

# Introduction to Gauge Theory

Ross Dempsey

Revised February 14, 2019

## Abstract

Twentieth century physics began with the shocking revolutions of quantum mechanics and special relativity. These discoveries, which at first confounded physical understanding, were eventually united in quantum field theory. Quantum field theory was immediately successful in describing quantum effects in electrodynamics. We now know that it also describes the weak and strong nuclear forces, albeit in a more complicated manner. This discovery, the Standard Model of particle physics, unexpectedly revealed a unifying principle known as gauge symmetry. In these notes, we will define and explain gauge symmetry in a classical setting, and show how the gauge principle leads to physical theories. We will also explore some of the effects which arise in quantum gauge theories.

## Contents

<b>1</b>	<b>Motivation</b>	<b>3</b>
1.1	Relativistic Electrodynamics . . . . .	4
1.2	Hamiltonian Electrodynamics . . . . .	5
<b>2</b>	<b>Manifolds and Bundles</b>	<b>8</b>
2.1	Manifolds . . . . .	8
2.2	Bundles . . . . .	11
2.3	Differential Forms . . . . .	14
<b>3</b>	<b>Connections on Bundles</b>	<b>18</b>
3.1	Vector Bundles . . . . .	18
3.2	Connections . . . . .	19
3.3	Curvature . . . . .	22
3.4	Line Bundles and Electrodynamics . . . . .	23

<b>4</b>	<b>Principal Bundles</b>	<b>24</b>
4.1	Lie Groups . . . . .	25
4.2	Lie Algebras . . . . .	28
4.3	Principal Bundles . . . . .	35
<b>5</b>	<b>Electrodynamics as a Gauge Theory</b>	<b>39</b>
<b>6</b>	<b>Yang-Mills Theories</b>	<b>41</b>
<b>7</b>	<b>Reduction of Symmetry</b>	<b>41</b>
<b>8</b>	<b>Renormalization of Gauge Couplings</b>	<b>41</b>
<b>9</b>	<b>Wilson Loops</b>	<b>41</b>
<b>10</b>	<b>Lattice Gauge Theory</b>	<b>41</b>

# 1 Motivation

Everyone knows at least one gauge theory: classical electromagnetism. Take a look at Maxwell's equations for the  $\mathbf{E}$  and  $\mathbf{B}$  fields:

$$\begin{aligned}\nabla \cdot \mathbf{E} &= 4\pi\rho & \nabla \cdot \mathbf{B} &= 0 \\ \nabla \times \mathbf{E} &= -\frac{1}{c} \frac{\partial \mathbf{B}}{\partial t} & \nabla \times \mathbf{B} &= 4\pi\mathbf{j}.\end{aligned}$$

These equations can be divided into two groups. Two of them involve source terms,  $\rho$  and  $\mathbf{j}$ . These are the equations which encode the real physics. The other two are constraint equations for the fields. These constraints can be made manifest by choosing a particular representation for the fields. By letting

$$\mathbf{E} = -\nabla\phi - \frac{1}{c} \frac{\partial \mathbf{A}}{\partial t}, \quad \mathbf{B} = \nabla \times \mathbf{A},$$

we automatically have

$$\begin{aligned}\nabla \times \mathbf{E} &= -\frac{1}{c} \nabla \times \frac{\partial \mathbf{A}}{\partial t} = -\frac{1}{c} \frac{\partial}{\partial t} (\nabla \times \mathbf{A}) = -\frac{1}{c} \frac{\partial \mathbf{B}}{\partial t}, \\ \nabla \cdot \mathbf{B} &= \nabla \cdot (\nabla \times \mathbf{A}) = 0.\end{aligned}$$

This representation comes with a caveat. The physical degrees of freedom are the fields  $\mathbf{E}$  and  $\mathbf{B}$ ; the potentials  $\phi$  and  $\mathbf{A}$  are not directly physical. This means that if we change  $\phi$  and  $\mathbf{A}$  without changing  $\mathbf{E}$  and  $\mathbf{B}$ , then we are looking at a different representation of the same physical situation. In fact, we can make such a change of representation with ease. If we add a gradient to  $\mathbf{A}$ ,  $\mathbf{A} \mapsto \mathbf{A} + \nabla\chi$ , then  $\mathbf{B} = \nabla \times \mathbf{A}$  is unchanged. To fix  $\mathbf{E}$  to be unchanged as well, we prescribe  $\phi \mapsto \phi - \frac{1}{c} \frac{\partial \chi}{\partial t}$ . In summary, we have

$$\begin{aligned}\mathbf{E} &\mapsto -\nabla \left( \phi - \frac{1}{c} \frac{\partial \chi}{\partial t} \right) - \frac{1}{c} \frac{\partial}{\partial t} (\mathbf{A} + \nabla\chi) = -\nabla\phi - \frac{1}{c} \frac{\partial \mathbf{A}}{\partial t} = \mathbf{E}, \\ \mathbf{B} &\mapsto \nabla \times (\mathbf{A} + \nabla\chi) = \nabla \times \mathbf{A} = \mathbf{B}.\end{aligned}$$

This is called a *gauge symmetry*. A gauge symmetry is an internal symmetry, in which a physical system is given a many-to-one mathematical representation. Additionally, gauge symmetries are local, a concept we will explore in much more detail later; here, we see locality from the spacetime dependence of the function  $\chi$ .

Gauge symmetry, presented in this way, is either a curiosity or a minor annoyance. We will show first that this symmetry is made manifest in the relativistic treatment of electrodynamics, lending a bit more credence to its importance. We will then look at the Hamiltonian formulation of electrodynamics and its quantum mechanical consequences, showing the centrality of the scalar and vector potentials and the significance of the gauge symmetry.

