathbb{C}^k \to \mathbb{C}^\ell$ satisfies $|Av| \geqslant C|v|$ for some constant $C > 0$ independent of $v \in \mathbb{C}^k$, so $Q_m(p)$ will satisfy such an estimate for every $p \in \mathbb{R}^n$ in the unit sphere, and we can take the same constant $C > 0$ at every point in the sphere since the sphere is compact. Moreover, $Q_m(p)$ is a homogeneous polynomial function of p with degree m, so if we now extend the estimate to allow p in the rest of $\mathbb{R}^n$ instead of just the unit sphere, the constant will scale with $|p|^m$, giving an estimate of the form

$$|Q_m(p)v| \geqslant C|p|^m|v| \qquad \text{for all } p \in \mathbb{R}^n,\ v \in \mathbb{C}^k,$$

with a constant $C > 0$ independent of p and v. Putting back the terms of degree less than m, $Q(p)$ will not satisfy such an estimate for all $p \in \mathbb{R}^n$, but (after increasing the constant $C > 0$ if necessary) it will when $|p|$ is sufficiently large, because the lower-degree terms then become negligible in comparison with Q_m. We therefore have constants $C, R > 0$ such that

$$|Q(p)v| \geqslant C|p|^m|v| \qquad \text{for all } p \in \mathbb{R}^n \backslash B^n_R(0),\ v \in \mathbb{C}^k.$$

After adjusting the constants further, we can also replace the polynomial on the right hand side with the one that appears in the definition of the Sobolev norms, and thus write

$$|Q(p)v|^2 \geqslant C(1 + |p|^2)^m|v|^2 \qquad \text{for all } p \in \mathbb{R}^n \backslash B^n_R(0),\ v \in \mathbb{C}^k. \tag{54.11}$$

Now suppose $\eta \in L^2(\mathbb{R}^n)$, $\xi \in H^s(\mathbb{R}^n)$ for some $s \geqslant 0$, and that the Fourier transform of the equation $D\eta = \xi$ is satisfied almost everywhere, i.e.

$$Q(p)\hat{\eta}(p) = \hat{\xi}(p) \qquad \text{for almost all } p \in \mathbb{R}^n.$$

Using (54.11), we now find

$$\begin{aligned}
\|\eta\|^2_{H^{s+m}} &= \int_{\mathbb{R}^n} (1 + |p|^2)^{s+m}|\hat{\eta}(p)|^2\, dp \\
&= \int_{B^n_R(0)} (1 + |p|^2)^{s+m}|\hat{\eta}(p)|^2\, dp + \int_{\mathbb{R}^n \backslash B^n_R(0)} (1 + |p|^2)^s(1 + |p|^2)^m|\hat{\eta}(p)|^2\, dp \\
&\leqslant (1 + R^2)^{s+m} \int_{B^n_R(0)} |\hat{\eta}(p)|^2\, dp + C_1 \int_{\mathbb{R}^n \backslash B^n_R(0)} (1 + |p|^2)^s|Q(p)\hat{\eta}(p)|^2 \\
&\leqslant C_2\|\eta\|^2_{L^2} + C_2\|\xi\|^2_{H^s},
\end{aligned}$$

where the values of the constants C_1, C_2 are positive and unimportant. Since the norms $|(x, y)| := |x| + |y|$ and $\|(x, y)\| := \sqrt{x^2 + y^2}$ on $\mathbb{R}^2$ are equivalent, we can find a new constant $C > 0$ such that the estimate obtained above gets rewritten as[114]

$$\|\eta\|_{H^{s+m}} \leqslant C\|D\eta\|_{H^s} + C\|\eta\|_{L^2}, \tag{54.12}$$

which is sometimes called a *fundamental elliptic estimate* for the operator D. It is a surprising result if you think about the meanings of the various norms involved: on the left hand side, the H^{s+m}-norm measures the L^2-norms of all partial derivatives of η up to order $s+m$. Some of these derivatives also appear on the right hand side, as the operator $D\eta$ involves derivatives up to order m, thus $\|D\eta\|_{H^s}$ measures the L^2-norms of certain linear combinations of partial derivatives up to order $s+m$, but for a typical operator D, it does not actually measure any of those derivatives individually. The magic of ellipticity is that the particular linear combinations measured by $\|D\eta\|_{H^s}$ nonetheless give complete control over $\|\eta\|_{H^{s+m}}$. As we will see, most of the important properties of elliptic operators follow from estimates of this type.

The estimate (54.12) is valid for any $\eta \in L^2(\mathbb{R}^n)$, so long as the expression $\|D\eta\|_{H^s}$ is suitably interpreted: let us say for this purpose that $\eta \in L^2(\mathbb{R}^n)$ is a **weak solution** to the equation $D\eta = \xi$ for some $\xi \in L^2(\mathbb{R}^n)$ if $Q(p)\hat{\eta}(p) = \hat{\xi}(p)$ for almost every $p \in \mathbb{R}^n$. If no function ξ with this property exists, e.g. because $Q\hat{\eta} \notin L^2(\mathbb{R}^n)$, then the norm

$$\|D\eta\|_{H^s} := \|(1+|p|^2)^{s/2} Q\hat{\eta}\|_{L^2}$$

will be infinite and the estimate (54.12) is thus free of content. But if $D\eta$ in this generalized sense happens to belong to one of the Sobolev spaces $H^s(\mathbb{R}^n)$, we obtain a more technical variant of Theorem 54.19:

Proposition 54.21. *For an elliptic operator $D : C^\infty(\mathbb{R}^n, \mathbb{F}^k) \to C^\infty(\mathbb{R}^n, \mathbb{F}^\ell)$ of order $m \geqslant 1$ with constant coefficients, if $\eta \in L^2(\mathbb{R}^n)$ is a weak solution to the equation $D\eta = \xi$ with $\xi \in H^s(\mathbb{R}^n)$ for some $s \geqslant 0$, then $\eta \in H^{s+m}(\mathbb{R}^n)$.* □

Proof of Theorem 54.19. The assumption that ξ is smooth with compact support implies that it belongs to $H^s(\mathbb{R}^n)$ for every $s \geqslant 0$, so if $\eta \in L^2(\mathbb{R}^n)$ is a weak solution to $D\eta = \xi$, we can choose $s > 0$ arbitrarily large and conclude from Proposition 54.21 that $\eta \in H^{s+m}(\mathbb{R}^n)$, proving that η also belongs to $H^s(\mathbb{R}^n)$ for all $s \geqslant 0$. By the Sobolev embedding theorem (Theorem 54.16), η is therefore smooth. This conclusion applies in particular if η is also assumed in the first place to be of class C^m and a solution to $D\eta = \xi$ in the classical sense. □

Remark 54.22. Instead of ellipticity, the hypothesis we actually used in the proofs of Proposition 54.21 and Theorem 54.19 was that

$$\sigma_D(\lambda) : E_p \to F_p \text{ is injective for all } \lambda \neq 0 \in T_p^*M,\ p \in M, \tag{54.13}$$

which is implied by ellipticity but is a weaker condition in general. It is also satisfied for instance by covariant derivatives $\nabla : \Gamma(E) \to \Omega^1(M, E)$ and Cauchy-Riemann type operators $D : \Gamma(E) \to \Gamma(\overline{\mathrm{Hom}}_{\mathbb{C}}(TM, E))$ on bundles $E \to M$ over a real or complex manifold of arbitrary dimension n, whereas these operators are only elliptic for $n = 1$ (see Exercises 53.33 and 54.13). These operators therefore satisfy corresponding regularity results: parallel sections of class C^1 of a vector bundle $E \to M$ with a connection are always smooth, and so are holomorphic sections of a holomorphic vector bundle $E \to M$ over a complex manifold, which can always be characterized as solutions to $D\eta = 0$ for a Cauchy-Riemann type operator D. One can also show that over closed manifolds, operators satisfying (54.13) have finite-dimensional kernels. What we lose if we only assume (54.13) instead of ellipticity is the second half of Theorem 53.32, the fact that coker(D) is also finite

[114] In lecture and in the first version of these notes I stated a weaker version of the estimate (54.12) in which $\|\eta\|_{H^s}$ appeared on the right hand side instead of $\|\eta\|_{L^2}$. The stronger estimate follows however from more-or-less the same proof, and is easier to apply.

dimensional, which guarantees that the inhomogeneous equation $D\eta = \xi$ can also be solved for sections ξ satisfying a finite set of conditions. Without this, one must generically expect the equation $D\eta = \xi$ to have no solutions; equations with this property are called *overdetermined.* On a technical level, what one gains by assuming $\sigma_D(\lambda)$ is not only injective but also surjective is that the symbol of the *formal adjoint* then satisfies the same condition: ellipticity is equivalent to the condition that *both* D and D^* satisfy (54.13). This fact will be used in the proof that $\dim \operatorname{coker}(D) < \infty$ for elliptic operators.

55. Proving finite dimensionality

Our goal in this lecture is to prove the first half of Theorem 53.32, the statement that the space of solutions to $D\eta = 0$ for an elliptic operator D over a closed manifold is always finite dimensional. In the process we will also prove a more technical result about the image of D that is an important prerequisite for showing $\dim \operatorname{coker}(D) < \infty$. The proof of the latter will be completed in the next lecture, where we shall also use it to derive the Hodge decomposition theorem.

55.1. Fredholm and compact operators. Throughout this section, X, Y, Z are Banach spaces, whose norms will be denoted either by $\|\cdot\|$ or by $\|\cdot\|_X$, $\|\cdot\|_Y$ etc. when we want to be extra clear about which is which. We denote by

$$\mathscr{L}(X, Y)$$

the Banach space of continuous linear operators $T : X \to Y$, which carries the **operator norm**

$$\|T\| := \sup_{x \in X \setminus \{0\}} \frac{\|Tx\|_Y}{\|x\|_X}.$$

For readers not accustomed to functional analysis, the following should be pointed out explicitly: every finite-dimensional vector space has a canonical topology for which linear subspaces are automatically closed subsets, but in infinite dimensions, subspaces need not be closed. For example, the space of compactly-supported smooth functions $C_0^\infty(\mathbb{R}^n)$ is dense in $L^2(\mathbb{R}^n)$, and thus forms a proper linear subspace whose closure is the whole space. The quotient of a Banach space X by a subspace $V \subset X$ has a natural Banach space structure if and only if V is closed in X.

Definition 55.1. An operator $T \in \mathscr{L}(X, Y)$ is **Fredholm** if there exist splittings of X and Y into direct sums of closed linear subspaces

$$X = V \oplus \ker(T), \qquad Y = \operatorname{im}(T) \oplus W$$

such that $\ker(T)$ and $W \cong \operatorname{coker}(T) = Y/\operatorname{im}(T)$ are both finite dimensional. The **Fredholm index** of T in this case is the integer

$$\operatorname{ind}(T) := \dim \ker(T) - \dim \operatorname{coker}(T).$$

Remark 55.2. One can show that $T \in \mathscr{L}(X, Y)$ is Fredholm if and only if $\dim \ker(T)$ and $\dim \operatorname{coker}(T)$ are both finite, i.e. the existence of the splittings in Definition 55.1 follows from these conditions automatically. The least obvious detail in this statement is that $\operatorname{im}(T)$ is a closed subspace, but in practice, most proofs that an operator T is Fredholm (including the one we will give for elliptic operators) involve at some step an explicit proof that $\operatorname{im}(T)$ is closed. For this reason, there is no need to worry about the distinction between Definition 55.1 and the seemingly weaker (but actually equivalent) definition found in some books.

The technical version of Theorem 53.32 on elliptic operators $D : \Gamma(E) \to \Gamma(F)$ over a closed manifold M is that for suitable Hilbert space completions $H^{k+m}(E)$ of $\Gamma(E)$ and $H^k(F)$ of $\Gamma(F)$, D defines a Fredholm operator from $H^{k+m}(E)$ to $H^k(F)$. The spaces $H^k(E)$ will be defined in §55.4, later in this lecture. The first step in the proof that $D : H^{k+m}(E) \to H^k(F)$ is Fredholm will then be to prove that its kernel is finite dimensional. The argument for this uses the following characterization of finite dimensionality:

LEMMA 55.3. *A Banach space X is finite dimensional if and only if closed and bounded subsets of X are compact, or equivalently, every bounded sequence in X has a convergent subsequence.* $\square$

One direction of Lemma 55.3 is a basic result that you hopefully learned in first-year analysis, often called the Bolzano-Weierstrass theorem. The other direction is fairly obvious in the setting of Hilbert spaces, which is the case we will use: in an infinite-dimensional Hilbert space, any countably infinite orthonormal set forms a bounded sequence that can have no convergent subsequence, thus there exist closed and bounded sets that are not compact.

DEFINITION 55.4. An operator $T \in \mathscr{L}(X, Y)$ is **compact** if for every bounded sequence $x_k \in X$, the sequence $Tx_k \in Y$ has a convergent subsequence, or equivalently, T maps every closed and bounded subset of X to a compact subset of Y.

In finite dimensions every linear map is compact, but Lemma 55.3 shows that for instance the identity map on an infinite-dimensional Banach space is not compact. The compact operators that we will be most interested in are natural inclusions of one Banach space into another.

EXAMPLE 55.5. For a bounded open subset $\mathcal{U} \subset \mathbb{R}^n$, the space $C^k(\mathcal{U})$ of C^k-functions on $\mathcal{U}$ with bounded derivatives up to order k admits a natural inclusion

$$C^k(\mathcal{U}) \hookrightarrow C^{k-1}(\mathcal{U})$$

for each $k \geqslant 1$, and this inclusion is compact. Indeed, for $k = 1$, any bounded sequence in $C^1(\mathcal{U})$ is both uniformly bounded and equicontinuous due to the bound on its first derivatives, so the Arzelà-Ascoli theorem guarantees the existence of a C^0-convergent subsequence. Applying the same argument to the partial derivatives up to order $k-1$ proves that any uniformly C^k-bounded sequence has a C^{k-1}-convergent subsequence. Note that the boundedness of the domain $\mathcal{U} \subset \mathbb{R}^n$ is crucial for the applicability of the Arzelà-Ascoli theorem, and the result is indeed false on unbounded domains: on $\mathbb{R}^n$ for instance, one can easily cook up C^1-bounded sequences without C^0-convergent subsequences just be composing a single function with a sequence of translations on $\mathbb{R}^n$ moving outward toward infinity.

Here is a useful functional-analytic tool for the first step in proving that an operator is Fredholm.

PROPOSITION 55.6. *Suppose $T : X \to Y$ and $K : X \to Z$ are continuous linear operators, K is compact, and there is an estimate of the form*

$$\|x\|_X \leqslant c\|Tx\|_Y + c\|Kx\|_Z \tag{55.1}$$

for some constant $c > 0$ independent of $x \in X$. Then $\ker(T) \subset X$ is finite dimensional and $\operatorname{im}(T) \subset Y$ is closed.

REMARK 55.7. Lest the hypothesis (55.1) should strike you as coming out of nowhere, I remind you that in the previous lecture we proved an estimate for elliptic operators D of order $m \in \mathbb{N}$ on $\mathbb{R}^n$ with constant coefficients that said

$$\|\eta\|_{H^{s+m}} \leqslant c\|D\eta\|_{H^s} + c\|\eta\|_{L^2}.$$

This has *almost* the same form as (55.1), with $D : H^{s+m}(\mathbb{R}^n) \to H^s(\mathbb{R}^n)$ in the role of $T : X \to Y$ and the inclusion $H^{s+m}(\mathbb{R}^n) \hookrightarrow L^2(\mathbb{R}^n)$ for $K : X \to Z$. That inclusion is defined by taking functions that have a certain number of derivatives (in a generalized sense) and forgetting some of them, just as with the inclusion $C^k(\mathcal{U}) \hookrightarrow C^{k-1}(\mathcal{U})$ in Example 55.5, which was compact if $\mathcal{U} \subset \mathbb{R}^n$ is bounded, due to the Arzelà-Ascoli theorem. The only trouble is that since $\mathbb{R}^n$ is not bounded, the inclusion $H^{s+m}(\mathbb{R}^n) \hookrightarrow L^2(\mathbb{R}^n)$ fails to be compact for the same reason that $C^k(\mathbb{R}^n) \hookrightarrow C^{k-1}(\mathbb{R}^n)$ does; we will rectify this in §55.2 below.

Proof of Proposition 55.6. To show that $\dim \ker(T) < \infty$, we apply Lemma 55.3 by proving that every sequence $x_k \in \ker(T)$ with $\|x_k\|_X$ bounded has a convergent subsequence. Indeed, since $K : X \to Z$ is compact, any sequence can be replaced with a subsequence (still denoted by x_k) so that the sequence $Kx_k \in Z$ converges, meaning in particular that Kx_k is a Cauchy sequence. The hypothesis (55.1) then gives

$$\|x_j - x_k\|_X \leqslant c\|Kx_j - Kx_k\|_Z,$$

implying that $x_k \in X$ is also a Cauchy sequence, so it converges.