## 1.1 Relativistic Electrodynamics

In relativistic electrodynamics, we treat charge and current density as components of a single four-vector, called the four-current. It is a simple exercise to show that the combination  $\begin{pmatrix} \rho c \\ \mathbf{j} \end{pmatrix}$  in fact forms a Lorentz vector, transforming in the appropriate way under a Lorentz transformation. We know from electrostatics that the scalar potential satisfies  $\nabla^2 \phi = -4\pi\rho$ . If we solve a similar equation for each component of the current density,  $\nabla^2 \mathbf{A} = -\frac{4\pi}{c}\mathbf{j}$ , then we obtain a vector potential  $\mathbf{A}$ . The combination of the scalar and the vector potentials forms a four-vector called the four-potential,

$$A^\mu = \begin{pmatrix} \phi \\ \mathbf{A} \end{pmatrix}.$$

The significance of the vector potential is not immediately clear from this definition of it. By integrating the Poisson equation, we have

$$\mathbf{A}(\mathbf{r}) = \int d^3\mathbf{r}' \frac{\mathbf{j}(\mathbf{r}')}{c|\mathbf{r} - \mathbf{r}'|}.$$

Taking the curl, we have

$$\nabla \times \mathbf{A} = \frac{1}{c} \int d^3\mathbf{r}' \frac{(\mathbf{r}' - \mathbf{r}) \times \mathbf{j}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3},$$

which is the Biot-Savart law for the magnetic field  $\mathbf{B}$ . In coordinates, we have

$$\begin{aligned} B_x &= \frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}, \\ B_y &= \frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}, \\ B_z &= \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}. \end{aligned}$$

In contrast, from electrostatics, the components of the electric field are  $E_i = -\frac{\partial \phi}{\partial x^i}$ . However, since this comes from electrostatics, it is not sensitive to terms which may arise from time dependence. If we take a leap of faith, and prescribe that the electric field is given by

$$E_i = -\frac{\partial \phi}{\partial x^i} - \frac{1}{c} \frac{\partial A_i}{\partial t},$$

then the electric and magnetic field components both arise as combinations of derivatives of the four potential. In fact, if we define the tensor

$$F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu = \partial^{[\mu} A^{\nu]},$$

then the field components are exactly its components:

$$F = \begin{pmatrix} 0 & E_x & E_y & E_z \\ -E_x & 0 & B_z & -B_y \\ -E_y & -B_z & 0 & B_x \\ -E_z & B_y & -B_x & 0 \end{pmatrix}.$$

Clearly, this tensor – known as the *field-strength tensor* – contains all the variables of physical importance. Additionally, it has a manifest symmetry. If we vary the four-potential by  $A^\mu \mapsto A^\mu + \partial^\mu \chi$ , then

$$F^{\mu\nu} \mapsto \partial^\mu(A^\nu + \partial^\nu \chi) - \partial^\nu(A^\mu + \partial^\mu \chi) = \partial^\mu A^\nu - \partial^\nu A^\mu = F^{\mu\nu}.$$

It is simple to verify that this transformation is exactly the same as the one we defined for the scalar and vector potentials individually, but now its covariant form is made clear.

It is worth noticing at this point that the gauge symmetry and the conservation of charge are cut from the same cloth: the antisymmetry of the field-strength tensor. The above argument follows because the added terms cancel, due to antisymmetry. To establish conservation of charge, we look at the equations of motion for the fields, which are given by

$$\partial_\mu F^{\mu\nu} = \frac{4\pi}{c} j^\nu.$$

If we take another derivative of this equation, then we find

$$\partial_\nu j^\nu = \frac{c}{4\pi} \partial_\mu \partial_\nu F^{\mu\nu} = 0,$$

by antisymmetry. But this is the continuity equation for charge:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{j} = 0.$$

## 1.2 Hamiltonian Electrodynamics

We will now shift gears from the physics of the EM field itself to its effect on charged particles. The Lorentz force law gives

$$\mathbf{F} = q \left( \mathbf{E} + \frac{\mathbf{v}}{c} \times \mathbf{B} \right).$$

It is not obvious how to form a Lagrangian, since the Lorentz force is velocity-dependent:

$$\mathbf{F} = q \left( \mathbf{E} + \frac{\dot{\mathbf{x}}}{c} \times \mathbf{B} \right) = q \left( -\nabla \phi - \frac{1}{c} \frac{d\mathbf{A}}{dt} + \frac{1}{c} \nabla(\mathbf{v} \cdot \mathbf{A}) \right).$$

The second equality is nontrivial; you should apply cross product identities and work it out for yourself. It turns out that the correct Lagrangian is

$$L(\mathbf{x}, \dot{\mathbf{x}}) = \frac{1}{2} m \dot{\mathbf{x}}^2 - q\phi(\mathbf{x}) + \frac{q}{c} \dot{\mathbf{x}} \cdot \mathbf{A}.$$

To see this, we form the Euler-Lagrange equations:

$$\frac{d}{dt} \left( m \dot{\mathbf{x}} + \frac{q}{c} \mathbf{A} \right) + q \left( \nabla \phi - \frac{1}{c} \nabla(\mathbf{v} \cdot \mathbf{A}) \right) = 0.$$

Rearranging this gives the Lorentz force law as written above.

Now that we have a Lagrangian, we can construct the Hamiltonian. The canonical momentum is

$$\mathbf{p} = \frac{\partial L}{\partial \dot{\mathbf{x}}} = m\dot{\mathbf{x}} + \frac{q}{c}\mathbf{A}.$$

The Hamiltonian is then

$$H = \mathbf{p} \cdot \dot{\mathbf{x}} - L = \frac{1}{2}m\dot{\mathbf{x}}^2 + q\phi(\mathbf{x}).$$

This seems to be missing information about the magnetic field. However, we have to express the Hamiltonian in terms of momentum, not velocity. Making this adjustment, we have

$$H = \frac{(\mathbf{p} - \frac{q}{c}\mathbf{A})^2}{2m} + q\phi(\mathbf{x}).$$

When we quantize the particle in an electromagnetic field, we use this Hamiltonian. The Schrödinger equation reads

$$\left[ \frac{1}{2m} \left( -i\hbar\nabla - \frac{q}{c}\mathbf{A} \right)^2 + q\phi(\mathbf{x}) \right] \psi(\mathbf{x}) = E\psi(\mathbf{x}).$$

Clearly there is some uncomfortable mixing of the gradient with the vector potential  $\mathbf{A}$ . We can remove this by defining

$$\psi(\mathbf{x}) = e^{\frac{iq}{\hbar c} \int_{\gamma} \mathbf{A} \cdot d\mathbf{x}} \tilde{\psi}(\mathbf{x}).$$

Substituting this in, the derivative acting on the exponential cancels the vector potential term, so  $\tilde{\psi}$  satisfies the normal Schrödinger equation. Thus, the effect of the vector potential is to add a phase to the wavefunction. Typically, a phase in a wavefunction is immaterial. However, if we move the particle in a closed path, then there is a phase  $e^{\frac{iq}{\hbar c} \oint \mathbf{A} \cdot d\mathbf{x}}$  which is nontrivial.

This phase is physical, but it also depends on the gauge-dependent quantity  $\mathbf{A}$ . This is reconciled by the fact that

$$\oint \mathbf{A} \cdot d\mathbf{x}$$

is in fact gauge-invariant. Indeed, it is the magnetic flux through the region enclosed by the path.

**Example 1.1.** Show that the time-dependent Schrödinger equation of a particle in an electromagnetic field is gauge invariant if the gauge transformations are amended to include a phase shift in the wavefunction.

**Solution:** Applying a gauge transformation to the time-dependent Schrödinger equation, we have

$$\left[ -\frac{\hbar^2}{2m} \left( -i\hbar\nabla - \frac{q}{c}(\mathbf{A} + \nabla\chi) \right)^2 + q \left( \phi - \frac{1}{c} \frac{\partial\chi}{\partial t} \right) \right] e^{i\lambda}\psi(\mathbf{x}) = i\hbar \frac{d}{dt} \left( e^{i\lambda}\psi(\mathbf{x}) \right),$$

where  $\lambda$  is some phase factor depending on the gauge function  $\chi$ . If the Schrödinger equation is to be gauge invariant, we must satisfy

$$\begin{aligned} \left( -i\hbar\nabla - \frac{q}{c}\nabla\chi \right) e^{i\lambda} &= 0, \\ i\hbar \frac{d}{dt} e^{i\lambda} &= -\frac{q}{c} \frac{\partial\chi}{\partial t} e^{i\lambda}. \end{aligned}$$

These equations are both satisfied by  $\lambda = \frac{q}{\hbar c}\chi$ . Therefore, the gauge transformation takes

$$\psi(\mathbf{x}) \mapsto e^{\frac{iq}{\hbar c}\chi}\psi(\mathbf{x}),$$

in addition to the usual transformations of  $\phi$  and  $\mathbf{A}$ .

The observation in the previous example allows us to reformulate what we mean by the gauge symmetry of electromagnetism. The symmetry of adding a gradient to the four-potential is somewhat difficult to put a finger on; exactly how much freedom does it entail? In comparison, the action of the gauge symmetry on the wavefunction is simple: we can multiply the wavefunction by a phase which varies from point to point in spacetime. Indeed, by shuffling constants we can write the gauge symmetry as

$$\begin{aligned}\psi(x^\mu) &\mapsto e^{i\lambda(x^\mu)}\psi(x^\mu), \\ A^\mu &\mapsto A^\mu + \frac{\hbar c}{q}\partial^\mu\lambda.\end{aligned}$$

Written in this way, we see that after choosing a phase  $e^{i\lambda(x^\mu)}$  at every point, the gauge transformation is fixed.

This is why we say that electromagnetism is a  $U(1)$  gauge theory. The group  $U(1)$ , meaning the unitary group over  $\mathbb{C}^1$ , is the group of complex phases (isomorphic to the circle group). A gauge transformation in electromagnetism is fixed by choosing an element of  $U(1)$  at every spacetime point.

This is a relatively simple idea; but not all groups are as simple as  $U(1)$ . In the following several sections, we will develop the theory of principal bundles, which are mathematical objects uniquely suited to describe symmetry groups acting locally on a spacetime manifold.

**Example 1.2.** The idea of gauge symmetry does not apply solely to physics. A local internal symmetry is also present in foreign exchange markets, as pointed out by [1]. Consider a discrete collection  $W$  of points, called countries, with a function  $\phi : W \times W \rightarrow \mathbb{R}$ , called the exchange rate. First argue that the “important” (i.e. profitable) quantities are not the values of  $\phi(w_1, w_2)$  but rather the arbitrage products

$$P(w_1, w_2, w_3) = \phi(w_1, w_2)\phi(w_2, w_3)\phi(w_3, w_1).$$

Then show that a gauge symmetry is given by

$$\phi(w_1, w_2) \mapsto \phi(w_1, w_2) \times \frac{\chi(w_1)}{\chi(w_2)}.$$

**Solution:** An exchange rate itself is not an important quantity. For example, at the time of writing,  $\phi(\text{USA}, \text{India}) = 72.47$  (meaning 1 USD = 72.47 rupee); this is just a definition of one currency in terms of the other. However, if we had three countries  $A, B, C$ , such that

$$\phi(A, B)\phi(B, C)\phi(C, A) \neq 1,$$

then by making a triangle of currency exchanges we could create money out of thin air (i.e., there is potential for arbitrage).