To study $\mathrm{im}(T) \subset Y$, we observe that the image of T does not change if we replace T with the map $X/\ker(T) \to Y : [x] \mapsto Tx$, which is also a continuous linear map between Banach spaces since the subspace $\ker(T)$ is in the mean time known to be closed (because it's finite dimensional). We can therefore assume without loss of generality that $T : X \to Y$ is injective, and consider a sequence $x_k \in X$ such that $Tx_k \in Y$ converges to some element $y \in Y$. There are two cases to consider.

Case 1: $\|x_k\|$ has a bounded subsequence.
By the compactness of K, we can then replace x_k with a subsequence such that $Kx_k \in Z$ converges, so both Kx_k and Tx_k are Cauchy sequences, and the estimate

$$\|x_j - x_k\|_X \leqslant c\|Tx_j - Tx_k\|_Y + c\|Kx_j - Kx_k\|_Z,$$

from (55.1) then implies that x_k is as well. It follows that x_k converges to some element $x \in X$, and since T is continuous, $Tx = y$, proving that $\mathrm{im}(T) \subset Y$ is closed.

Case 2: $\|x_k\| \to \infty$.
Consider the bounded sequence $x'_k := x_k/\|x_k\| \in X$, which satisfies $Tx'_k = \frac{1}{\|x_k\|}Tx_k \to 0$ since $Tx_k \to y$ and $\|x_k\| \to \infty$. By the compactness of K, we can replace this with a subsequence such that Kx'_k converges, and (55.1) then implies as in the previous case that x'_k also converges to some element $x' \in X$, which necessarily has $\|x'\| = 1$. But the continuity of T then implies $Tx' = 0$, so x' is a nontrivial element of $\ker(T)$, which is a contradiction. □

55.2. The Rellich-Kondrashov compactness theorem for $H_0^s(\mathcal{U})$. The strategy we've mapped out for proving $\dim\ker(D) < \infty$ requires a compact operator, and a candidate for such an operator that is relevant to our setting was mentioned in Remark 55.7: it is the inclusion $H^{s+m} \hookrightarrow L^2$. Bounded sequences in $H^{s+m}(\mathbb{R}^n)$, however, need not have L^2-convergent subsequences, as shown by the example of a sequence of the form $f_k(x) := f(x + v_k)$ for a fixed nontrivial function $f \in H^{s+m}(\mathbb{R}^n)$ and translations $v_k \in \mathbb{R}^n$ tending to infinity. The compact inclusion $C^k(\mathcal{U}) \hookrightarrow C^{k-1}(\mathcal{U})$ of Example 55.5, in which compactness follows from the Arzelà-Ascoli theorem, suggests that one might have more success considering functions defined only on a bounded domain $\mathcal{U} \subset \mathbb{R}^n$.

Definition 55.8. For an open subset $\mathcal{U} \subset \mathbb{R}^n$ and $s \geqslant 0$, the Hilbert space

$$H_0^s(\mathcal{U}) \subset H^s(\mathbb{R}^n)$$

is defined as the closure in $H^s(\mathbb{R}^n)$ of the space $C_0^\infty(\mathcal{U})$ of C^∞-functions with compact support in $\mathcal{U}$. In light of the injective map

$$H_0^s(\mathcal{U}) \hookrightarrow L^2(\mathcal{U}) : f \mapsto f|_{\mathcal{U}},$$

we shall often regard elements of $H_0^s(\mathcal{U})$ as functions defined on $\mathcal{U}$ that admit trivial extensions of class H^s over $\mathbb{R}^n$.

Note that since $C_0^\infty(\mathcal{U})$ is dense in $L^2(\mathcal{U})$,

$$H_0^0(\mathcal{U}) = L^2(\mathcal{U}),$$

and there are obvious inclusions

$$H_0^s(\mathcal{U}) \hookrightarrow H_0^t(\mathcal{U}) \qquad \text{whenever } s > t.$$

The main result about Sobolev inclusions of this type is known as the Rellich-Kondrashov compactness theorem, often abbreviated simply as Rellich's theorem. You should think of it as a Sobolev analogue of what the Arzelà-Ascoli theorem gives us for the inclusion $C^k(\mathcal{U}) \hookrightarrow C^{k-1}(\mathcal{U})$: if we bound derivatives up to a certain order over a bounded domain but then *forget* the highest-order derivatives, boundedness becomes compactness.

THEOREM 55.9 (Rellich-Kondrashov for $H^s(\mathcal{U})$). *If $\mathcal{U} \subset \mathbb{R}^n$ is a bounded open subset, then the inclusion $H_0^s(\mathcal{U}) \hookrightarrow H_0^t(\mathcal{U})$ is a compact operator for every $s > t \geqslant 0$.*

We shall prove Theorem 55.9 in three steps. The first is the observation that the set of compact operators from X to Y is a closed subset of $\mathscr{L}(X, Y)$:

EXERCISE 55.10. Prove that if $A_k \in \mathscr{L}(X, Y)$ is a sequence of compact operators with $\|A_k - A\| \to 0$ for some $A \in \mathscr{L}(X, Y)$, then A is also compact.
Hint: Prove that if $x_j \in X$ is a sequence for which $A_k x_j \in Y$ is a convergent sequence for every fixed k, then Ax_j is a Cauchy sequence in Y.

The second step is to define a family of compact operators $j^R : H_0^s(\mathcal{U}) \to H^t(\mathbb{R}^n)$ for $R > 0$ that can be shown in the third step to converge as $R \to \infty$ to the inclusion $H_0^s(\mathcal{U}) \hookrightarrow H^t(\mathbb{R}^n)$. The compactness of the operators j^R will depend on the assumption that $\mathcal{U} \subset \mathbb{R}^n$ is bounded, but it will not require $s > t$. Given $R > 0$, define a map $L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n) : f \mapsto f^R$ such that the Fourier transform of f^R is

$$\widehat{f^R}(p) := \begin{cases} \hat{f}(p) & \text{if } |p| \leqslant R, \\ 0 & \text{otherwise.} \end{cases}$$

The new function f^R belongs to $H^t(\mathbb{R}^n)$ for every $t \geqslant 0$, and $f \mapsto f^R$ is a continuous linear operator $L^2(\mathbb{R}^n) \to H^t(\mathbb{R}^n)$ since

$$\begin{aligned} \|f^R\|_{H^t}^2 &= \int_{\mathbb{R}^n} (1 + |p|^2)^t |\widehat{f^R}(p)|^2 \, dp = \int_{B_R^n(0)} (1 + |p|^2)^t |\hat{f}(p)|^2 \, dp \\ &\leqslant (1 + R^2)^t \int_{B_R^n(0)} |\hat{f}(p)|^2 \, dp \leqslant (1 + R^2)^t \|\hat{f}\|_{L^2}^2 = (1 + R^2)^t \|f\|_{L^2}^2. \end{aligned}$$

Composing it with the inclusion $H^s(\mathbb{R}^n) \hookrightarrow L^2(\mathbb{R}^n)$ for any $s \geqslant 0$ allows us also to view $f \mapsto f^R$ as a continuous operator $H^s(\mathbb{R}^n) \to H^t(\mathbb{R}^n)$, which can then be restricted to the subspace $H_0^s(\mathcal{U}) \subset H^s(\mathbb{R}^n)$ for any open set $\mathcal{U} \subset \mathbb{R}^n$, though there is no reason to expect the image of this restriction to lie in $H_0^t(\mathcal{U})$.

LEMMA 55.11. *For any bounded open subset $\mathcal{U} \subset \mathbb{R}^n$ and any $s, t \geqslant 0$ and $R > 0$, the operator $H_0^s(\mathcal{U}) \to H^t(\mathbb{R}^n) : f \mapsto f^R$ is compact.*

Our proof of this lemma will require a standard result in the theory of L^p-spaces, the Banach-Alaoglu theorem. For L^2-functions in particular on a measurable subset $\mathcal{U} \subset \mathbb{R}^n$, a sequence $f_k \in L^2(\mathcal{U})$ is called **weakly convergent** to a function $f \in L^2(\mathcal{U})$, written

$$f_k \overset{L^2}{\rightharpoonup} f,$$

if for every fixed $g \in L^2(\mathcal{U})$, $\langle g, f_k\rangle_{L^2} \to \langle g, f\rangle_{L^2}$. Another way of saying this is that $f_k \rightharpoonup f$ means $\phi(f_k) \to \phi(f)$ for every continuous linear functional $\phi : L^2(\mathcal{U}) \to \mathbb{C}$. The analogous condition for a sequence x_k in $\mathbb{R}^n$ would imply that the individual coordinates of x_k each converge, since coordinate functions are examples of continuous linear functionals on $\mathbb{R}^n$, so weak convergence in finite dimensions just means convergence in the usual topology. In $L^2(\mathcal{U})$, however, it is a strictly weaker condition, and thus more easily satisfied:

Theorem 55.12 (Banach-Alaoglu for L^2). *Every bounded sequence in $L^2(\mathcal{U})$ has a weakly convergent subsequence.*

Proof sketch. Given a bounded sequence $f_k \in L^2(\mathcal{U})$, the sequence $\langle g, f_k\rangle_{L^2} \in \mathbb{R}$ is bounded for each $g \in L^2(\mathcal{U})$ and thus has a convergent subsequence. For any countable set $X \subset L^2(\mathcal{U})$, the Cantor diagonal trick can then be used to replace f_k with a subsequence so that $\langle g, f_k\rangle_{L^2}$ converges for every $g \in X$, and it follows from this that $\langle g, f_k\rangle_{L^2}$ also converges for all g in the L^2-closure of X. The result then follows from the fact that $L^2(\mathcal{U})$ admits a countable dense subset, i.e. it is separable. To construct such a subset over $\mathbb{R}^n$, one can for instance consider all finite sums of functions that take constant values in some fixed countable dense subset of V on sets of the form $[a_1, b_1] \times \ldots \times [a_n, b_n] \subset \mathbb{R}^n$ for $a_j, b_j \in \mathbb{Q}$ and vanish everywhere else. The restrictions of such functions to $\mathcal{U}$ form a countable dense subset of $L^2(\mathcal{U})$. □

Proof of Lemma 55.11. A bounded sequence f_k in $H_0^s(\mathcal{U})$ is also bounded in $L^2(\mathcal{U})$, so by the Banach-Alaoglu theorem, we can replace it with a subsequence such that f_k is weakly convergent in $L^2(\mathcal{U})$ to some $f_\infty \in L^2(\mathcal{U})$. This has the following consequence: since $\mathcal{U}$ is bounded, the complex exponential $x \mapsto e^{2\pi i p\cdot x}$ is an L^2-function on $\mathcal{U}$ for each $p \in \mathbb{R}^n$, and it follows that the Fourier transforms of f_k converge pointwise:

$$\widehat{f}_k(p) = \int_{\mathbb{R}^n} e^{-2\pi i p\cdot x} f_k(x)\, dx = \int_{\mathcal{U}} e^{-2\pi i p\cdot x} f_k(x)\, dx \to \int_{\mathcal{U}} e^{-2\pi i p\cdot x} f_\infty(x)\, dx = \widehat{f}_\infty(p).$$

Here we are also implicitly using the fact that, again since $\mathcal{U}$ is bounded, the L^2-functions f_k and f_∞ are also in L^1, so their Fourier transforms can be computed from the usual integral formula and are continuous functions of $p \in \mathbb{R}^n$. They also satisfy a uniform C^0-bound, because by the Cauchy-Schwarz inequality,

$$\|f_k\|_{L^1} \leqslant \|1\|_{L^2(\mathcal{U})} \cdot \|f_k\|_{L^2}$$

is bounded, and $|\widehat{f}_k(p)| \leqslant \|f_k\|_{L^1}$ for all k and $p \in \mathbb{R}^n$.

Having reduced $f_k \in H_0^s(\mathcal{U})$ to a weakly L^2-convergent subsequence, we claim that for each $R > 0$ and $t \geqslant 0$, the corresponding Fourier truncations f_k^R form a Cauchy sequence in $H^t(\mathbb{R}^n)$. Indeed, taking j, k large, we have

$$\begin{aligned}\|f_j^R - f_k^R\|_{H^t}^2 &= \int_{\mathbb{R}^n} (1 + |p|^2)^t |\widehat{f_j^R}(p) - \widehat{f_k^R}(p)|^2\, dp = \int_{B_R^n(0)} (1 + |p|^2)^t |\widehat{f}_j(p) - \widehat{f}_k(p)|^2\, dp \\ &\leqslant (1 + R^2)^t \int_{B_R^n(0)} |\widehat{f}_j(p) - \widehat{f}_k(p)|^2\, dp.\end{aligned}$$

For any sequence of values for j and k tending to ∞, the integrands in this last expression are uniformly bounded and converge pointwise to 0, so by the dominated convergence theorem, the integrals also converge to 0. □

The third step does not require working on a bounded domain, but here is where it becomes essential to assume $s > t$.

LEMMA 55.13. *For $s, t \geqslant 0$ and $R > 0$, let $j^R : H^s(\mathbb{R}^n) \to H^t(\mathbb{R}^n)$ denote the map $f \mapsto f^R$. If $s > t$, then j^R converges in the operator norm on $\mathscr{L}(H^s(\mathbb{R}^n), H^t(\mathbb{R}^n))$ to the inclusion $H^s(\mathbb{R}^n) \hookrightarrow H^t(\mathbb{R}^n)$ as $R \to \infty$.*

PROOF. Let $j : H^s(\mathbb{R}^n) \hookrightarrow H^t(\mathbb{R}^n)$ denote the inclusion. Given $f \in H^s(\mathbb{R}^n)$, we have

$$\begin{aligned}\|jf - j^R f\|_{H^t}^2 = \|f - f^R\|_{H^t}^2 &= \int_{\mathbb{R}^n} (1 + |p|^2)^t |\hat{f}(p) - \widehat{f^R}(p)|^2 \, dp \\ &= \int_{\mathbb{R}^n \setminus B_R^n(0)} (1 + |p|^2)^t |\hat{f}(p)|^2 \, dp = \int_{\mathbb{R}^n \setminus B_R^n(0)} \frac{(1 + |p|^2)^s}{(1 + |p|^2)^{s-t}} |\hat{f}(p)|^2 \, dp \\ &\leqslant \frac{1}{(1 + R^2)^{s-t}} \int_{\mathbb{R}^n \setminus B_R^n(0)} (1 + |p|^2)^s |\hat{f}(p)|^2 \, dp \leqslant \frac{1}{(1 + R^2)^{s-t}} \|f\|_{H^s}^2,\end{aligned}$$

thus $\|j - j^R\| \leqslant \frac{1}{(1+R^2)^{\frac{s-t}{2}}} \to 0$ as $R \to \infty$. □

PROOF OF THEOREM 55.9. The map $i^R : H_0^s(\mathcal{U}) \to H^t(\mathbb{R}^n) : f \mapsto f^R$ is compact by Lemma 55.11, and it is also the composition of the inclusion $H_0^s(\mathcal{U}) \hookrightarrow H^s(\mathbb{R}^n)$ with the map $j^R : H^s(\mathbb{R}^n) \to H^t(\mathbb{R}^n)$ from Lemma 55.13, and therefore converges as $R \to \infty$ to the inclusion $H_0^s(\mathcal{U}) \hookrightarrow H^t(\mathbb{R}^n)$, whose image is in $H_0^t(\mathcal{U})$. Exercise 55.10 then implies that this limiting map is compact. □

REMARK 55.14. If you remember the result in Theorem 54.14 about Fourier transforms of convolutions, you may notice that for $R > 0$, the function $f^R : \mathbb{R}^n \to V$ could also have been defined as $f^R = \rho^R * f$, where $\rho^R : \mathbb{R}^n \to \mathbb{C}$ is the function whose Fourier transform is the characteristic function of $B_R^n(0)$, i.e.

$$\rho^R(x) := \int_{B_R^n(0)} e^{2\pi i p \cdot x} \, dp.$$

This family of functions forms an approximate identity, thus f^R can be viewed as a family of smoothings of f that converge to f as $R \to \infty$. A very similar argument gives the compactness of the inclusion $H^s(\mathbb{T}^n) \hookrightarrow H^t(\mathbb{T}^n)$ for Sobolev spaces on the torus defined in terms of Fourier series. The smoothing operator $f \mapsto f^R$ in that case can be defined analogously by truncating the Fourier series, and its compactness is then easier to prove because truncated Fourier series are finite, and operators of finite rank are always compact.