It would make no difference to currency exchanges if every country were to make an arbitrary rescaling of its currency. For example, if the United States started using the dime as the fundamental unit of currency, then we would say  $\phi(\text{USA}, \text{India}) = 7.247$  and there would be no real change. If every country  $w$  scales up the value of its currency by  $\chi(w)$ , then the exchange rates scale as

$$\phi(A, B) \mapsto \frac{\chi(A)}{\chi(B)},$$

and the arbitrage potential is manifestly unaffected.

For a fuller discussion of this concept, including the importance of time variation in the exchange rates, see [1].

## 2 Manifolds and Bundles

In the last section, we described gauge symmetry as a local and internal symmetry. In the next few sections, we will be developing mathematical machinery to handle this kind of symmetry. The general approach will be to take a spacetime manifold, and attach to each point a full symmetry group, so that we can choose a gauge by choosing an element of the symmetry group at each point.

### 2.1 Manifolds

A manifold is a generalization of familiar  $n$ -dimensional space. In  $\mathbb{R}^n$ , the coordinates for a given point are obvious; points are labeled by their coordinates. For a manifold, we allow a much more general starting point: a topological space. A topological space is given by a set of points,  $X$ , together with a specification of the open subsets of  $X$ , satisfying some consistency conditions.

The freedom to choose the open sets may seem unfamiliar. Typically, we are given a metric  $d(x, y)$  on a space, and then the open sets  $U$  are ones for which, for all  $x \in U$ , there exists  $\epsilon > 0$  such that  $B(x, \epsilon) \subset U$ . Intuitively, open sets are ones which do not contain their boundaries; every point is in the interior.

This is a particular topology known as the metric topology. It is not the only topology we can choose for a given set of points. For example, consider the discrete topology, in which all subsets of  $X$  are open. In particular, singleton sets  $\{x\} \subset X$  are open. This would only happen in a metric topology if  $d(x, y) > \epsilon$  for some fixed  $\epsilon > 0$  and all  $y \in X$ , meaning that  $x$  has a ball around it containing no other points. Thus, we think of the discrete topology as the topology in which every point is isolated.

This example shows that specifying a topology is akin to specifying the shape of a set, without specifying its exact metric structure. Indeed, there are topologies which cannot be derived from a metric, though we will not be especially concerned with these. Think of a topological space as a stretchy sort of object, where only non-metric concepts like continuity make sense.



**Example 2.1.** A topology must satisfy the following two constraints:

- (1) Any union of open sets,  $\bigcup_{i \in I} U_i$  (where  $I$  is an arbitrary index set), is open.
- (2) Any finite intersection of open sets,  $\bigcap_{i=1}^n U_i$ , is open.

Show that any metric topology satisfies these constraints.

**Solution:** Let

$$x \in \bigcup_{i \in I} U_i,$$

where all the  $U_i$  are open. Then there is some  $j \in I$  for which  $x \in U_j$ . Since  $U_j$  is open, it follows that there exists  $\epsilon > 0$  for which  $B(x, \epsilon) \subset U_j$ , and it follows that  $B(x, \epsilon) \subset \bigcup_{i \in I} U_i$ , showing that the union is open.

Likewise, let

$$x \in \bigcap_{i=1}^n U_i.$$

Then  $x \in U_i$  for all  $i = 1, \dots, n$ , and so there exist numbers  $\epsilon_i > 0$  such that  $B(x, \epsilon_i) \subset U_i$  for all  $i = 1, \dots, n$ . Let  $\epsilon = \min(\epsilon_1, \dots, \epsilon_n)$ . Then  $B(x, \epsilon) \subset \bigcap_{i=1}^n U_i$ , showing that the intersection is closed.

We can have functions  $f : X \rightarrow Y$  from one topological space to another. A topology is sufficient to define when a function is continuous; we say  $f$  is continuous if, whenever  $V \subset Y$  is an open set, so too is  $f^{-1}(V) \subset X$ . You should show that this aligns with the typical  $\delta$ - $\epsilon$  definition of continuity for functions  $f : \mathbb{R} \rightarrow \mathbb{R}$ . If there is a bijection  $f : X \rightarrow Y$  between topological spaces, such that both  $f$  and  $f^{-1}$  are continuous, then we say  $f$  is a *homeomorphism*. When two topological spaces are homeomorphic, they are the same in a topological sense.

Topological spaces are a very wide class of objects, and this class contains some unfriendly creatures. For a topological space  $X$  to be a manifold, we have several extra demands. First, we require it to be Hausdorff, a technical constraint on the topology which will not concern us. More importantly, we require it to be locally homeomorphic to a Euclidean space. By locally homeomorphic, we mean there exists an open cover  $\{U_i\}$  (i.e., a collection of open sets  $U_i$  such that  $\bigcup_i U_i = X$ ) such that each  $U_i$  is homeomorphic to an open subset  $V_i \subset \mathbb{R}^n$ . The functions  $f_i : U_i \rightarrow V_i$  implementing the homeomorphisms are called a *coordinate chart*, and the set of all these functions is called a *coordinate atlas*.

The most trivial example of a manifold is  $\mathbb{R}^n$  itself. It forms a topological space under its metric topology, and an open cover is given by a single open set,  $\mathbb{R}^n$  itself. A chart on  $\mathbb{R}^n$  is simply the identity map.