55.3. Elliptic estimates with nonconstant coefficients. The main task in proving that elliptic operators on closed manifolds have finite-dimensional kernel will now be to generalize the estimate $\|\eta\|_{H^{s+m}} \leqslant C\|D\eta\|_{H^s} + C\|\eta\|_{L^2}$, which was established in §54.4 for elliptic operators on trivial bundles over $\mathbb{R}^n$ with constant coefficients. The idea will be to formulate such an estimate in a context where the second term on the right hand side can be understood as a compact inclusion, thus establishing the hypothesis of Proposition 55.6. We must first relax the condition that D has constant coefficients.

The Fourier transform is not a convenient tool to use for functions defined on domains other than $\mathbb{R}^n$, so it will be useful to adjust our perspective on the Sobolev norms. Recall that for $k \geqslant 0$ an integer and $f \in \mathscr{S}(\mathbb{R}^n)$, our usual definition of $\|f\|_{H^k}$ is equivalent to a norm that measures the L^2-norms of all partial derivatives of f up to order k. The latter can also be defined for arbitrary smooth functions defined only on an open subset $\mathcal{U} \subset \mathbb{R}^n$, and we will thus set

$$\|f\|_{H^k(\mathcal{U})} := \sum_{|\alpha| \leqslant k} \|\partial^\alpha f\|_{L^2(\mathcal{U})} := \sum_{|\alpha| \leqslant k} \left(\int_{\mathcal{U}} |\partial^\alpha f(x)|^2 \, dx \right)^{1/2} \tag{55.2}$$

in this situation. If f belongs to $H^k_0(\mathcal{U})$, then this definition of the H^k-norm of f is different from but equivalent (in the sense of defining the same topology) to our usual definition, and it has the advantage that for any open subset $\mathcal{V} \subset \mathcal{U}$, we can also define $\|f\|_{H^k(\mathcal{V})}$ for functions that are not compactly supported in $\mathcal{V}$, and write $\|f\|_{H^k(\mathcal{V})} \leqslant \|f\|_{H^k(\mathcal{U})}$. One can also define a Hilbert space $H^k(\mathcal{U})$ on which (55.2) is the natural norm: the quickest way is to take the subspace of $L^2(\mathcal{U})$ obtained as the closure in the H^k-norm of the space of smooth functions f on $\mathcal{U}$ with $\|f\|_{H^k(\mathcal{U})} < \infty$. This space is typically larger than $H^k_0(\mathcal{U})$, but the latter space will suffice for our purposes, and we will only refer to $\|f\|_{H^k(\mathcal{U})}$ in cases where f is smooth, so that the meaning of $\partial^\alpha f$ is clear.

Lemma 55.15. *Suppose $\mathcal{U} \subset \mathbb{R}^n$ is an open subset and $D = \sum_{|\alpha|\leqslant m} c_\alpha \partial^\alpha : C^\infty(\mathcal{U}, \mathbb{F}^k) \to C^\infty(\mathcal{U}, \mathbb{F}^\ell)$ is an elliptic differential operator of order $m \in \mathbb{N}$ between trivial vector bundles over $\mathcal{U}$. Then for each integer $s \geqslant 0$, every point $p \in \mathcal{U}$ has a neighborhood $p \in \mathcal{V} \subset \mathcal{U}$ such that for any smooth function $\varphi : \mathcal{U} \to \mathbb{R}$ with compact support in $\mathcal{V}$, an estimate of the form*

$$\|\varphi\eta\|_{H^{s+m}(\mathcal{V})} \leqslant C\|D(\varphi\eta)\|_{H^s(\mathcal{V})} + C\|\eta\|_{H^{s+m-1}(\mathcal{V})}$$

holds for all smooth functions $\eta : \mathcal{U} \to \mathbb{F}^k$ and a constant $C > 0$ independent of η.

Proof. The idea is to deduce the result from the case of constant coefficients by choosing the neighborhood $\mathcal{V} \subset \mathcal{U}$ of p small enough so that D is only a small perturbation of the operator

$$D_0 := \sum_{|\alpha|\leqslant m} c_\alpha(p)\partial^\alpha : C^\infty(\mathcal{V}, \mathbb{F}^k) \to C^\infty(\mathcal{V}, \mathbb{F}^\ell),$$

which has constant coefficients and matches D at p. Indeed, let us assume that the bound

$$|c_\alpha(q) - c_\alpha(p)| < \epsilon \qquad \text{for all } q \in \mathcal{V} \text{ and } |\alpha| = m$$

holds for some constant $\epsilon > 0$, which we will be free to make arbitrarily small by shrinking $\mathcal{V}$. Since $\varphi\eta$ and $D_0(\varphi\eta)$ can each be regarded as smooth functions on $\mathbb{R}^n$ with compact support in $\mathcal{V}$, the existing estimate for the case with constant coefficients gives

$$\begin{aligned} \|\varphi\eta\|_{H^{s+m}(\mathcal{V})} = \|\varphi\eta\|_{H^{s+m}(\mathbb{R}^n)} &\leqslant C_1\|D_0(\varphi\eta)\|_{H^s(\mathbb{R}^n)} + C_1\|\varphi\eta\|_{L^2(\mathbb{R}^n)} \\ &= C_1\|D_0(\varphi\eta)\|_{H^s(\mathcal{V})} + C_1\|\varphi\eta\|_{L^2(\mathcal{V})}, \end{aligned} \tag{55.3}$$

in which the constant $C_1 > 0$ comes directly from the estimate (54.12), so it depends on the choice of point p (which determines the operator D_0) but not in any way on the choice of neighborhood $\mathcal{V}$ or function φ. This is important to note, because we will later want to adjust $\mathcal{V}$ and φ in a way that depends on the value of C_1. The term $C_1\|\varphi\eta\|_{L^2(\mathcal{V})}$ is clearly bounded by $C_2\|\eta\|_{H^{s+m-1}(\mathcal{V})}$ for some constant $C_2 > 0$, and we shall not worry about it any further.

By the triangle inequality,

$$\|D_0(\varphi\eta)\|_{H^s(\mathcal{V})} \leqslant \|D(\varphi\eta)\|_{H^s(\mathcal{V})} + \|(D_0 - D)(\varphi\eta)\|_{H^s(\mathcal{V})}.$$

For the first term, we can write

$$D(\varphi\eta) = \sum_{|\alpha|\leqslant m} c_\alpha \partial^\alpha(\varphi\eta) = \varphi\, D\eta + D_1\eta,$$

where $D_1 : C^\infty(\mathcal{U}, \mathbb{F}^k) \to C^\infty(\mathcal{U}, \mathbb{F}^\ell)$ is a differential operator defined by collecting all the terms in $c_\alpha\partial^\alpha(\varphi\eta)$ that involve derivatives of φ, meaning that η is differentiated fewer than m times, and D_1 thus has order at most $m - 1$. It follows that we can write

$$\|D_1\eta\|_{H^s(\mathcal{V})} \leqslant C_2\|\eta\|_{H^{s+m-1}(\mathcal{V})}$$

after possibly increasing the constant $C_2 > 0$. For $(D_0 - D)(\varphi\eta)$, we can similarly write

$$(D_0 - D)(\varphi\eta) = \sum_{|\alpha|=m} (c_\alpha(p) - c_\alpha)\partial^\alpha(\varphi\eta) + D_2\eta$$

for an operator $D_2 : C^\infty(\mathcal{U}, \mathbb{F}^k) \to C^\infty(\mathcal{U}, \mathbb{F}^\ell)$ of order at most $m-1$ whose coefficients depend on c_α and derivatives of φ, so after increasing $C_2 > 0$ again, we have $\|D_2\eta\|_{H^s(\mathcal{V})} \leqslant C_2\|\eta\|_{H^{s+m-1}(\mathcal{V})}$ and thus

$$\|(D_0 - D)(\varphi\eta)\|_{H^s(\mathcal{V})} \leqslant \Big\| \sum_{|\alpha|=m} (c_\alpha(p) - c_\alpha)\partial^\alpha(\varphi\eta) \Big\|_{H^s(\mathcal{V})} + C_2\|\eta\|_{H^{s+m-1}(\mathcal{V})}.$$

The H^s-norm on the right hand side is a sum of L^2-norms of derivatives up to order s, most of which can also be absorbed into the term $C_2\|\eta\|_{H^{s+m-1}(\mathcal{V})}$ at the cost of increasing $C_2 > 0$ further; the only terms that cannot be dealt with in this way are those in which $\partial^\alpha(\varphi\eta)$ (but not $c_\alpha(p) - c_\alpha$) is differentiated s times, and we thus obtain

$$\begin{aligned}\|(D_0 - D)(\varphi\eta)\|_{H^s(\mathcal{V})} &\leqslant \sum_{|\beta|=s}\sum_{|\alpha|=m} \big\|(c_\alpha(p) - c_\alpha)\partial^\beta\partial^\alpha(\varphi\eta)\big\|_{L^2(\mathcal{V})} + C_2\|\eta\|_{H^{s+m-1}(\mathcal{V})} \\ &\leqslant \epsilon C_3\|\varphi\eta\|_{H^{s+m}(\mathcal{V})} + C_2\|\eta\|_{H^{s+m-1}(\mathcal{V})},\end{aligned}$$

where $C_3 > 0$ is a combinatorial constant depending only on s, m and n. After increasing $C_2 > 0$ one more time, the entire computation that began with (55.3) can now be summarized as

$$\|\varphi\eta\|_{H^{s+m}(\mathcal{V})} \leqslant C_1\|D(\varphi\eta)\|_{H^s(\mathcal{V})} + \epsilon C_1C_3\|\varphi\eta\|_{H^{s+m}(\mathcal{V})} + C_2\|\eta\|_{H^{s+m-1}(\mathcal{V})}.$$

While the constant $C_2 > 0$ has been increased several times and depends on φ (though not on η), $C_1 > 0$ still depends only on the choice of point $p \in \mathcal{U}$ and the integers s, m, n, while $C_3 > 0$ depends only the latter. We are thus free to shrink $\mathcal{V}$ and change φ accordingly without altering C_1 or C_3, and can arrange by doing so that

$$\epsilon C_1C_3 \leqslant \frac{1}{2}.$$

This makes it possible to subtract $\epsilon C_1C_2\|\varphi\eta\|_{H^{s+m}(\mathcal{V})}$ from both sides so that the H^{s+m}-norm appears only on the left hand side, and the stated estimate is thus established. $\square$

55.4. Sobolev spaces on compact manifolds. It is now time to define Sobolev spaces of sections of vector bundles over a compact manifold.

If $\mathcal{U} \subset \mathbb{R}^n$ is an open subset, let us say that a function $f : \mathcal{U} \to \mathbb{F}^k$ is of **class** H^s_{loc} on $\mathcal{U}$ and write

$$f \in H^s_{\mathrm{loc}}(\mathcal{U})$$

if its product with arbitrary smooth functions compactly supported in $\mathcal{U}$ is in $H^s_0(\mathcal{U})$, i.e.

$$\varphi f \in H^s_0(\mathcal{U}) \qquad \text{for every } \varphi \in C^\infty_0(\mathcal{U}).$$

We can endow $H^s_{\mathrm{loc}}(\mathcal{U})$ with a natural topology such that a sequence $f_k \in H^s_{\mathrm{loc}}(\mathcal{U})$ converges to $f \in H^s_{\mathrm{loc}}(\mathcal{U})$ if and only if

$$\varphi f_k \to \varphi f \text{ in } H^s_0(\mathcal{U}) \qquad \text{for every } \varphi \in C^\infty_0(\mathcal{U}).$$

The functions φ that you should imagine in this definition are cutoff functions that equal 1 on some open subset but vanish near the boundary of $\mathcal{U}$, so convergence in H^s_{loc} controls the H^s-norm on arbitrary compact subsets of $\mathcal{U}$, but without worrying about what happens at the boundary. In general, the space $H^s_{\mathrm{loc}}(\mathcal{U})$ is larger than $H^s_0(\mathcal{U})$, in fact too large to be a Hilbert space in any obvious way, since e.g. it contains all smooth functions on $\mathcal{U}$, including unbounded functions that admit no smooth extensions over the rest of $\mathbb{R}^n$.

Now if $E \to M$ is a smooth vector bundle over an n-manifold, any open subset $\mathcal{U}_\alpha \subset M$ admitting a chart $x_\alpha : \mathcal{U}_\alpha \to x_\alpha(\mathcal{U}_\alpha) := \mathcal{O}_\alpha \subset \mathbb{R}^n$ and local trivialization $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^k$ associates to each section $\eta : M \to E$ a **local representative**

$$\eta^\alpha := \mathrm{pr}_2 \circ \Phi_\alpha \circ \eta \circ x_\alpha^{-1} : \mathcal{O}_\alpha \to \mathbb{F}^k,$$

where $\mathrm{pr}_2 : \mathcal{U}_\alpha \times \mathbb{F}^k \to \mathbb{F}^k$ denotes the obvious projection. We define

$$H^s_{\mathrm{loc}}(E) := \left\{ \text{sections } \eta : M \to E \;\middle|\; \text{all local representatives } \eta^\alpha \text{ are of class } H^s_{\mathrm{loc}} \right\},$$

and call this the space of **H^s_{loc}-sections** of E. It is a vector space, and can also be endowed with a natural topology for which a sequence $\eta_k \in H^s_{\mathrm{loc}}(E)$ converges to $\eta \in H^s_{\mathrm{loc}}(E)$ if and only if each of its local representatives converges in H^s_{loc} to the corresponding local representative of η. This makes $H^s_{\mathrm{loc}}(E)$ into a topological vector space, but it is not naturally a Hilbert space for arbitrary bundles $E \to M$. It turns out however that if M is compact, then $H^s_{\mathrm{loc}}(E)$ can be endowed with a Hilbert space norm that depends on some choices but is canonical up to equivalence. The choices required are as follows:

(1) A finite open covering $\{\mathcal{U}_\alpha \subset M\}_{\alpha \in I}$ of M;
(2) A partition of unity $\{\varphi_\alpha : M \to [0,1]\}_{\alpha \in I}$ subordinate to the open covering $\{\mathcal{U}_\alpha\}_{\alpha \in I}$;
(3) For each $\alpha \in I$, a chart $x_\alpha : \mathcal{U}_\alpha \to \mathbb{R}^n$ whose image $\mathcal{O}_\alpha := x_\alpha(\mathcal{U}_\alpha)$ is a bounded subset of $\mathbb{R}^n$;
(4) For each $\alpha \in I$, a local trivialization $\Phi_\alpha : E|_{\mathcal{U}_\alpha} \to \mathcal{U}_\alpha \times \mathbb{F}^k$.

Note that if M is compact, then each of the open sets $\mathcal{U}_\alpha \subset M$ automatically has compact closure, and the functions φ_α therefore have compact support in $\mathcal{U}_\alpha$. The condition that every $\mathcal{O}_\alpha = x_\alpha(\mathcal{U}_\alpha)$ is bounded can then always be achieved after replacing $\{\mathcal{U}_\alpha\}_{\alpha \in I}$ with another open covering $\{\mathcal{U}'_\alpha\}_{\alpha \in I}$ consisting of slightly smaller subsets $\mathcal{U}'_\alpha$ with closure in $\mathcal{U}_\alpha$. Conversely, if we require $\{\mathcal{U}_\alpha\}_{\alpha \in I}$ to be a finite open covering by sets with compact closure, then this is only possible if M is compact. We can now associate to each section $\eta : M \to E$ and each $\alpha \in I$ a local representative $\eta^\alpha : \mathcal{O}_\alpha \to \mathbb{F}^k$ determined by the chart x_α and trivialization Φ_α, and writing

$$\psi_\alpha := \varphi_\alpha \circ x_\alpha^{-1} : \mathcal{O}_\alpha \to [0,1],$$

every $\eta \in H^s_{\mathrm{loc}}(E)$ is finite in the norm

$$\|\eta\|_{H^s(M)} := \sum_{\alpha \in I} \|\psi_\alpha \eta^\alpha\|_{H^s(\mathcal{O}_\alpha)},$$

where the norms on the right hand side can be defined via the usual Fourier transform prescription after extending $\psi_\alpha \eta^\alpha : \mathcal{O}_\alpha \to \mathbb{F}^k$ to a function on $\mathbb{R}^n$ that is trivial outside the compact set $\mathrm{supp}(\psi_\alpha) \subset \mathcal{O}_\alpha$. We will usually abbreviate $\|\eta\|_{H^s} := \|\eta\|_{H^s(M)}$ when the domain we're working on is clear from context. One can similarly use the H^s-inner products for compactly supported functions on the domains $\mathcal{O}_\alpha \subset \mathbb{R}^n$ to define an inner product on $H^s_{\mathrm{loc}}(E)$ that induces a norm equivalent to $\|\cdot\|_{H^s(M)}$; we will have no need to make direct use of this inner product, and thus leave the details as an exercise.