A more interesting example is the circle  $S^1$  as a one-dimensional manifold. We can put a topology on the circle by first giving it a metric, under which the distance between two points is the angle between them, and then taking the metric topology. However, there is no continuous map from

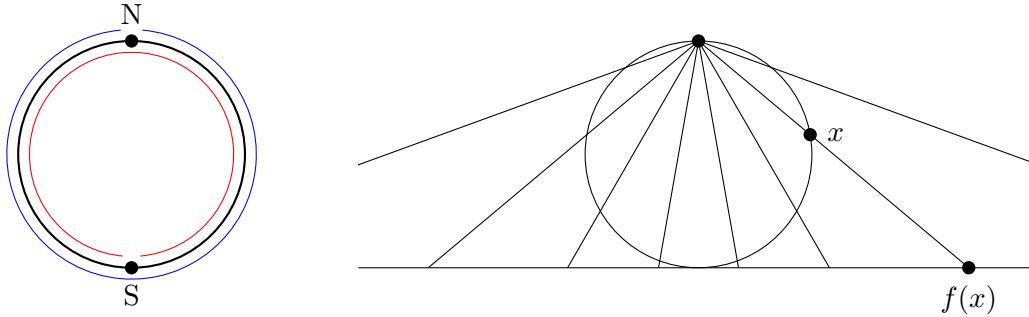


Figure 1: Via stereographic projection, we can map all but one point of a circle to the real line.

the circle to the real line (you should verify this by attempting to construct one), so we need to be more creative in constructing an atlas. Let  $N$  and  $S$  be the north and south points of the circle, and form an open cover by taking the sets  $\{S^1 - N, S^1 - S\}$ . We can map both of these sets to the real line by stereographic projection, as shown in Figure 1. This defines a coordinate atlas on the circle, giving it the structure of a manifold.

Since manifolds are locally homeomorphic to  $\mathbb{R}^n$ , we can require them to inherit desired properties of  $\mathbb{R}^n$ . For example, we almost always require a manifold to be differentiable. Note that we cannot directly require the functions  $f_i$  to be differentiable, because the domain is a topological space, which does not have metric structure. Rather, we require the *transition functions* to be differentiable. The transition functions are

$$f_j \circ f_i^{-1} : f(U_i \cap U_j) \rightarrow f(U_j).$$

Check for yourself that  $f_j \circ f_i^{-1}$  is well-defined throughout  $f(U_i \cap U_j)$ . These functions describe how to connect two different coordinate charts which lie over the same point.

**Example 2.2.** Show that the transition function for the coordinate atlas we defined on the circle maps  $x \mapsto R^2/x$  for some parameter  $R$ .

**Solution:** The transition function is defined over  $f_1(U_1 \cap U_2)$ , which is  $\mathbb{R} - \{0\}$ . To compute the transition function, we have to perform an inverse stereographic projection, followed by a stereographic projection from the opposite side of the circle. This process is shown in Figure 2.

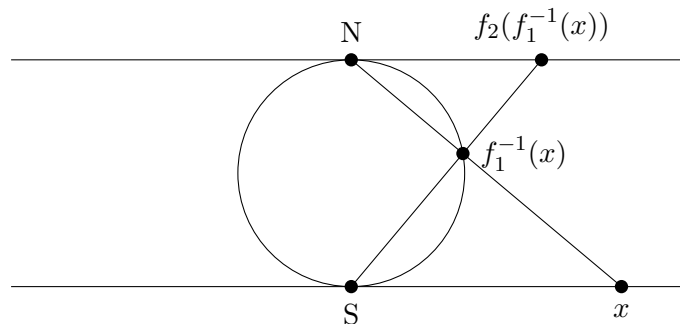


Figure 2

Note that the triangle between points  $N$ ,  $S$ , and  $f_1^{-1}(x)$  is right. Additionally,  $f_2(f_1^{-1}(x)) \propto \tan \angle NSf^{-1}(x)$ , and  $x \propto \tan \angle SNf^{-1}(x)$ . These angles are complementary, so  $f_2(f_1^{-1}(x)) \propto x^{-1}$ .

We can make even more stringent requirements than differentiability. A manifold is smooth if all its transition functions are infinitely differentiable. We will require all manifolds to be smooth in these notes.

## 2.2 Bundles

A bundle is relatively simple to define: it is a map  $\pi : E \rightarrow B$  from a manifold  $E$  to a manifold  $B$ .

There is more here than meets the eye. We call  $E$  the total space,  $B$  the base space, and  $\pi$  the projection. Conceptually, a bundle is a manifold  $B$  to which we attach *fibers*,  $\pi^{-1}(b)$ , over each point  $b \in B$ . For example, there is the trivial bundle where  $E = B \times F$ , and  $\pi : B \times F \rightarrow B$  is the canonical projection. We call  $F$  the fiber space.