Proposition 55.16. *For a vector bundle $E \to M$ over a compact manifold M, the equivalence class of the norm $\|\cdot\|_{H^s(M)}$ on $H^s_{\mathrm{loc}}(E)$ is independent of choices, and a sequence of sections is convergent in this norm if and only if it is H^s_{loc}-convergent. Moreover, $H^s_{\mathrm{loc}}(E)$ with the norm $\|\cdot\|_{H^s(M)}$ is complete.*

Proof sketch. We outline the proof assuming $s \geqslant 0$ is an integer, which is the case we'll need in practice. The main step is to prove that there are continuous linear maps on H^s defined by composing with smooth coordinate transformations and multiplication by smooth functions. More precisely, if $\mathcal{O}_\alpha, \mathcal{O}_\beta \subset \mathbb{R}^n$ are two bounded open subsets and $\psi : \mathcal{O}_\alpha \to \mathcal{O}_\beta$ is a diffeomorphism whose derivatives of all orders are bounded, then there is a continuous linear map

$$H_0^s(\mathcal{O}_\beta) \to H_0^s(\mathcal{O}_\alpha) : f \mapsto f \circ \psi,$$

and similarly, there is a continuous linear map

$$H_0^s(\mathcal{O}_\alpha) \to H_0^s(\mathcal{O}_\alpha) : f \mapsto gf$$

for any smooth function $g : \mathcal{O}_\alpha \to \mathbb{R}$ whose derivatives of all orders are bounded. Both statements can be proved first under the assumption that f is smooth with compact support by using the chain rule and product rule to write out the derivatives up to order s of $f \circ \psi$ and gf, and bounding their L^2-norms. They then extend via density to statements about the spaces $H_0^s(\mathcal{O}_\alpha)$ and $H_0^s(\mathcal{O}_\beta)$. With this understood, if $\{\mathcal{U}_\alpha, \varphi_\alpha, x_\alpha, \Phi_\alpha\}_{\alpha \in I}$ and $\{\mathcal{U}_\beta, \varphi_\beta, x_\beta, \Phi_\beta\}_{\beta \in J}$ are two choices of the data required for defining the norm on $H^s_{\text{loc}}(E)$, then since any $\eta \in H^s_{\text{loc}}(E)$ can be written as a finite sum $\sum_{(\alpha,\beta) \in I \times J} \varphi_\alpha \varphi_\beta \eta$ of sections with compact support in intersections of the form $\mathcal{U}_\alpha \cap \mathcal{U}_\beta$ for $\alpha \in I$ and $\beta \in J$, it suffices to take a function with support in $\mathcal{U}_\alpha \cap \mathcal{U}_\beta$ and compare the H^s-norms of its local representatives with respect to the choices $\alpha \in I$ and $\beta \in J$. The continuity of the two transformations described above makes it possible to bound each in terms of the other.

The completeness of $H^s_{\text{loc}}(E)$ with either of these norms now follows from the completeness of $H_0^s(\mathcal{O})$ for each open set $\mathcal{O} \subset \mathbb{R}^n$, since any Cauchy sequence in $H^s_{\text{loc}}(E)$ will have local representatives that are also Cauchy sequences. □

In light of Proposition 55.16, we shall from now on denote

$$H^s(E) := H^s_{\text{loc}}(E) \qquad \text{whenever } M \text{ is compact,}$$

and refer to this as the space of H^s**-sections** of $E \to M$.

Remark 55.17. There is obviously no problem in defining the space $H^s_{\text{loc}}(E)$ when M is noncompact, but the prescription described above for defining a norm on this space will not work since M cannot be covered by finitely many open subsets with compact closure. This does not mean that Hilbert spaces $H^s(E)$ of sections of $E \to M$ cannot be defined when M is noncompact, but doing so generally requires some extra choices, the exact nature of which depends on the intended application, and the condition $\|\eta\|_{H^s} < \infty$ then typically imposes asymptotic conditions on η, so that not every $\eta \in H^s_{\text{loc}}(E)$ satisfies it, and $H^s(E)$ is therefore a smaller space. Such issues need to be considered quite carefully in any application of Sobolev spaces on noncompact manifolds, but fortunately they are irrelevant for our presentation of Hodge theory.

Every property of the Sobolev spaces $H^s(\mathbb{R}^n)$ and $H_0^s(\mathcal{U})$ we have studied so far can be extended to the setting of a vector bundle over a compact manifold, usually by converting local coordinate computations into global results via a partition of unity. Without getting into the details, here is a summary of general properties, followed by a translation of the local estimate from Lemma 55.15 into the present setting.

Theorem 55.18. *For a smooth vector bundle $E \to M$ over a compact manifold, the Sobolev spaces $H^s(E)$ for $s \geqslant 0$ have the following properties:*

(1) The space $\Gamma(E)$ of smooth sections is a dense subspace of $H^s(E)$.
(2) (Sobolev embedding theorem) If $2s > n := \dim M$, then there is a continuous inclusion

$$H^{s+k}(E) \hookrightarrow C^k(E)$$

for each integer $k \geqslant 0$, where $C^k(E)$ is the space of C^k-sections of $E \to M$.

(3) (Rellich-Kondrashov compactness theorem) For $s > t \geqslant 0$, the inclusion

$$H^s(E) \hookrightarrow H^t(E)$$

is compact.

□

THEOREM 55.19 (global elliptic estimate). *For any elliptic operator $D : \Gamma(E) \to \Gamma(F)$ of order $m \in \mathbb{N}$ between vector bundles E, F over a closed manifold M, and every integer $s \geqslant 0$, the estimate*

$$\|\eta\|_{H^{s+m}} \leqslant C\|D\eta\|_{H^s} + C\|\eta\|_{H^{s+m-1}}$$

is satisfied for all $\eta \in H^{s+m}(E)$, with a constant $C > 0$ independent of η.

PROOF. We can cover M with finitely many neighborhoods $\mathcal{U}_\alpha \subset M$ that are small enough for the local estimate of Lemma 55.15 to be valid on smooth sections of the form $\varphi_\alpha \eta$ with $\eta \in \Gamma(E)$ and $\operatorname{supp}(\varphi_\alpha) \subset \mathcal{U}_\alpha$, and use this open covering in the definition of the H^{s+m}-norm. The terms that appear on the right hand side of the local estimate are then easy to bound in terms of the global norms $\|D\eta\|_{H^s(M)}$ and $\|\eta\|_{H^{s+m-1}(M)}$ for sections of F and E respectively, and this estalishes the stated estimate for all smooth sections $\eta \in \Gamma(E)$. It is then also true for all $\eta \in H^{s+m}(E)$ since $\Gamma(E)$ is dense in the latter, and for any sequence of smooth sections η_j converging in $H^{s+m}(E)$ to η, one has $D\eta_j \to D\eta$ in $H^s(F)$ and $\eta_j \to \eta$ in $H^{s+m-1}(E)$. □

Together with the compactness of the inclusion $H^{s+m}(E) \hookrightarrow H^{s+m-1}(E)$, Theorem 55.19 establishes the main hypothesis of Proposition 55.6 and thus proves:

COROLLARY 55.20. *For every elliptic operator $D : \Gamma(E) \to \Gamma(F)$ of order $m \in \mathbb{N}$ between vector bundles E, F over a closed manifold M, and every integer $s \geqslant 0$, the unique extension of D to a continuous linear operator $H^{s+m}(E) \to H^s(F)$ has finite-dimensional kernel and closed image.* □

REMARK 55.21. One sometimes also sees the estimate in Theorem 55.19 stated with a different norm in place of H^{s+m-1} on the right hand side, e.g. [**War83**, **Ebe**] both put $\|\eta\|_{H^s}$ in place of $\|\eta\|_{H^{s+m-1}}$. Actually, one can use a tool called the "Peter-Paul inequality" to deduce from Theorem 55.19 a seemingly stronger estimate of the form

$$\|\eta\|_{H^{s+m}} \leqslant C\|D\eta\|_{H^s} + C\|\eta\|_{L^2},$$

also valid for all $\eta \in H^{s+m}(E)$. Indeed, the Peter-Paul inequality implies that for smooth sections $\eta \in \Gamma(E)$, $\|\eta\|_{H^{s+m-1}}$ can be bounded by $\epsilon\|\eta\|_{H^{s+m}} + C'\|\eta\|_{L^2}$, where $\epsilon > 0$ can be made arbitrarily small at the cost of allowing $C' > 0$ to be large. Replacing $\|\eta\|_{H^{s+m-1}}$ in Theorem 55.19 with this expression and assuming $\epsilon > 0$ sufficiently small makes it possible to pull the H^{s+m}-norm to the left hand side, leaving only the L^2-norm behind. For our purposes, however, this stronger estimate does not offer any advantages that are not already present in Theorem 55.19, since the inclusion $H^{s+m}(E) \hookrightarrow H^{s+m-1}(E)$ is already compact.

56. The Hodge decomposition theorem

In this lecture we fill in the remaining bits of analysis needed for proving that elliptic operators on closed manifolds are Fredholm operators, and then use this in §56.3 to prove the Hodge decomposition theorem. We will conclude in §56.4 with an application of Hodge theory that constrains the topology of Riemannian manifolds under an assumption about their curvature.

56.1. Elliptic regularity. In §54.4 we proved a regularity result saying that for an elliptic operator D of order $m \in \mathbb{N}$ on $\mathbb{R}^n$ with constant coefficients, weak solutions $\eta \in L^2(\mathbb{R}^n)$ to the equation $D\eta = \xi$ with $\xi \in H^s(\mathbb{R}^n)$ are in $\eta \in H^{s+m}(\mathbb{R}^n)$; informally, the solution is "m steps more regular" than the right hand side of the equation. We will need a version of this result for elliptic operators on manifolds before the main theorem about the Fredholm property can be proved. In this context, we will also need a new interpretation of the term "weak solutions"; our previous definition of this notion required the Fourier transform, and thus made sense for functions on $\mathbb{R}^n$, but not for sections of a bundle over a manifold. This issue will be clarified in §56.1.1 below, but first, here is a statement of the result we are aiming for:

THEOREM 56.1. *Suppose $E, F \to M$ are vector bundles over a smooth manifold without boundary, and $D : \Gamma(E) \to \Gamma(F)$ is an elliptic operator of order $m \in \mathbb{N}$. For any $\xi \in H^s_{\mathrm{loc}}(F)$ for an integer $s \geqslant 0$, every weak solution $\eta \in L^2_{\mathrm{loc}}(E)$ to the equation $D\eta = \xi$ is in $H^{s+m}_{\mathrm{loc}}(E)$. In particular, if ξ is smooth, then so is η.*

The conclusion about the case when ξ is smooth follows from the Sobolev embedding theorem: any section that is of class H^{s+k} for some $s > n/2$ and every integer $k \geqslant 0$ is also of class C^k for every k, and therefore smooth. Note that Theorem 56.1 is of a purely local nature, so it does not require M to be compact, and by choosing coordinates and trivializations, we will be able to assume without loss of generality that M is an open domain $\mathcal{U} \subset \mathbb{R}^n$ on which both bundles are trivial, so D takes the form $\sum_{|\alpha| \leqslant m} c_\alpha \partial^\alpha$ for some smooth matrix-valued coefficient functions $c_\alpha : \mathcal{U} \to \mathrm{Hom}(\mathbb{F}^k, \mathbb{F}^\ell)$. The novel features in comparison with §54.4 are that the coefficients c_α can now be nonconstant, and we are dealing with functions on an open subset $\mathcal{U} \subset \mathbb{R}^n$ rather than all of $\mathbb{R}^n$.

The global elliptic estimate $\|\eta\|_{H^{s+m}} \leqslant C\|D\eta\|_{H^s} + C\|\eta\|_{H^{s+m-1}}$ obtained in the previous lecture makes Theorem 56.1 look plausible, but we still need to work a bit before reaching such a conclusion, because as it stands, our proof of the estimate requires first knowing that η is of class H^{s+m}. In the constant coefficients case, the Fourier transform provided a convenient shortcut that made such hypotheses unnecessary, but when nonconstant coefficients are present, they force extra terms depending on derivatives of η to appear, and for this reason we stated and proved Lemma 55.15 only for *smooth* sections. A density argument extends its validity to functions of class H^{s+m}, but some new ideas will be required before we can relax this regularity assumption further.

Realistically, it would require more than one full lecture to give a complete proof of Theorem 56.1, so we will explain the main ideas but leave some of the details as nontrivial exercises. If you prefer to concentrate on geometric rather than analytical issues, you may want to treat elliptic regularity as a black box provided by analysts and skip ahead to the proof of the Fredholm property in §56.2, though you should first read §56.1.1 so that you know what the term "weak solution" actually means.

Here is a quick sketch of the proof of Theorem 56.1. One of the main tools we will use is the notion of *difference quotients*, which are functions of the form

$$d_j^h f(x) := \frac{f(x + he_j) - f(x)}{h}$$

defined from a function f on some open domain in $\mathbb{R}^n$, where $h \in \mathbb{R}\backslash\{0\}$ is usually assumed small and $e_1, \ldots, e_n \in \mathbb{R}^n$ denotes the standard basis. The classical notion of differentiability of f amounts to the condition that the functions $d_j^h f$ are well behaved in the limit as $h \to 0$, and we will see in §56.1.4 that a similar property holds in the context of Sobolev spaces: a function of class H^s_{loc} is also of class H^{s+1}_{loc} if and only if its difference quotients are uniformly H^s-bounded on compact subsets as $h \to 0$. For weak solutions of the equation $D\eta = \xi$, the way to obtain uniform bounds on $\|d_j^h \eta\|_{H^s}$ will then be via the local elliptic estimate of Lemma 55.15, in which the regularity assumption for ξ allows us to derive a similar uniform bound on Sobolev norms of $d_j^h\xi = d_j^h(D\eta)$, an object that is not identical but closely related to $D(d_j^h\eta)$. In order to apply this strategy toward improving a weak solution η of class $L^2_{\mathrm{loc}} = H^0_{\mathrm{loc}}$, we will need a version of Lemma 55.15 that has $\|\varphi\eta\|_{H^0}$ on the left hand side, which means the right hand side needs to involve Sobolev H^s-norms with $s < 0$, a notion that we have avoided considering thus far. We will give a proper definition for these spaces in §56.1.2.

REMARK 56.2. The proof sketched here is somewhat different from the one that I sketched (even more briefly) in the lecture, and on closer examination, that proof probably cannot be made to work in quite the way I had in mind. The essential idea was to take the equation $D\eta = \xi$ with $\eta, \xi \in L^2_{\mathrm{loc}}$ and use mollifiers (i.e. approximate identities) to smoothen both η and ξ so that the local elliptic estimate can be applied. This is a standard technique, and the details can be found e.g. in [**Ebe**], but doing it properly requires a rather technical lemma due to Friedrichs about the commutator of a mollifier with a differential operator, the proof of which seems disproportionately tedious for the problem at hand. Working with difference quotients is easier, so that is the approach I've settled on for these notes.