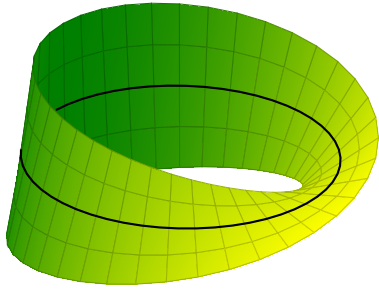
Most bundles we are interested are fiber bundles. Fiber bundles are bundles which are locally equivalent to the trivial bundle. This is similar in nature to the requirement that a manifold be locally homeomorphic to Euclidean space. The role of the coordinate chart is filled by the local trivialization. The idea behind this is that, at every point  $x \in B$ , there should be a neighborhood  $\pi(x) \in U \subset B$  such that  $\pi^{-1}(U)$  looks like a trivial bundle. Formally, we require that there exists a map  $\phi$  for which the diagram

$$\begin{array}{ccc} \pi^{-1}(U) & \overset{\phi}{\dashrightarrow} & U \times F \\ \downarrow \pi & \swarrow \text{proj} & \\ U & & \end{array}$$

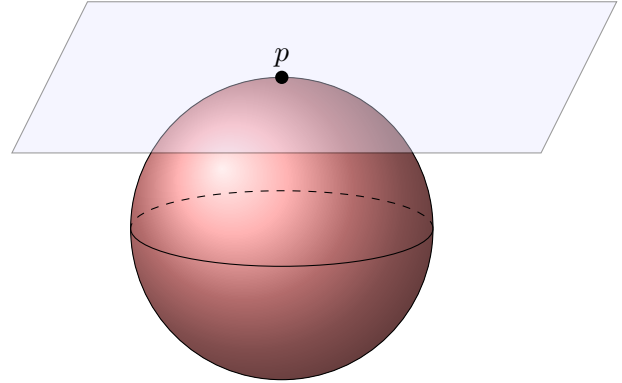
commutes. To understand the meaning of this, follow the arrows in both ways. This says that projecting the fibers  $\pi^{-1}(U)$  down to  $U$  is equivalent to mapping the fibers under  $\phi$  to the trivial bundle  $U \times F$ , and then projecting that bundle down to  $U$ .

Even though a fiber bundle is locally trivial, it need not be globally so. A canonical example of a nontrivial fiber bundle is the Möbius strip. Figure 3a shows a Möbius strip, with a circle marked out. We identify the strip with  $E$  and the circle with  $B$ , and let  $\pi$  be a projection along the grid lines down to  $B$ . The fiber space is a segment of the real line. Clearly  $E \not\cong B \times F$ , since  $B \times F$  would be cylindrical. However, if we take any point along  $B$  and look at a small neighborhood of it, the Möbius strip looks like the trivial bundle in that neighborhood, which is what makes this a fiber bundle.

A *section* of a bundle is a map  $s : B \rightarrow E$  for which  $\pi \circ s$  is the identity on  $B$ . This is a formal way of saying that  $s$  maps points  $x \in B$  to their fibers  $\pi^{-1}(x)$ . For example, a section of a trivial bundle  $E = B \times F$  is given by a function  $B \rightarrow F$ . Going backwards, any function between manifolds can be thought of as a section of a trivial bundle.



(a) A Möbius strip is a fiber bundle with base space  $B = S^1$ .



(b) The tangent space  $T_pM$ , where  $M$  is a sphere.

Figure 3

An important example of a fiber bundle is the tangent bundle on a manifold. The tangent bundle for a manifold  $M$  is one which associates to every point  $p \in M$  its tangent space  $T_pM$ . The tangent space  $T_pM$  is, intuitively, the space of tangent directions to the manifold at  $p$ . When  $M$  is an  $n$ -dimensional real manifold, we have  $T_pM \cong \mathbb{R}^n$ . We think of the tangent space as lying on the manifold at  $p$ , as in Figure 3b.

More precisely, the tangent space is composed of directional derivatives at  $p$ . A directional derivative has no immediate meaning on a manifold, since it does not come equipped with a metric structure of its own. However, via the coordinate atlas, the manifold inherits the structure of  $\mathbb{R}^n$ . That is, given any smooth function  $\phi : M \rightarrow \mathbb{R}$ , we have  $\phi \circ f^{-1} : f(U) \rightarrow \mathbb{R}$ , where  $U$  is an open subset of the manifold and  $f$  is a coordinate chart on it. Since  $f(U) \subset \mathbb{R}^n$ , we can pick a vector  $\mathbf{v} \in \mathbb{R}^n$  and define the directional derivative  $\mathbf{v} \cdot \nabla(\phi \circ f^{-1})$ . The corresponding vector in  $T_pM$  is defined abstractly as the map

$$v : \phi \mapsto \mathbf{v} \cdot \nabla(\phi \circ f^{-1}).$$

Defined in this way, vectors are coordinate-free objects. To give them coordinates, we have to pick a basis for  $T_pM$ . This can be done by using a corresponding basis  $\mathbf{e}_i$  in  $\mathbb{R}^n$ , where  $\mathbf{e}_i$  is the unit vector in the  $x^i$  direction. We can then express a vector  $v$  as  $\sum v^i \mathbf{e}_i$ , where  $v^i$  are numbers and  $\mathbf{e}_i$  are basis vectors in  $T_pM$ .

If we change the coordinates on  $\mathbb{R}^n$ , we will also change the coordinates of vectors in  $T_pM$ . To see how this works, note that  $\mathbf{e}_i$  is simply the directional derivative in the  $x^i$  direction, or  $\frac{\partial}{\partial x^i}$ . Thus, we have

$$v = \sum v^i \frac{\partial}{\partial x^i}.$$

If we make a coordinate transformation  $x^i(\tilde{x}^j)$ , then we have

$$v = \sum v^i \frac{\partial}{\partial x^i} = \sum_i v^i \left( \sum_j \frac{\partial \tilde{x}^j}{\partial x^i} \frac{\partial}{\partial \tilde{x}^j} \right) = \sum_j \left( \sum_i v^i \frac{\partial \tilde{x}^j}{\partial x^i} \right) \frac{\partial}{\partial \tilde{x}^j}.$$

Thus, we should identify the new coordinates as

$$\tilde{v}^j = \sum_i v^i \frac{\partial \tilde{x}^j}{\partial x^i}.$$

This is exactly the transformation law we require for a contravariant vector when we define them in terms of coordinates.