56.1.1. *Distributions and weak derivatives.* Suppose $\mathcal{U} \subset \mathbb{R}^n$ is an open subset and $f : \mathcal{U} \to V$ is a function. How do we say that f satisfies a differential equation without explicitly mentioning any derivatives of f? The possible answers to this question all involve integration: e.g. in the existence-uniqueness theory for ODEs, one transforms a differential equation into an integral equation in order to apply the Banach fixed point theorem. The Fourier transform gave rise to another nice approach in Lecture 54, but it only makes sense when $\mathcal{U} = \mathbb{R}^n$, and otherwise the Fourier transform is not defined. A different idea is to use integration by parts. To formulate the appropriate definition, we only need to assume that f is **locally integrable**, i.e. it is of class L^1 on all compact subsets of $\mathcal{U}$, written $f \in L^1_{\mathrm{loc}}(\mathcal{U})$. Every such function determines a continuous linear map

$$\Lambda_f : C_0^\infty(\mathcal{U}) \to V : \varphi \mapsto \int_{\mathcal{U}} \varphi(x) f(x)\, dx, \tag{56.1}$$

where $C_0^\infty(\mathcal{U})$ is the space of smooth real-valued functions with compact support in $\mathcal{U}$, equipped with a very strong topology in which convergence $\varphi_k \to \varphi$ means C^∞_{loc}-convergence with the additional constraint that all of the φ_k have support inside the same compact subset. Linear maps $\Lambda : C_0^\infty(\mathcal{U}) \to V$ that are continuous with respect to this topology are called (V-valued) **distributions** on $\mathcal{U}$. The space of distributions is regarded as an enlargement of the space of locally integrable functions, and its elements are thus also sometimes called **generalized functions**. A popular example of a distribution that is not representable as a function is the so-called **Dirac δ-function**, which physicists are fond of describing as a "function" $\delta : \mathbb{R} \to \mathbb{R}$ that vanishes on $\mathbb{R}\backslash\{0\}$ and is normalized to satisfy $\int_{\mathbb{R}} \delta(x)\, dx = 1$, so that $\int_{\mathbb{R}} \delta(x)\varphi(x)\, dx = \varphi(0)$ for any $\varphi \in C_0^\infty(\mathbb{R})$. No such function actually exists, but δ can be described mathematically as the distribution

$$\delta : C_0^\infty(\mathbb{R}) \to \mathbb{R} : \varphi \mapsto \varphi(0).$$

The **partial derivatives** of a distribution Λ are distributions $\partial_j\Lambda$ defined by

$$(\partial_j\Lambda)(\varphi) := -\Lambda(\partial_j\varphi).$$

The motivation for this definition comes from the example of Λ_f, defined as in (56.1): if f is a C^1-function, then integration by parts implies

$$\partial_j\Lambda_f = \Lambda_{\partial_j f}.$$

For $f, g \in L^1_{\text{loc}}(\mathcal{U})$, we call g a **weak partial derivative** of f in the jth direction and write $\partial_j f = g$ if the distributions Λ_f, Λ_g satisfy $\partial_j \Lambda_f = \Lambda_g$. Concretely, this condition means

$$\int_{\mathcal{U}} \varphi(x) g(x)\, dx = -\int_{\mathcal{U}} \partial_j \varphi(x)\, f(x)\, dx \qquad \text{for all } \varphi \in C_0^\infty(\mathcal{U}), \tag{56.2}$$

and one also says in this situation that the equation $\partial_j f = g$ is satisfied **in the sense of distributions**. In general, the distribution Λ_f defined by a locally integrable function f determines f almost everywhere, so if a weak derivative $\partial_j f$ exists, then it can also be considered unique so long as we regard weak derivatives as *equivalence classes* of functions defined almost everywhere, rather than specific functions. Weak differentiation thus produces well-defined elements $\partial_j f \in L^1_{\text{loc}}(\mathcal{U})$, but in general it does not make sense to speak of the value $\partial_j f(x)$ at any specific point $x \in \mathcal{U}$. On the other hand, if f is of class C^1, then integration by parts implies that its weak partial derivatives can be identified naturally with its classical partial derivatives, which are specific continuous functions. Weak derivatives can also exist for functions that are not classically differentiable: the standard example is $f(x) = |x|$ on $\mathbb{R}$, whose weak derivative $f' \in L^1_{\text{loc}}(\mathbb{R})$ is exactly what you think it should be—it only needs to be defined almost everywhere on $\mathbb{R}$, so there is no need to worry about the value of $f'(0)$. Note that in this example, f' itself does not have a weak derivative (see Exercise 56.3 below); the distribution $\Lambda_{f'}$ does of course have a derivative that is another distribution, but as the example of the Dirac δ-function shows, not all distributions $\Lambda : C_0^\infty(\mathcal{U}) \to V$ can be represented by locally integrable functions.

Distributional derivatives can be iterated arbitrarily many times, thus one can similarly define the notion of a weak higher-order derivative $\partial^\alpha f \in L^1_{\text{loc}}(\mathcal{U})$ of $f \in L^1_{\text{loc}}(\mathcal{U})$ for any multi-index α; if it exists, then it is characterized uniquely almost everywhere via the condition

$$\int_{\mathcal{U}} \varphi(x) \partial^\alpha f(x)\, dx = (-1)^{|\alpha|} \int_{\mathcal{U}} \partial^\alpha \varphi(x)\, f(x)\, dx \qquad \text{for all } \varphi \in C_0^\infty(\mathcal{U}).$$

Exercise 56.3. What is the second derivative of $f(x) = |x|$ in the sense of distributions?

A real-valued distribution $\Lambda : C_0^\infty(\mathcal{U}) \to \mathbb{R}$ can be multiplied by a smooth function $\psi : \mathcal{U} \to \mathbb{R}$, giving rise to another real-valued distribution $\psi\Lambda : C_0^\infty(\mathcal{U}) \to \mathbb{R}$ defined by

$$(\psi\Lambda)(\varphi) := \Lambda(\psi\varphi) \in \mathbb{R},$$

which makes sense because multiplication by ψ defines a continuous linear map $C_0^\infty(\mathcal{U}) \to C_0^\infty(\mathcal{U})$. To extend this to vector-valued distributions, suppose V, W are finite-dimensional real vector spaces, $\Lambda : C_0^\infty(\mathcal{U}) \to V$ is a V-valued distribution, and $\psi : \mathcal{U} \to \operatorname{Hom}(V, W)$ is a smooth function. Choosing a basis $A_1, \ldots, A_n \in \operatorname{Hom}(V, W)$ and writing $\psi = \psi^i A_i$, we can then define a W-valued distribution $\psi\Lambda : C_0^\infty(\mathcal{U}) \to W$ by

$$(\psi\Lambda)(\varphi) := A_i \Lambda(\psi^i \varphi) \in W.$$

You can easily check that (1) this definition does not depend on the choice of basis, and (2) it does what you think it should in the case where Λ is represented by a V-valued function $f \in L^1_{\text{loc}}(\mathcal{U})$, i.e. $\psi\Lambda$ is then represented by the W-valued function $\psi f \in L^1_{\text{loc}}(\mathcal{U})$.

We now have enough definitions in place to consider distributional solutions to linear PDEs with smooth coefficients. For a differential operator $D = \sum_{|\alpha| \leqslant m} c_\alpha \partial^\alpha$ with smooth coefficients $c_\alpha : \mathcal{U} \to \operatorname{Hom}(\mathbb{F}^k, \mathbb{F}^\ell)$ and two distributions Λ, Λ' valued in $\mathbb{F}^k$ and $\mathbb{F}^\ell$ respectively, we say that Λ is a **weak solution** to the equation $D\Lambda = \Lambda'$ if the equation $\sum_{|\alpha| \leqslant m} c_\alpha \partial^\alpha \Lambda = \Lambda'$ is satisfied in the sense of distributions. If $\Lambda = \Lambda_f$ and $\Lambda' = \Lambda_g$ for two locally integrable vector-valued functions f and g, we then call f a **weak solution** to the equation $Df = g$.

The following exercise shows that the notion of weak derivatives defined here is consistent with what was introduced in §54.4 in terms of Fourier transforms.

EXERCISE 56.4. Show that for $f, g \in L^2(\mathbb{R}^n)$ and a multi-index α, $\partial^\alpha f = g$ in the sense of distributions if and only if $\widehat{g}(p) = (2\pi i p)^\alpha \widehat{f}(p)$ for almost every $p \in \mathbb{R}^n$.

EXERCISE 56.5. Show that for an integer $k \geqslant 0$, $H^k(\mathbb{R}^n)$ is the space of functions $f \in L^2(\mathbb{R}^n)$ such that for every multi-index with $|\alpha| \leqslant k$, a weak derivative $\partial^\alpha f$ exists and belongs to $L^2(\mathbb{R}^n)$.

For sections of a vector bundle $E \to M$ over a manifold, the notions of local integrability and weak derivatives carry over by choosing local charts and trivializations: a section is then in $L^1_{\mathrm{loc}}(E)$ if and only if all its local representatives are of class L^1_{loc}, and for a differential operator $D : \Gamma(E) \to \Gamma(F)$, a section $\eta \in L^1_{\mathrm{loc}}(E)$ is a weak solution to $D\eta = \xi$ for some $\xi \in L^1_{\mathrm{loc}}(F)$ if its local representatives are all weak solutions to the corresponding local expressions of this equation. In the global setting, we will not consider solutions that are distributions not representable by functions or sections—we will only need to allow that level of generality in the local picture after choosing coordinates and trivializations.

EXERCISE 56.6. Assume $D : \Gamma(E) \to \Gamma(F)$ is a differential operator and $D^* : \Gamma(F) \to \Gamma(E)$ is its formal adjoint for some choice of bundle metrics on E, F and volume element on M. Show that for $\xi \in L^1_{\mathrm{loc}}(F)$, a section $\eta \in L^1_{\mathrm{loc}}(E)$ is a weak solution to $D\eta = \xi$ if and only if the relation

$$\int_M \langle \varphi, \xi \rangle \, d\mathrm{vol} = \int_M \langle D^*\varphi, \eta \rangle \, d\mathrm{vol}$$

holds for all smooth sections $\varphi \in \Gamma(F)$ with compact support.

56.1.2. *The Sobolev spaces H^{-s}.* this part will be written someday

56.1.3. *Elliptic estimates revisited.* this part will be written someday

56.1.4. *Difference quotients.* this part will be written someday

56.1.5. *Proof of the regularity theorem.* this part will be written someday

56.2. The Fredholm property for elliptic operators. We are now in a position to prove the main theorem about elliptic operators on closed manifolds, the statement of which was previewed in Theorem 53.32. We will establish first a technical version that includes Sobolev spaces in the statement, and follow its proof with a less technical corollary in which Sobolev spaces are required for the proof, but the statement only involves smooth objects.

In the following, $E, F \to M$ are smooth vector bundles endowed with bundle metrics $\langle \ , \ \rangle$, while M is endowed with a volume form or volume element $d\mathrm{vol}$, and we denote

$$\langle \xi, \eta \rangle_{L^2} := \int_M \langle \xi, \eta \rangle \, d\mathrm{vol}$$

for two sections ξ, η of E or F. With these choices in place, every differential operator $D : \Gamma(E) \to \Gamma(F)$ has a formal adjoint $D^* : \Gamma(F) \to \Gamma(E)$ determined by the condition

$$\langle \xi, D\eta \rangle_{L^2} = \langle D^*\xi, \eta \rangle_{L^2}$$

for all $\eta \in \Gamma(E)$ and $\xi \in \Gamma(F)$ with compact support.

THEOREM 56.7. *Assume M is closed, $D : \Gamma(E) \to \Gamma(F)$ is an elliptic differential operator of order $m \in \mathbb{N}$, and $D^* : \Gamma(F) \to \Gamma(E)$ is its formal adjoint with respect to choices of bundle metrics and a volume element as described above. Then:*

(1) *There exist finite-dimensional spaces of smooth sections*

$$\ker D \subset \Gamma(E), \qquad \ker D^* \subset \Gamma(F)$$

which are the kernels of the unique extensions of D and D^* to continuous linear operators $D : H^{k+m}(E) \to H^k(F)$ and $D^* : H^{k+m}(F) \to H^k(E)$ for every integer $k \geqslant 0$.

(2) *For each integer* $k \geqslant 0$, *the image* $\operatorname{im}(D) \subset H^k(F)$ *of the extended operator* $D : H^{k+m}(E) \to H^k(F)$ *is a closed subspace, and*

$$\ker D^* = \left\{ \xi \in L^2(F) \;\middle|\; \langle \xi, D\eta \rangle_{L^2} = 0 \text{ for all } \eta \in H^{k+m}(E) \right\}.$$

Similarly, $\operatorname{im}(D^*) \subset H^k(E)$ *is closed and* $\ker D$ *can be characterized as its* L^2*-orthogonal complement.*

(3) *For each integer* $k \geqslant 0$ *and the extended operators* $D : H^{k+m}(E) \to H^k(F)$ *and* $D^* : H^{k+m}(F) \to H^k(E)$*, we have*

$$H^k(F) = \operatorname{im}(D) \oplus \ker(D^*), \qquad H^k(E) = \operatorname{im}(D^*) \oplus \ker(D),$$

where the summands of each splitting are closed linear subspaces that are L^2*-orthogonal to each other.*

Corollary 56.8. *For each integer* $k \geqslant 0$ *in the setting of Theorem 56.7, the continuous linear operators* $D : H^{k+m}(E) \to H^k(F)$ *and* $D^* : H^{k+m}(F) \to H^k(E)$ *are Fredholm, and the quotient projections* $H^k(F) \to \operatorname{coker}(D) = H^k(F)/\operatorname{im}(D)$ *and* $H^k(E) \to \operatorname{coker}(D^*) = H^k(E)/\operatorname{im}(D^*)$ *restrict to isomorphisms*

$$\ker(D^*) \xrightarrow{\cong} \operatorname{coker}(D), \qquad \ker(D) \xrightarrow{\cong} \operatorname{coker}(D^*),$$

implying in particular that $\operatorname{ind}(D^*) = -\operatorname{ind}(D)$*. Moreover, none of the dimensions of these spaces depend on the choice of integer* $k \geqslant 0$. □

Proof of Theorem 56.7. Note that D is elliptic if and only if D^* is elliptic, thus everything proved about D or D^* will be equally valid if their roles are reversed. Corollary 55.20 established already that the operator $D : H^{k+m}(E) \to H^k(F)$ has finite-dimensional kernel and closed image. Elliptic regularity (Theorem 56.1) implies moreover that every element in the kernel is a smooth section, thus belonging to $H^{k+m}(E)$ for every integer $k \geqslant 0$, and the kernel of the operator is therefore a fixed subspace of $\Gamma(E)$ independent of the choice of k.

For the second statement, observe that since spaces of smooth sections are dense in spaces of H^m-sections, the defining relation $\langle \xi, D\eta \rangle_{L^2} = \langle D^*\xi, \eta \rangle_{L^2}$ for the formal adjoint is valid for all $\eta \in H^m(E)$ and $\xi \in H^m(F)$. If $\eta \in H^{k+m}(E)$ and $\xi \in \ker(D^*) \subset \Gamma(F)$, it therefore follows that

$$\langle \xi, D\eta \rangle_{L^2} = \langle D^*\xi, \eta \rangle_{L^2} = 0,$$

hence ξ lies in the L^2-orthogonal complement of the image of $D : H^{k+m}(E) \to H^k(F)$. Conversely, if $\xi \in L^2(F)$ and $\langle \xi, D\eta \rangle_{L^2} = 0$ for all $\eta \in H^{k+m}(E)$, then this holds in particular for all smooth sections η, so by Exercise 56.6, ξ is a weak solution to the equation $D^*\xi = 0$. Theorem 56.1 then implies that ξ is smooth and belongs to the space $\ker(D^*)$.

Having shown that the closed subspaces $\operatorname{im}(D) \subset H^k(F)$ and $\ker(D^*) \subset H^k(F)$ are L^2-orthogonal to each other, the third statement will follow if we can show that they also span $H^k(F)$, i.e. $\ker(D^*) + \operatorname{im}(D) = H^k(F)$. In the case $k = 0$, this follows immediately from a general fact about Hilbert spaces: for every Hilbert space $\mathcal{H}$ and a closed subspace $V \subset \mathcal{H}$, $\mathcal{H} = V \oplus V^\perp$. The cases $k \in \mathbb{N}$ now follows as a consequence of regularity: given $\xi \in H^k(F) \subset L^2(F)$, the case $k = 0$ gives us $\xi = D\eta + \xi'$ for some $\eta \in H^m(E)$ and $\xi' \in \ker(D^*)$, but then ξ' is smooth and $D\eta = \xi - \xi'$ is therefore of class H^k, implying $\eta \in H^{k+m}(E)$ and thus proving the result. □

Here is the less technical corollary that was promised:

Corollary 56.9. *In the setting of Theorem 56.7, the elliptic operators* $D : \Gamma(E) \to \Gamma(F)$ *and* $D^* : \Gamma(F) \to \Gamma(E)$ *both have finite-dimensional kernels, and we have*

$$\Gamma(F) = \operatorname{im}(D) \oplus \ker(D^*), \qquad \Gamma(E) = \operatorname{im}(D^*) \oplus \ker(D),$$

where in both of these direct sums, each summand is the orthogonal complement of the other with respect to the L^2-pairing on the space of smooth sections $\Gamma(E)$ or $\Gamma(F)$. In particular, the quotient projections $\Gamma(F) \to \operatorname{coker}(D) = \Gamma(F)/\operatorname{im}(D)$ and $\Gamma(E) \to \operatorname{coker}(D^*) = \Gamma(E)/\operatorname{im}(D^*)$ restrict to isomorphisms

$$\ker(D^*) \xrightarrow{\cong} \operatorname{coker}(D), \qquad \ker(D) \xrightarrow{\cong} \operatorname{coker}(D^*).$$

Proof. Given $\xi \in \Gamma(F)$, the $k = 0$ case of Theorem 56.7 gives $\xi = D\eta + \xi'$ for some $\eta \in H^m(E)$ and a unique $\xi' \in \ker(D^*)$. Then $D\eta = \xi - \xi'$ is smooth, so regularity implies η is also smooth, thus proving $\Gamma(F) = \operatorname{im}(D) \oplus \ker(D^*)$. Since the image of the continuously extended operator $D : H^m(E) \to L^2(F)$ is closed in $L^2(F)$, and $\Gamma(E)$ is dense in $H^m(E)$, this image is also the closure in $L^2(F)$ of the image of $D : \Gamma(E) \to \Gamma(F)$, and it follows that both have the same orthogonal complement in $L^2(F)$. By Theorem 56.7, that complement is $\ker(D^*)$, which is contained in $\Gamma(F)$, and is therefore the L^2-orthogonal complement of $\operatorname{im}(D) \subset \Gamma(F)$. What remains to be shown is that $\operatorname{im}(D) \subset \Gamma(F)$ is likewise the L^2-orthogonal complement of $\ker(D^*)$ in $\Gamma(F)$, i.e. every $\xi \in \Gamma(F)$ that is L^2-orthogonal to $\ker(D^*)$ is in the image of $D : \Gamma(E) \to \Gamma(F)$. As shown above, every $\xi \in \Gamma(F)$ can be written as $D\eta + \xi'$ for some $\eta \in \Gamma(E)$ and $\xi' \in \ker(D^*)$, and if ξ is also L^2-orthogonal to ξ', it follows that

$$0 = \langle \xi, \xi' \rangle_{L^2} = \langle D\eta + \xi', \xi' \rangle_{L^2} = \langle \xi', \xi' \rangle_{L^2} = \|\xi'\|^2_{L^2},$$

and thus $\xi' = 0$, implying $\xi = D\eta$. □

56.3. The Hodge decomposition. Now let's apply the results of the previous section to the Laplace-Beltrami operator $\Delta = dd^* + d^*d : \Omega^k(M) \to \Omega^k(M)$ on a closed Riemannian manifold (M, g). Thanks to Corollary 56.9, there will no longer be any need to mention Sobolev spaces, which is fortuitous since de Rham cohomology (also denoted by H with a superscript) is about to make a reappearance.

Recall from Lecture 53 that the Riemannian metric g determines a positive bundle metric $\langle\ ,\ \rangle$ on $\Lambda^k T^*M$ for each $k = 0, \dots, n$, as well as a volume form $d\mathrm{vol} \in \Omega^n(M)$ or (if M is not oriented) volume element. We can thus define an L^2-pairing for differential k-forms $\alpha, \beta \in \Omega^k(M)$,

$$\langle \alpha, \beta \rangle_{L^2} := \int_M \langle \alpha, \beta \rangle \, d\mathrm{vol} \in \mathbb{R},$$

and since d^* and d are formal adjoints with respect to this pairing, Δ is its own formal adjoint. It follows via Corollary 56.9 that $\Omega^k(M)$ splits into the direct sum of the kernel and image of $\Delta : \Omega^k(M) \to \Omega^k(M)$, which are each other's L^2-orthogonal complements in $\Omega^k(M)$. To get further mileage out of this splitting, we observe that for any harmonic k-form ω,

$$0 = \langle \omega, \Delta\omega \rangle_{L^2} = \langle \omega, dd^*\omega + d^*d\omega \rangle_{L^2} = \langle d^*\omega, d^*\omega \rangle_{L^2} + \langle d\omega, d\omega \rangle_{L^2} = \|d^*\omega\|^2_{L^2} + \|d\omega\|^2_{L^2},$$

implying $d\omega \equiv d^*\omega \equiv 0$. One sees immediately from the definition of Δ that the converse is also true, thus:

Proposition 56.10. *A k-form $\omega \in \Omega^k(M)$ on a closed Riemannian manifold (M, g) is harmonic if and only if it is both closed and **co-closed**, where the latter means $d^*\omega \equiv 0$.* □

Remark 56.11. It's worth mentioning at this point that we could have chosen to *define* the term "harmonic k-form" to mean a k-form that is closed and co-closed, without mentioning the Laplace-Beltrami operator, but if we had done this, then it would be very far from obvious that the space of harmonic forms is finite dimensional. In most cases, d and d^* are non-elliptic operators with infinite-dimensional kernels, and there is no obvious geometric reason why the intersection of those two infinite-dimensional spaces should be finite dimensional. That this nonetheless holds is the consequence of a deep interplay between geometry and analysis, resulting from the fact that Δ is elliptic.

We similarly call $\omega \in \Omega^k(M)$ **co-exact** if $\omega = d^*\alpha$ for some $\alpha \in \Omega^{k+1}(M)$. The next result gives a dual interpretation of the k-forms in $\operatorname{im}(\Delta)$: they are uniquely sums of exact and co-exact forms.

PROPOSITION 56.12. *On a closed Riemannian manifold (M, g), the image of $\Delta : \Omega^k(M) \to \Omega^k(M)$ splits into the direct sum of the images of the operators $d : \Omega^{k-1}(M) \to \Omega^k(M)$ and $d^* : \Omega^{k+1}(M) \to \Omega^k(M)$, which are L^2-orthogonal to each other.*

PROOF. For any $\alpha \in \Omega^{k-1}(M)$ and $\beta \in \Omega^{k+1}(M)$, we have $\langle d\alpha, d^*\beta\rangle_{L^2} = \langle d^2\alpha, \beta\rangle_{L^2} = 0$, thus $\operatorname{im}(d)$ and $\operatorname{im}(d^*)$ are L^2-orthogonal subspaces of $\Omega^k(M)$. Since $\Delta\omega = d(d^*\omega) + d^*(d\omega)$ for every $\omega \in \Omega^k(M)$, the image of $\Delta : \Omega^k(M) \to \Omega^k(M)$ is clearly contained in $\operatorname{im}(d) + \operatorname{im}(d^*)$, and we claim conversely that $\operatorname{im}(d) + \operatorname{im}(d^*) \subset \operatorname{im}(\Delta)$. To see this, pick any $\alpha \in \Omega^{k-1}(M)$, $\beta \in \Omega^{k+1}(M)$ and $\omega \in \ker(\Delta) \subset \Omega^k(M)$, and observe

$$\langle d\alpha + d^*\beta, \omega\rangle_{L^2} = \langle \alpha, d^*\omega\rangle_{L^2} + \langle \beta, d\omega\rangle_{L^2} = 0,$$

since by Proposition 56.10, ω is both closed and co-closed. As an element of the L^2-orthogonal complement of $\ker(\Delta)$, $d\alpha + d^*\beta$ is therefore in the image of Δ. □

Now we put it all together.

THEOREM 56.13 (Hodge decomposition theorem). *On any closed Riemannian n-manifold (M, g), for each $k = 0, \dots, n$, the space of smooth differential k-forms splits into a direct sum of three L^2-orthogonal subspaces*

$$\Omega^k(M) = \ker\left(\Omega^k(M) \xrightarrow{\Delta} \Omega^k(M)\right) \oplus \operatorname{im}\left(\Omega^{k-1}(M) \xrightarrow{d} \Omega^k(M)\right) \oplus \operatorname{im}\left(\Omega^{k+1}(M) \xrightarrow{d^*} \Omega^k(M)\right),$$

in which the first summand is the space of closed and co-closed k-forms,

$$\ker(\Delta) = \ker\left(\Omega^k(M) \xrightarrow{d} \Omega^{k+1}(M)\right) \cap \ker\left(\Omega^k(M) \xrightarrow{d^*} \Omega^{k-1}(M)\right),$$

and the sum of the first two summands is the space of all closed k-forms,

$$\ker(\Delta) \oplus \operatorname{im}(d) = \ker\left(\Omega^k(M) \xrightarrow{d} \Omega^{k-1}(M)\right).$$

PROOF. The only detail we haven't already proved is that $\ker(\Delta) \oplus \operatorname{im}(d) = \ker(d)$. It is clear that the former space is contained in the latter, and in light of the splitting of $\Omega^k(M)$ into $\ker(\Delta) \oplus \operatorname{im}(d) \oplus \operatorname{im}(d^*)$, it suffices to prove that $d : \Omega^k(M) \to \Omega^{k+1}(M)$ restricts injectively to the subspace $\operatorname{im}(d^*) \subset \Omega^k(M)$. Indeed, if $\omega \in \Omega^k(M)$ is nontrivial and is equal to $d^*\alpha$ for some $\alpha \in \Omega^{k+1}(M)$, then

$$\langle \alpha, d\omega\rangle_{L^2} = \langle d^*\alpha, \omega\rangle_{L^2} = \langle \omega, \omega\rangle_{L^2} = \|\omega\|_{L^2}^2 > 0,$$

implying $d\omega \neq 0$. □

COROLLARY 56.14. *For each $k = 0, \dots, n$ on a closed Riemannian manifold (M, g), the map $\Omega^k(M) \supset \ker(\Delta) \to H^k_{\mathrm{dR}}(M) = \ker(d)/\operatorname{im}(d) : \omega \mapsto [\omega]$ is an isomorphism.* □

56.4. The Bochner technique. We saw a couple of applications of the Hodge decomposition theorem in §53.2, but those were results about the topology of smooth manifolds, in which the Riemannian metric could be chosen arbitrarily. I'd like to conclude this discussion with another application that demonstrates an interaction between topology and *Riemannian* geometry, in the spirit of the Gauss-Bonnet theorem.

Recall that on a pseudo-Riemannian manifold (M, g), the Ricci tensor $\operatorname{Ric} \in \Gamma(T^0_2 M)$ is symmetric, and can thus be identified with a quadratic form on the tangent space at each point. We write

$$\operatorname{Ric} \geqslant 0 \qquad \text{or} \qquad \operatorname{Ric} > 0$$

on some region $\mathcal{U} \subset M$ if that quadratic form is nonnegative or positive-definite respectively at every point in $\mathcal{U}$, i.e. $\mathrm{Ric}(X, X) \geqslant 0$ for all $X \in T_p M$ at points $p \in \mathcal{U}$, with strict inequality for $X \neq 0$ in the second case. For example, if $\dim M = 2$, then the relation $\mathrm{Ric} = K_G \cdot g$ from Exercise 36.8 implies that $\mathrm{Ric} \geqslant 0$ if and only if the Gaussian curvature is nonnegative.

THEOREM 56.15. *For any closed connected Riemannian n-manifold (M, g) satisfying* $\mathrm{Ric} \geqslant 0$ *everywhere,* $\dim H^1_{\mathrm{dR}}(M) \leqslant n$*, and if additionally* $\mathrm{Ric} > 0$ *at some point, then* $H^1_{\mathrm{dR}}(M)$ *is trivial.*

EXAMPLE 56.16. By de Rham's theorem, $H^1_{\mathrm{dR}}(M)$ for any smooth manifold M is isomorphic to $H^1(M; \mathbb{R})$, the singular cohomology with real coefficients. The computation of the latter for closed oriented surfaces Σ is a standard topic in algebraic topology: if you haven't seen this but have seen how to compute their fundamental groups, you get $H^1(\Sigma; \mathbb{R})$ by taking the abelianization of $\pi_1(\Sigma)$ and tensoring the result with $\mathbb{R}$, hence $\dim H^1(\Sigma, \mathbb{R})$ is the number of generators in the standard presentation of $\pi_1(\Sigma)$. Concretely, if Σ_g denotes the closed oriented surface of genus $g \geqslant 0$, which includes S^2 and $\mathbb{T}^2$ as the special cases Σ_0 and Σ_1 respectively, then

$$\dim H^1_{\mathrm{dR}}(\Sigma_g) = 2g.$$

Theorem 56.15 thus tells us something about surfaces that we could also have deduced from the Gauss-Bonnet theorem: Σ_g for $g \geqslant 2$ does not admit any Riemannian metric with everywhere nonnegative Ricci curvature, and the only case in which the Ricci curvature can be everywhere nonnegative and somewhere positive is $g = 0$.

Theorem 56.15 is due to Bochner, and its proof follows a trick known as the *Bochner technique*, which is also responsible for various other global results relating curvature and topology on closed Riemannian manifolds. The main ingredient needed is a so-called *Weitzenböck formula*, which relates the Laplace-Beltrami operator Δ to another second-order differential operator that is defined in a somewhat more general context, and is also a variation on the Laplace operator.

Suppose $(E, \langle\ ,\ \rangle)$ is a Euclidean vector bundle over the Riemannian manifold (M, g), with a choice of metric connection ∇. This together with the Levi-Cività connection on TM induces a natural connection on

$$F := \mathrm{Hom}(TM, E) \cong T^*M \otimes E,$$

which is likewise compatible with the bundle metric induced on $T^*M \otimes E$ by the metrics on TM and E. With this data in place, the first-order differential operator $\nabla : \Gamma(E) \to \Gamma(F)$ has a formal adjoint

$$\nabla^* : \Gamma(F) \to \Gamma(E),$$

and it will be useful to write down an explicit formula for it. Observe that for any $\lambda \in \Gamma(F) = \Gamma(T^*M \otimes E) = \Omega^1(M, E)$, the covariant derivative $\nabla\lambda$ is a section of $\mathrm{Hom}(TM, F) = T^*M \otimes T^*M \otimes F$, which can be transformed via a musical isomorphism into a section of $T^*M \otimes TM \otimes E$ and then contracted via the map

$$T^*M \otimes TM \otimes E \to E : \alpha \otimes X \otimes v \mapsto \alpha(X)v.$$

The composition of this contraction with the aforementioned musical isomorphism defines a kind of trace, which we will denote by

$$\mathrm{tr}_g : T^*M \otimes T^*M \otimes E \to E : \alpha \otimes \beta \otimes v \mapsto \alpha(\beta^\sharp)v = \langle \alpha, \beta \rangle v,$$

and if we regard elements $\omega \in T^*_p M \otimes T^*_p M \otimes E_p$ as bilinear maps $\omega : T_p M \times T_p M \to E$, the trace can be computed as

$$\mathrm{tr}_g(\omega) = \sum_{j=1}^{n} \omega(e_j, e_j) \in E_p$$

for any choice of orthonormal basis $e_1, \dots, e_n \in T_pM$. We can now define the **divergence** of a bundle-valued 1-form $\lambda \in \Omega^1(M, E)$ by

$$\operatorname{div}(\lambda) := \operatorname{tr}_g(\nabla\lambda) \in \Gamma(E).$$

Choosing an orthonormal basis $e_1, \dots, e_n \in T_pM$ at a point $p \in M$, this definition gives

$$\operatorname{div}(\lambda)(p) = \sum_{j=1}^{n} (\nabla_{e_j}\lambda)(e_j) \in E_p. \tag{56.3}$$

LEMMA 56.17. *For every $\lambda \in \Omega^1(M, E) = \Gamma(F)$, $\nabla^*\lambda = -\operatorname{div}(\lambda)$.*

PROOF. We can use the bundle metric $\langle\ ,\ \rangle$ on E to define a fiberwise bilinear pairing $\langle\ ,\ \rangle : E \oplus F \to T^*M$ via the map

$$E \otimes F = E \otimes (T^*M \otimes E) \to T^*M : v \otimes \lambda \otimes w \mapsto \langle v, w\rangle\lambda =: \langle v, \lambda \otimes w\rangle,$$

thus giving a pairing $\langle \eta, \lambda\rangle \in \Omega^1(M)$ for any $\eta \in \Gamma(E)$ and $\lambda \in \Omega^1(M, E)$. The main step is to show that this pairing relates the divergence on $\Omega^1(M, E)$ to the divergence of vector fields via a Leibniz rule, namely

$$\operatorname{div}\left(\langle \eta, \lambda\rangle^\sharp\right) = \langle \nabla\eta, \lambda\rangle + \langle \eta, \operatorname{div}(\lambda)\rangle \qquad \text{for all } \eta \in \Gamma(E),\ \lambda \in \Omega^1(M, E). \tag{56.4}$$

To prove that this holds at any given point $p \in M$, let us choose an orthonormal frame $e_1, \dots, e_n$ for TM on some neighborhood $\mathcal{U} \subset M$ of p such that each e_j satisfies $\nabla e_j = 0$ at p, and let $e_*^j := (e_j)_\flat$, so that $e_*^1, \dots, e_*^n$ is the dual frame for T^*M over $\mathcal{U}$. On this neighborhood, we can then write $\lambda = e_*^j \otimes \lambda_j$ for unique sections $\lambda_1, \dots, \lambda_n \in \Gamma(E|_{\mathcal{U}})$, and the assumption that the e_j are parallel at p implies $\nabla_Y\lambda = e_*^j \otimes \nabla_Y\lambda_j$ for any $Y \in T_pM$, thus

$$\operatorname{div}(\lambda) = \sum_{j=1}^{n} (\nabla_{e_j}\lambda)(e_j) = \sum_{j=1}^{n} (e_*^i \otimes \nabla_{e_j}\lambda_i)(e_j) = \sum_{j=1}^{n} e_*^i(e_j)\nabla_{e_j}\lambda_i = \sum_{j=1}^{n} \nabla_{e_j}\lambda_j \qquad \text{at } p.$$

Note that in these expressions, there are *two* summations whenever both of the indices i and j appear; the summation over i is implied due to the Einstein convention, but we've written the summation over j explicitly since j does not appear in an upper/lower pair. (Some similar situations will occur in further calculations below, and you should always assume the Einstein convention is in effect when upper/lower pairs of indices appear, but does not apply e.g. to a pair of matching lower indices.)