The tangent bundle ties together the tangent spaces at all points of a manifold  $M$  into a single object, denoted  $TM$ . A section of  $TM$  is simply a vector field on  $M$ . This is the cleanest way to think about vector fields on manifolds, and it is important to get comfortable with it. Put another way, a vector field is a map from each point on a manifold to an element of the tangent space at that point.

There is a similar construction, called the cotangent bundle. The cotangent bundle associates every point  $p \in M$  with its cotangent bundle,  $T_p^*M$ , defined as the dual space of  $T_pM$ . (Recall that the dual  $V^*$  of a vector space  $V$  is the vector space of linear functionals on  $V$ ). More concretely, an element  $\omega \in T_p^*M$  is a function  $T_pM \rightarrow \mathbb{R}$  with the properties

$$\omega(\alpha v) = \alpha \omega(v), \quad \omega(v + w) = \omega(v) + \omega(w).$$

The elements of the cotangent space can be built from functions on the manifold. Recall that the tangent space is composed of directional derivatives. Directional derivatives act linearly on functions, and we can turn this statement around to say that functions act linearly on directional derivatives. Concretely, given a function  $f : M \rightarrow \mathbb{R}$ , we have an element  $df \in T_p^*M$ , where

$$df(v) = v(f), \quad \forall v \in T_pM.$$

We call the map  $f \mapsto df$  the differential.

A basis for  $T_p^*M$  is given by  $dx^i$ ,  $i = 1, \dots, n$ ; this is true simply because these are linearly independent (check this) and  $\dim T_p^*M = \dim T_pM = n$ . The action of  $dx^i$  on a vector  $v$  is given by  $dx^i(v) = v(x^i) = v^i$ . To write  $df$  in this basis, we use

$$df(v) = v(f) = \sum v^i e_i(f) = \sum v^i \frac{\partial f}{\partial x^i}.$$

It follows that

$$df = \sum_i \frac{\partial f}{\partial x^i} dx^i,$$

as we would expect.

If we change the coordinates via a transformation  $x^i(\tilde{x}^j)$ , then we have

$$df = \sum_i \frac{\partial f}{\partial x^i} dx^i = \sum_i \frac{\partial f}{\partial x^i} \left( \sum_j \frac{\partial x^i}{\partial \tilde{x}^j} d\tilde{x}^j \right) = \sum_j \left( \sum_i \frac{\partial f}{\partial x^i} \frac{\partial x^i}{\partial \tilde{x}^j} \right) d\tilde{x}^j.$$

It follows that the components of  $df$  under the transformation are

$$(df)_j = \sum_i (df)_i \frac{\partial x^i}{\partial \tilde{x}^j}.$$

This is the transformation law for a covariant vector.

This discussion shows that contravariant vector fields are sections of the tangent bundle  $TM$ , and covariant vector fields are sections of the cotangent bundle  $T^*M$ . This is the coordinate-free approach to vectors. We can use this approach to build up tensors of any rank. A tensor at a point  $p$  of rank  $(r, s)$  is an element of

$$\underbrace{TM \otimes \cdots \otimes TM}_r \otimes \underbrace{T^*M \otimes \cdots \otimes T^*M}_s.$$

This coordinate-free approach to tensors has the advantage of focusing on the intrinsic structure of the manifold, rather than being bogged down in indices.

### 2.3 Differential Forms

The cotangent bundle is the simplest example of a space of differential forms. For an  $n$ -dimensional manifold, we define the spaces  $\Omega^p(M)$ , for  $p = 0, \dots, n$ , by

$$\Omega^p(M) = \underbrace{T_p^*M \wedge \cdots \wedge T_p^*M}_p.$$

The wedge product  $\wedge$  of two vector spaces  $V$  and  $W$  is defined as the vector space spanned by all objects of the form  $v \wedge w$ , with  $v \in V$  and  $w \in W$ , where  $v \wedge w = -w \wedge v$ . We call  $\Omega^p(M)$  the space of  $p$ -forms on  $M$ .

A  $p$ -form on  $M$  is equivalent to a totally antisymmetric tensor of rank  $(0, p)$ . A totally antisymmetric tensor is fully specified if we choose its components for strictly increasing index values. For example, if we have a totally antisymmetric tensor  $F$  of rank  $(0, 2)$  in four dimensions, then it is specified by its components  $F_{01}, F_{02}, F_{03}, F_{12}, F_{13}, F_{23}$ . Generalizing this argument, we see that  $\dim \Omega^p(M) = \binom{n}{p}$ .

For example, take  $n = 3$ . The differential forms have the structure:

$$\begin{aligned} \text{0-forms : } & f \\ \text{1-forms : } & a_x dx + a_y dy + a_z dz \\ \text{2-forms : } & A_z dx \wedge dy + A_y dz \wedge dx + A_x dy \wedge dz \\ \text{3-forms : } & F dx \wedge dy \wedge dz. \end{aligned}$$

We connect differential forms together via a map  $d : \Omega^p(M) \rightarrow \Omega^{p+1}(M)$ . We have already seen the example  $d : \Omega^0(M) \rightarrow \Omega^1(M)$ , which we called the differential. It maps

$$f \mapsto \sum_i \frac{\partial f}{\partial x^i} dx^i.$$

To define the exterior derivative for higher  $p$ -forms, we make the following definitions:  $d(df) = 0$  for any smooth function  $f$ , and  $d(\alpha \wedge \beta) = (d\alpha) \wedge \beta + (-1)^p \alpha \wedge (d\beta)$ , where  $\alpha$  is a  $p$ -form. These

two facts uniquely specify the map  $d$ . For example, we can compute the exterior derivative of a 1-form as follows:

$$\begin{aligned} d\left(\sum_i a_i dx^i\right) &= \sum_i (d(a_i) \wedge dx^i + a_i d(dx^i)) \\ &= \sum_i \sum_j \left(\frac{\partial a_i}{\partial x^j} dx^j\right) \wedge dx^i \\ &= \sum_{i < j} \left(\frac{\partial a_j}{\partial x^i} - \frac{\partial a_i}{\partial x^j}\right) dx^i \wedge dx^j. \end{aligned}$$