We also have $\langle \eta, \lambda\rangle = \langle \eta, e_*^j \otimes \lambda_j\rangle = \langle \eta, \lambda_j\rangle e_*^j$ and thus $\langle \eta, \lambda\rangle^\sharp = \sum_{j=1}^n \langle \eta, \lambda_j\rangle e_j$. Applying the Leibniz rule for divergences of vector fields (Exercise 53.1) and the assumption that $\nabla e_j = 0$ at p now gives

$$\begin{aligned}
\operatorname{div}\left(\langle \eta, \lambda\rangle^\sharp\right) &= \sum_{j=1}^{n} d\big(\langle \eta, \lambda_j\rangle\big)(e_j) = \sum_{j=1}^{n} \big(\langle \nabla_{e_j}\eta, \lambda_j\rangle + \langle \eta, \nabla_{e_j}\lambda_j\rangle\big) \\
&= \langle e_*^i, e_*^j\rangle \cdot \langle \nabla_{e_i}\eta, \lambda_j\rangle + \left\langle \eta, \sum_{j=1}^{n} \nabla_{e_j}\lambda_j \right\rangle = \langle e_*^i \otimes \nabla_{e_i}\eta, e_*^j \otimes \lambda_j\rangle + \langle \eta, \operatorname{div}(\lambda)\rangle \\
&= \langle \nabla\eta, \lambda\rangle + \langle \eta, \operatorname{div}(\lambda)\rangle
\end{aligned}$$

as claimed.

Since integrals of divergences of compactly supported vector fields times $d\mathrm{vol}$ always vanish, it follows from (56.4) that

$$\int_M \langle \nabla\eta, \lambda\rangle\, d\mathrm{vol} + \int_M \langle \eta, \operatorname{div}(\lambda)\rangle\, d\mathrm{vol} = 0$$

for any $\eta \in \Gamma(E)$ and $\lambda \in \Omega^1(M, E)$ with compact support. $\square$

Definition 56.18. On a Euclidean vector bundle E with metric connection ∇ over a Riemannian manifold (M, g), the **Bochner Laplacian** is the second-order differential operator

$$\nabla^*\nabla : \Gamma(E) \to \Gamma(E) : \eta \mapsto -\operatorname{div}(\nabla\eta) = -\operatorname{tr}_g(\nabla^2\eta).$$

The Bochner Laplacian is also an elliptic operator, though we will not need to exploit any of the deeper consequences of this fact. We will make use of the observation that it defines a nonnegative quadratic form on $\Gamma(E)$, since

$$\langle \eta, \nabla^*\nabla\eta \rangle_{L^2} = \langle \nabla\eta, \nabla\eta \rangle_{L^2} = \|\nabla\eta\|_{L^2}^2 \geqslant 0$$

for all compactly supported smooth sections $\eta \in \Gamma(E)$, with equality if and only if η is parallel. For the purposes of Theorem 56.15, the vector bundle on which we need to consider $\nabla^*\nabla$ is T^*M, with its natural bundle metric and connection induced by g and the Levi-Cività connection. More generally, $\nabla^*\nabla$ can also be defined on $\Lambda^k T^*M \to M$ for any $k \geqslant 0$. As clarified by the next exercise, the Levi-Cività connection on TM induces a natural connection on $\Lambda^k T^*M$ that respects the bundle metric. Recall that every covariant rank k tensor field $\omega \in \Gamma(T_k^0 M)$ can be "antisymmetrized" to produce an alternating k-form $\operatorname{Alt}(\omega) \in \Omega^k(M)$ defined by

$$\operatorname{Alt}(\omega)(X_1, \ldots, X_k) := \frac{1}{k!} \sum_{\sigma \in S_k} (-1)^{|\sigma|} \omega(X_{\sigma(1)}, \ldots, X_{\sigma(k)}),$$

so in particular, $\operatorname{Alt}(\omega) = \omega$ if and only if ω is alternating. We can regard Alt as a smooth linear bundle map $(T^*M)^{\otimes k} = T_k^0 M \to T_k^0 M$ defining a fiberwise-linear projection to the subbundle $\Lambda^k T^*M \subset T_k^0 M$.

Exercise 56.19. Assume M is a smooth manifold with an affine connection ∇, and let ∇ also denote the natural connection induced on the tensor bundle $T_k^0 M = (T^*M)^{\otimes k}$ for each $k \in \mathbb{N}$.

(a) Given integers $1 \leqslant i < j \leqslant k$, let $\tau : T_k^0 M \to T_k^0 M$ denote the bundle map

$$(\tau\omega)(X_1, \ldots, X_i, \ldots, X_j, \ldots, X_k) := \omega(X_1, \ldots, X_j, \ldots, X_i, \ldots, X_k).$$

Show that τ is a parallel section of $\operatorname{End}(T_k^0 M)$ with respect to the natural connection induced by ∇.

(b) Show that $\operatorname{Alt} : T_k^0 M \to T_k^0 M$ is also a parallel section of $\operatorname{End}(T_k^0 M)$, and conclude that for every $X \in \mathfrak{X}(M)$ and $\omega \in \Omega^k(M) \subset \Gamma(T_k^0 M)$, $\nabla_X \omega \in \Gamma(T_k^0 M)$ is also in $\Omega^k(M)$. This shows that the connection on $T_k^0 M$ has a natural restriction to the subbundle $\Lambda^k T^*M$, i.e. its parallel transport maps preserve the subbundle.

(c) Show that the connection on $\Lambda^* T^*M = \bigoplus_{k=0}^n \Lambda^k T^*M$ resulting from part (b) satisfies the Leibniz rule

$$\nabla_X(\alpha \wedge \beta) = \nabla_X \alpha \wedge \beta + \alpha \wedge \nabla_X \beta \qquad \text{for all } \alpha, \beta \in \Omega^*(M) \text{ and } X \in \mathfrak{X}(M).$$

(d) Show that if the original affine connection on M is compatible with a pseudo-Riemannian metric, then the induced connection on $\Lambda^* TM$ is compatible with the induced bundle metric on $\Lambda^* T^*M$ (defined via Lemma 53.8), and the bundle map $\operatorname{Alt} : T_k^0 M \to T_k^0 M$ is self-adjoint on each fiber with respect to the induced bundle metric on $T_k^0 M$. In particular, Alt can be interpreted as a fiberwise orthogonal projection.
Caution: Recall from the proof of Lemma 53.8 that the natural bundle metrics on $\Lambda^k T^*M$ and $T_k^0 M$ are not identical in general, but are related by a combinatorial factor.

(e) Show that if the original affine connection on M is symmetric, then for the induced connection on $\Lambda^k T^*M$ and any $\omega \in \Omega^k(M) = \Gamma(\Lambda^k T^*M)$, the covariant derivative $\nabla\omega \in \Omega^1(M, \Lambda^k T^*M) = \Gamma(T^*M \otimes \Lambda^k T^*M) \subset \Gamma(T^0_{k+1}M)$ satisfies

$$\mathrm{Alt}(\nabla\omega) = \frac{1}{k+1} d\omega.$$

Hint: Use geodesic normal coordinates (see Exercise 34.12).

In light of Exercise 56.19, we shall assume in the following that for a Riemannian manifold (M, g), the exterior product bundle $\Lambda^* T^*M$ is endowed with the natural connection induced by the Levi-Cività connection, which is then compatible with the natural bundle metric on $\Lambda^* T^*M$, and also enables us to compute exterior derivatives by composing Alt with ∇ and multiplying by a combinatorial factor. Since $\Lambda^k T^*M$ is a subbundle of $T^*M \otimes \Lambda^{k-1} T^*M$, which is naturally isomorphic to $\mathrm{Hom}(TM, \Lambda^{k-1} T^*M)$, we can interpret any k-form $\omega \in \Omega^k(M)$ as a $\Lambda^{k-1} T^*M$-valued 1-form, and thus define its divergence

$$\mathrm{div}(\omega) = \mathrm{tr}_g(\nabla\omega) \in \Gamma(\Lambda^{k-1} T^*M) = \Omega^{k-1}(M) \qquad \text{for } \omega \in \Omega^k(M) \subset \Omega^1(M, \Lambda^{k-1} T^*M),$$

as a special case of what was defined above for arbitrary bundle-valued 1-forms.

LEMMA 56.20. *For $\beta \in \Omega^k(M)$ on a Riemannian manifold (M, g), $d^*\beta = -\mathrm{div}(\beta)$.*

PROOF. The result follows from an extra observation added to the Leibniz rule in (56.4), but we should first be clear on exactly what bundle metrics we are using. Let $E := \Lambda^{k-1} T^*M$ and $F := \mathrm{Hom}(TM, \Lambda^{k-1} T^*M) = T^*M \otimes \Lambda^{k-1} T^*M$, which are subbundles of $T^0_{k-1}M$ and $T^0_k M$ respectively, while F also contains $\Lambda^k T^*M$ as a distinguished subbundle. The bundle metric g on TM induces a dual bundle metric on T^*M and resulting tensor product bundle metrics on $T^0_k M$ and $T^0_{k-1}M$ as explained in the proof of Lemma 53.8; we shall denote these bundle metrics by $\langle\ ,\ \rangle_\otimes$, and recall that the natural bundle metric on $\Lambda^k T^*M$ is related to it by

$$\langle\ ,\ \rangle = \frac{1}{k!} \langle\ ,\ \rangle_\otimes \qquad \text{on } \Lambda^k T^*M.$$

The correct bundle metric to use on $F = T^*M \otimes \Lambda^{k-1} T^*M$ is determined via the tensor product from the bundle metrics we use on T^*M and $\Lambda^{k-1} T^*M$, thus it is also not identical to the bundle metric on the larger bundle $T^0_k M = T^*M \otimes T^0_{k-1}M$, but is related to it by

$$\langle\ ,\ \rangle = \frac{1}{(k-1)!} \langle\ ,\ \rangle_\otimes \qquad \text{on } F.$$

Now for $\alpha \in \Omega^{k-1}(M) = \Gamma(E)$ and $\beta \in \Omega^k(M) \subset \Gamma(F)$, (56.4) gives

$$\mathrm{div}\left(\langle \alpha, \beta \rangle^\sharp\right) = \langle \nabla\alpha, \beta \rangle + \langle \alpha, \mathrm{div}(\beta) \rangle,$$

where the two bundle metrics appearing on the right hand side are those of F and E respectively. We can rewrite the first one in terms of the bundle metric on $T^0_k M$ and then use the fact that $\mathrm{Alt} : T^0_k M \to T^0_k M$ is a fiberwise orthogonal projection, whose kernel is therefore orthogonal to the alternating form $\beta \in \Omega^k(M) \subset \Gamma(T^0_k M)$, thus

$$\langle \nabla\alpha, \beta \rangle = \frac{1}{(k-1)!} \langle \nabla\alpha, \beta \rangle_\otimes = \frac{1}{(k-1)!} \langle \mathrm{Alt}(\nabla\alpha), \beta \rangle_\otimes = \frac{1}{k!} \langle d\alpha, \beta \rangle_\otimes = \langle d\alpha, \beta \rangle,$$

where we've used the formula in Exercise 56.19 for the exterior derivative, and replaced $\langle\ ,\ \rangle_\otimes$ in the last step with the natural bundle metric on $\Lambda^k T^*M$. We've thus proved that for any $\alpha \in \Omega^{k-1}(M)$ and $\beta \in \Omega^k(M)$, $\langle d\alpha, \beta \rangle + \langle \alpha, \mathrm{div}(\beta) \rangle$ is the divergence of a vector field that has compact support if α and β do, so the result follows. $\square$

The next result is the main ingredient beyond the Hodge decomposition theorem that is needed for the proof of Theorem 56.15.

PROPOSITION 56.21 (Weitzenböck formula). *For* $\omega \in \Omega^1(M)$, $\Delta\omega = \nabla^*\nabla\omega + \mathrm{Ric}(\omega^\sharp, \cdot)$.

PROOF. Given $p \in M$, we can choose an orthonormal frame $e_1, \ldots, e_n$ for TM near p such that each e_j is parallel at p, and consider the dual frame formed by the 1-forms $e_*^j := (e_j)_\flat$ for $j = 1, \ldots, n$. We will use these frames to prove via a direct computation that the formula holds at p.

Given $\omega \in \Omega^1(M) = \Gamma(T^*M)$, $\nabla\omega \in \Omega^1(M, T^*M) = \Gamma(T^*M \otimes T^*M) = \Gamma(T_2^0 M)$ can be written near p as

$$\nabla\omega = e_*^j \otimes \nabla_{e_j}\omega,$$

and since the 1-forms e_*^j are parallel at p, we then have

$$\nabla_X(\nabla\omega) = e_*^j \otimes \nabla_X\nabla_{e_j}\omega \qquad \text{for } X \in T_pM,$$

implying

$$\nabla^*\nabla\omega = -\operatorname{div}(\nabla\omega) = -\sum_{i=1}^n \left(\nabla_{e_i}(\nabla\omega)\right)(e_i) = -\sum_{i=1}^n \left(e_*^j \otimes \nabla_{e_i}\nabla_{e_j}\omega\right)(e_i) = -\sum_{i=1}^n e_*^j(e_i)\nabla_{e_i}\nabla_{e_j}\omega$$
$$= -\sum_{j=1}^n \nabla_{e_j}\nabla_{e_j}\omega.$$

In order to carry out a similar computation of $\Delta\omega$ at p, we need local formulas for d and d^*. We claim: for any $\alpha \in \Omega^*(M)$, on the neighborhood of p where our orthonormal frame is defined,

$$d\alpha = e_*^j \wedge \nabla_{e_j}\alpha \qquad \text{and} \qquad d^*\alpha = -\sum_{j=1}^n \iota_{e_j}(\nabla_{e_j}\alpha). \tag{56.5}$$

Indeed, assuming $\alpha \in \Omega^k(M)$, we have $\nabla\alpha = e_*^j \otimes \nabla_{e_j}\alpha$, and can thus apply Exercise 56.19 and one of the standard formulas relating the wedge product and tensor product to compute

$$d\alpha = (k+1)\operatorname{Alt}(\nabla\alpha) = (k+1)\operatorname{Alt}(e_*^j \otimes \nabla_{e_j}\alpha) = \frac{(k+1)!}{k!1!}\operatorname{Alt}(e_*^j \otimes \nabla_{e_j}\alpha) = e_*^j \wedge \nabla_{e_j}\alpha.$$

For the second formula in (56.5), observe that if $\alpha \in \Omega^k(M)$ is identified with a bundle-valued 1-form $\alpha \in \Omega^1(M, \Lambda^{k-1}T^*M)$, its evaluation on a tangent vector $X \in T_qM$ at some point $q \in M$ is equivalent to the interior product $\iota_X\alpha \in \Lambda^{k-1}T_q^*M$, thus (56.3) gives

$$\operatorname{div}(\alpha) = \sum_{j=1}^n \iota_{e_j}(\nabla_{e_j}\alpha),$$

which implies the result in light of Lemma 56.20.

The remaining ingredient we need is a formula for commuting the covariant derivatives of $\omega \in \Omega^1(M)$ in different directions; this is of course how the curvature will appear. Since the e_j are parallel at p, their Lie brackets $[e_i, e_j] = \nabla_{e_i}e_j - \nabla_{e_j}e_i$ also vanish at this point, so any vector field $X \in \mathfrak{X}(M)$ satisfies

$$\nabla_{e_i}\nabla_{e_j}X - \nabla_{e_j}\nabla_{e_i}X = R(e_i, e_j)X \qquad \text{at } p.$$

If we choose X so that $\nabla X = 0$ at p, then computing second derivatives of the real-valued function $\omega(X)$ at p will produce no contributions from first derivatives of X, thus working again at p so that $[e_i, e_j] = 0$,

$$0 = (\mathcal{L}_{e_i}\mathcal{L}_{e_j} - \mathcal{L}_{e_j}\mathcal{L}_{e_i} - \mathcal{L}_{[e_i,e_j]})\big(\omega(X)\big) = (\nabla_{e_i}\nabla_{e_j}\omega - \nabla_{e_j}\nabla_{e_i}\omega)(X) + \omega(\nabla_{e_i}\nabla_{e_j}X - \nabla_{e_j}\nabla_{e_i}X)$$
$$= (\nabla_{e_i}\nabla_{e_j}\omega - \nabla_{e_j}\nabla_{e_i}\omega)(X) + \omega(R(e_i,e_j)X) \qquad \text{at } p,$$

implying the formula

$$(\nabla_{e_i}\nabla_{e_j}\omega - \nabla_{e_j}\nabla_{e_i}\omega)(X) = -\omega(R(e_i,e_j)X), \tag{56.6}$$

which is valid for all $X \in T_pM$ since a vector field that is parallel at p can be chosen to have any value at p.

With these ingredients in place, we compute at the point p and use the Einstein summation convention wherever matching upper/lower pairs of indices appear:

$$\begin{aligned}\Delta\omega - \nabla^*\nabla\omega = dd^*\omega + d^*d\omega - \nabla^*\nabla\omega &= -d\left(\sum_{j=1}^n \iota_{e_j}(\nabla_{e_j}\omega)\right) + d^*\left(e_*^j \wedge \nabla_{e_j}\omega\right) + \sum_{j=1}^n \nabla_{e_j}\nabla_{e_j}\omega \\ &= -\sum_{j=1}^n e_*^i \wedge \nabla_{e_i}\left(\iota_{e_j}(\nabla_{e_j}\omega)\right) - \sum_{i=1}^n \iota_{e_i}\left(\nabla_{e_i}(e_*^j \wedge \nabla_{e_j}\omega)\right) + \sum_{j=1}^n \nabla_{e_j}\nabla_{e_j}\omega \\ &= -\sum_{j=1}^n e_*^i \wedge \iota_{e_j}\left(\nabla_{e_i}\nabla_{e_j}\omega\right) - \sum_{i=1}^n \iota_{e_i}\left(e_*^j \wedge \nabla_{e_i}\nabla_{e_j}\omega\right) + \sum_{j=1}^n \nabla_{e_j}\nabla_{e_j}\omega,\end{aligned}$$

where we've used the equalities $\nabla_{e_i}\left(\iota_{e_j}(\nabla_{e_j}\omega)\right) = \iota_{e_j}\left(\nabla_{e_i}\nabla_{e_j}\omega\right)$ and $\nabla_{e_i}\left(e_*^j \wedge \nabla_{e_j}\omega\right) = e_*^j \wedge \nabla_{e_i}\nabla_{e_j}\omega$, both of which result from the assumption that $\nabla e_j = 0$ at p. Next, recall from Exercise 14.7 in the first semester that the interior product satisfies a graded Leibniz rule with respect to the wedge product: $\iota_X(\alpha\wedge\beta) = \iota_X\alpha \wedge \beta + (-1)^{|\alpha|}\alpha \wedge \iota_X\beta$. This can be applied to the second of the three summations above, and will cause some cancellations if we first exchange the order of the derivatives in the first summation and relabel the indices: using (56.6), the first summation can be rewritten as

$$-\sum_{j=1}^n e_*^i \wedge \iota_{e_j}\left(\nabla_{e_i}\nabla_{e_j}\omega\right) = -\sum_{i=1}^n e_*^j \wedge \iota_{e_i}\left(\nabla_{e_j}\nabla_{e_i}\omega\right) = -\sum_{i=1}^n e_*^j \wedge \iota_{e_i}\Big(\nabla_{e_i}\nabla_{e_j}\omega + \omega\big(R(e_i,e_j)(\cdot)\big)\Big).$$

We can now put this together with the other two summations, apply the Leibniz rule for ι_{e_i}, cancel redundant terms and use the symmetries of the Riemann tensor so that the remaining summation produces the Ricci tensor:

$$\begin{aligned}\Delta\omega - \nabla^*\nabla\omega &= -\sum_{i=1}^n e_*^j \wedge \iota_{e_i}\Big(\omega\big(R(e_i,e_j)(\cdot)\big)\Big) - \sum_{i=1}^n \iota_{e_i}(e_*^j)\,\nabla_{e_i}\nabla_{e_j}\omega + \sum_{j=1}^n \nabla_{e_j}\nabla_{e_j}\omega \\ &= -\sum_{i=1}^n \omega\big(R(e_i,e_j)e_i\big)\,e_*^j - \sum_{j=1}^n \nabla_{e_j}\nabla_{e_j}\omega + \sum_{j=1}^n \nabla_{e_j}\nabla_{e_j}\omega = -\sum_{i=1}^n \omega\big(R(e_i,e_j)e_i\big)\,e_*^j \\ &= -\sum_{i=1}^n \langle\omega^\sharp, R(e_i,e_j)e_i\rangle\,e_*^j = -\sum_{i=1}^n \operatorname{Riem}(\omega^\sharp, e_i, e_j, e_i)\,e_*^j = \sum_{i=1}^n \operatorname{Riem}(\omega^\sharp, e_j, e_i, e_i)\,e_*^j \\ &= \sum_{i=1}^n \operatorname{Riem}(e_i, e_i, \omega^\sharp, e_j)\,e_*^j = \operatorname{Ric}(\omega^\sharp, e_j)\,e_*^j = \sum_{j=1}^n \operatorname{Ric}(\omega^\sharp, \langle e_j, \cdot\rangle e_j) \\ &= \operatorname{Ric}(\omega^\sharp, \cdot) \in T_p^*M.\end{aligned}$$

□

Proof of Theorem 56.15. Assuming $\mathrm{Ric} \geqslant 0$ everywhere, we will show that the space of harmonic 1-forms on M satisfies the dimension bounds stated in the theorem, so that the result follows from the isomorphism of $H^*_{\mathrm{dR}}(M)$ with the space of harmonic forms. If $\omega \in \Omega^1(M)$ is harmonic, then by the Weitzenböck formula, we have

$$0 = \int_M \langle \omega, \Delta\omega \rangle \, d\mathrm{vol} = \int_M \langle \omega, \nabla^*\nabla\omega \rangle \, d\mathrm{vol} + \int_M \langle \omega, \mathrm{Ric}(\omega^\sharp, \cdot) \rangle \, d\mathrm{vol}.$$

By the defining property of the formal adjoint, the first term on the right is $\int_M \langle \nabla\omega, \nabla\omega \rangle \, d\mathrm{vol} = \int_M |\nabla\omega|^2 \, d\mathrm{vol}$, and the second term can be understood by applying a musical isomorphism to both sides of the inner product, giving

$$\langle \omega, \mathrm{Ric}(\omega^\sharp, \cdot) \rangle = \langle \omega^\sharp, \mathrm{Ric}^\sharp(\omega^\sharp) \rangle = \mathrm{Ric}(\omega^\sharp, \omega^\sharp),$$

thus

$$0 = \int_M |\nabla\omega|^2 \, d\mathrm{vol} + \int_M \mathrm{Ric}(\omega^\sharp, \omega^\sharp) \, d\mathrm{vol}.$$

Both terms on the right hand side are now manifestly nonnegative, so they must both vanish, implying in particular that ω is a parallel section of $\Lambda^1 T^*M = T^*M$. In general, the space of parallel sections of a vector bundle over a connected manifold can never be very large: each such section is uniquely determined by its value at one point, and we conclude that the space of harmonic 1-forms on M cannot be larger than the rank of the bundle T^*M, which is n. If additionally $\mathrm{Ric} > 0$ at some point (and therefore on a nonempty open set), it follows that $\omega^\sharp$ must vanish on that set, and since ω is parallel, this implies that ω vanishes identically, proving that the space of harmonic 1-forms is trivial. $\square$

Bibliography

[AF03] R. A. Adams and J. J. F. Fournier, *Sobolev spaces*, 2nd ed., Pure and Applied Mathematics (Amsterdam), vol. 140, Elsevier/Academic Press, Amsterdam, 2003.

[Bär] C. Bär, *Lorentzian Geometry*. lecture notes, available at https://www.math.uni-potsdam.de/professuren/geometrie/lehre/lehrmaterialien/.

[Bau06] H. Baum, *Eine Einführung in die Differentialgeometrie* (2006). Notes from a course at HU Berlin, available at https://www.mathematik.hu-berlin.de/~baum/Skript/diffgeo1.pdf.

[Ber03] M. Berger, *A panoramic view of Riemannian geometry*, Springer-Verlag, Berlin, 2003.

[Boo77] B. Booss, *Topologie und Analysis*, Springer-Verlag, Berlin-New York, 1977 (German). Einführung in die Atiyah-Singer-Indexformel; Hochschultext.

[BB85] B. Booss and D. D. Bleecker, *Topology and analysis*, Universitext, Springer-Verlag, New York, 1985. The Atiyah-Singer index formula and gauge-theoretic physics; Translated from the German by Bleecker and A. Mader.

[BT82] R. Bott and L. W. Tu, *Differential forms in algebraic topology*, Graduate Texts in Mathematics, vol. 82, Springer-Verlag, New York-Berlin, 1982.

[Bou] N. Boumal, *An introduction to optimization on smooth manifolds*. Available at https://sma.epfl.ch/~nboumal/book/; to appear in Cambridge University Press.

[Bre93] G. E. Bredon, *Topology and geometry*, Springer-Verlag, New York, 1993.

[CR53] E. Calabi and M. Rosenlicht, *Complex analytic manifolds without countable base*, Proc. Amer. Math. Soc. **4** (1953), 335–340.

[CE12] K. Cieliebak and Y. Eliashberg, *From Stein to Weinstein and back: symplectic geometry of affine complex manifolds*, American Mathematical Society Colloquium Publications, vol. 59, American Mathematical Society, Providence, RI, 2012.

[dR84] G. de Rham, *Differentiable manifolds*, Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 266, Springer-Verlag, Berlin, 1984. Forms, currents, harmonic forms; Translated from the French by F. R. Smith; With an introduction by S. S. Chern.

[DK90] S. K. Donaldson and P. B. Kronheimer, *The geometry of four-manifolds*, Oxford Mathematical Monographs, The Clarendon Press Oxford University Press, New York, 1990. Oxford Science Publications.

[DK00] J. J. Duistermaat and J. A. C. Kolk, *Lie groups*, Universitext, Springer-Verlag, Berlin, 2000.

[Ebe] J. Ebert, *A lecture course on the Atiyah-Singer index theorem*. notes available at https://ivv5hpp.uni-muenster.de/u/jeber_02/skripten/mainfile.pdf.

[Gar11] D. J. H. Garling, *Clifford algebras: an introduction*, London Mathematical Society Student Texts, vol. 78, Cambridge University Press, Cambridge, 2011.

[Gro85] M. Gromov, *Pseudoholomorphic curves in symplectic manifolds*, Invent. Math. **82** (1985), no. 2, 307–347.

[Hal15] B. Hall, *Lie groups, Lie algebras, and representations*, 2nd ed., Graduate Texts in Mathematics, vol. 222, Springer, Cham, 2015. An elementary introduction.

[Hat02] A. Hatcher, *Algebraic topology*, Cambridge University Press, Cambridge, 2002.

[Hir94] M. W. Hirsch, *Differential topology*, Springer-Verlag, New York, 1994.

[Jän05] K. Jänich, *Topologie*, 8th ed., Springer-Verlag, Berlin, 2005 (German).

[Jos17] J. Jost, *Riemannian geometry and geometric analysis*, 7th ed., Universitext, Springer, Cham, 2017.

[Kel75] J. L. Kelley, *General topology*, Graduate Texts in Mathematics, vol. 27, Springer-Verlag, New York, 1975. Reprint of the 1955 edition [Van Nostrand, Toronto, Ont.]

[Kob95] S. Kobayashi, *Transformation groups in differential geometry*, Classics in Mathematics, Springer-Verlag, Berlin, 1995. Reprint of the 1972 edition.

[LM89] H. B. Lawson Jr. and M.-L. Michelsohn, *Spin geometry*, Princeton Mathematical Series, vol. 38, Princeton University Press, Princeton, NJ, 1989.

[Lee11] J. M. Lee, *Introduction to topological manifolds*, 2nd ed., Graduate Texts in Mathematics, vol. 202, Springer, New York, 2011.

[Lee13a] ———, *Introduction to smooth manifolds*, 2nd ed., Graduate Texts in Mathematics, vol. 218, Springer, New York, 2013.
[Lee13b] ———, *Axiomatic geometry*, Pure and Applied Undergraduate Texts, vol. 21, American Mathematical Society, Providence, RI, 2013.
[Lee18] ———, *Introduction to Riemannian manifolds*, Graduate Texts in Mathematics, vol. 176, Springer, Cham, 2018. Second edition.
[LL01] E. H. Lieb and M. Loss, *Analysis*, 2nd ed., Graduate Studies in Mathematics, vol. 14, American Mathematical Society, Providence, RI, 2001.
[MS17] D. McDuff and D. Salamon, *Introduction to symplectic topology*, 3rd ed., Oxford University Press, 2017.
[Mil56] J. Milnor, *Construction of universal bundles. II*, Ann. of Math. (2) **63** (1956), 430–436.
[Mil76] ———, *Curvatures of left invariant metrics on Lie groups*, Advances in Math. **21** (1976), no. 3, 293–329.
[Mil97] J. W. Milnor, *Topology from the differentiable viewpoint*, Princeton Landmarks in Mathematics, Princeton University Press, Princeton, NJ, 1997. Based on notes by David W. Weaver; Revised reprint of the 1965 original.
[MS74] J. W. Milnor and J. D. Stasheff, *Characteristic classes*, Princeton University Press, Princeton, N. J., 1974. Annals of Mathematics Studies, No. 76.
[Moi77] E. E. Moise, *Geometric topology in dimensions* 2 *and* 3, Springer-Verlag, New York-Heidelberg, 1977. Graduate Texts in Mathematics, Vol. 47.
[MS39] S. B. Myers and N. E. Steenrod, *The group of isometries of a Riemannian manifold*, Ann. of Math. (2) **40** (1939), no. 2, 400–416.
[nLa] nLab, *Shrinking lemma*. exposition based on a blog post by Matt Rosenzweig, available at https://ncatlab.org/nlab/show/shrinking+lemma.
[Olv86] P. J. Olver, *Applications of Lie groups to differential equations*, Graduate Texts in Mathematics, vol. 107, Springer-Verlag, New York, 1986.
[RS80] M. Reed and B. Simon, *Methods of modern mathematical physics. I*, 2nd ed., Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York, 1980. Functional analysis.
[Roe98] J. Roe, *Elliptic operators, topology and asymptotic methods*, 2nd ed., Pitman Research Notes in Mathematics Series, vol. 395, Longman, Harlow, 1998.
[Rud69] M. E. Rudin, *A new proof that metric spaces are paracompact*, Proc. Amer. Math. Soc. **20** (1969), 603.
[Sal16] D. A. Salamon, *Measure and integration*, EMS Textbooks in Mathematics, European Mathematical Society (EMS), Zürich, 2016. MR3469972
[Sán01] M. Sánchez, *Geodesic connectedness of semi-Riemannian manifolds*, Proceedings of the Third World Congress of Nonlinear Analysts, Part 5 (Catania, 2000), 2001, pp. 3085–3102.
[Sei67] H.-J. Seifert, *Global connectivity by timelike geodesics*, Z. Naturforsch **22a** (1967), 1356–1360.
[Spi99a] M. Spivak, *A comprehensive introduction to differential geometry*, 3rd ed., Vol. 1, Publish or Perish Inc., Houston, TX, 1999.
[Spi99b] ———, *A comprehensive introduction to differential geometry*, 3rd ed., Vol. 3, Publish or Perish Inc., Houston, TX, 1999.
[Ste51] N. Steenrod, *The Topology of Fibre Bundles*, Princeton University Press, Princeton, N. J., 1951.
[Tho54] R. Thom, *Quelques propriétés globales des variétés différentiables*, Comment. Math. Helv. **28** (1954), 17–86 (French).
[Tro92] A. J. Tromba, *Teichmüller theory in Riemannian geometry*, Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 1992.
[Tu17] L. W. Tu, *Differential geometry*, Graduate Texts in Mathematics, vol. 275, Springer, Cham, 2017. Connections, curvature, and characteristic classes.
[Var04] V. S. Varadarajan, *Supersymmetry for mathematicians: an introduction*, Courant Lecture Notes in Mathematics, vol. 11, New York University, Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence, RI, 2004.
[War83] F. W. Warner, *Foundations of differentiable manifolds and Lie groups*, Graduate Texts in Mathematics, vol. 94, Springer-Verlag, New York-Berlin, 1983. Corrected reprint of the 1971 edition.
[Wen18] C. Wendl, *Topology I and II* (2018). Notes from the course at HU Berlin, available at https://www.mathematik.hu-berlin.de/~wendl/Winter2018/Topologie2/lecturenotes.pdf.
[Wen20] ———, *Lebesgue, Fourier and Sobolev (notes for functional analysis)* (2020). Available at https://www.mathematik.hu-berlin.de/~wendl/Winter2020/FunkAna/lecturenotes.pdf.
[Wen19] ———, *Integration auf Untermannigfaltigkeiten* (2019). Skript zur Vorlesung Analysis III an der HU Berlin, verfügbar unter https://www.mathematik.hu-berlin.de/~wendl/Winter2019/Analysis3/Skript_DifferentialFormen.pd