If we specialize to three dimensions, the exterior derivative becomes recognizable. Consider its action on 0-forms, 1-forms, and 2-forms:

$$\begin{aligned} f &\mapsto \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy + \frac{\partial f}{\partial z} dz \\ a_x dx + a_y dy + a_z dz &\mapsto \left(\frac{\partial a_y}{\partial x} - \frac{\partial a_x}{\partial y}\right) dx \wedge dy + \left(\frac{\partial a_x}{\partial z} - \frac{\partial a_z}{\partial x}\right) dz \wedge dx + \left(\frac{\partial a_z}{\partial y} - \frac{\partial a_y}{\partial z}\right) dy \wedge dz \\ A_z dx \wedge dy + A_y dz \wedge dx + A_x dy \wedge dz &\mapsto \left(\frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}\right) dx \wedge dy \wedge dz. \end{aligned}$$

Remarkably, the exterior derivative seems to reproduce the gradient, curl, and divergence. The only discrepancy is that the “curl” maps 1-forms to 2-forms, while we expect it to map vectors to vectors, and the “divergence” maps 2-forms to 3-forms, when we expect it to map vectors to scalars.

This concern is resolved by Hodge duality. Since  $\dim \Omega^p(M) = \binom{n}{p} = \binom{n}{n-p} = \dim \Omega^{n-p}(M)$ , we can construct an isomorphism between  $\Omega^p(M)$  and  $\Omega^{n-p}(M)$ . In the case of three dimensions, Hodge duality relates 0-forms to 3-forms and 1-forms to 2-forms. More precisely, the Hodge star operator acts on an orthonormal basis  $(dx^1, \dots, dx^n)$  by

$$\star(dx^1 \wedge \dots \wedge dx^k) = dx^{k+1} \wedge \dots \wedge dx^n.$$

If the metric is not positive-definite (e.g., the Minkowski metric), we have to be a bit more careful. A general expression for the Hodge star is

$$\star(dx^{i_1} \wedge \dots \wedge dx^{i_k}) = \frac{\sqrt{\det g}}{(n-k)!} \epsilon_{j_{k+1} \dots j_n}^{i_1 \dots i_k} dx^{j_{k+1}} \wedge \dots \wedge dx^{j_n},$$

where  $\epsilon_{i_1 \dots i_n}$  is the Levi-Civita symbol and indices have been raised using the metric.

**Example 2.3.** Let  $a = a_\mu dx^\mu$  and  $b = b_\mu dx^\mu$  be 1-forms in four dimensions, and let the metric be the Minkowski metric  $\eta_{\mu\nu}$  with signature  $(+, -, -, -)$ . Show that the inner product  $a^\mu b_\mu$  can be expressed as  $\star(a \wedge \star b)$ .

**Solution:** Using the expression above, the Hodge star acting on the basis vectors is given by

$$\begin{aligned} \star dt &= dx \wedge dy \wedge dz, & \star dx &= dt \wedge dy \wedge dz, \\ \star dy &= -dt \wedge dx \wedge dz, & \star dz &= dt \wedge dx \wedge dy. \end{aligned}$$

Therefore,

$$a \wedge \star b = (a_0 b_0 - a_1 b_1 - a_2 b_2 - a_3 b_3)(dt \wedge dx \wedge dy \wedge dz).$$

Since  $\star(dt \wedge dx \wedge dy \wedge dz) = 1$ , the result follows.

In general, for any two  $k$ -forms  $a, b$ , the  $n$ -form  $a \wedge \star b$  is the inner product  $\langle a, b \rangle$  (defined with the metric) multiplied by the volume form  $\omega = \sqrt{|\det g|} dx^1 \wedge \cdots \wedge dx^n$ .

The exterior derivative has an important property. If we apply it twice, it is identically zero:  $d^2 = 0$ . To show this, take a  $p$ -form written in Einstein notation as  $a_{i_1 \dots i_p} dx^{i_1} \wedge \cdots \wedge dx^{i_p}$ . Then

$$\begin{aligned} d^2(a_{i_1 \dots i_p} dx^{i_1} \wedge \cdots \wedge dx^{i_p}) &= d \left( \frac{\partial a_{i_1 \dots i_p}}{\partial x^i} dx^i \wedge dx^{i_1} \wedge \cdots \wedge dx^{i_p} \right) \\ &= \frac{\partial^2 a_{i_1 \dots i_p}}{\partial x^i \partial x^j} dx^j \wedge dx^i \wedge dx^{i_1} \wedge \cdots \wedge dx^{i_p} \\ &= \frac{1}{2} \left( \frac{\partial^2 a_{i_1 \dots i_p}}{\partial x^i \partial x^j} - \frac{\partial^2 a_{i_1 \dots i_p}}{\partial x^j \partial x^i} \right) dx^j \wedge dx^i \wedge dx^{i_1} \wedge \cdots \wedge dx^{i_p} \\ &= 0. \end{aligned}$$

This proof shows that  $d^2 = 0$  is a consequence of Clairaut's theorem. Specializing again to three dimensions, we can unpack  $d^2 = 0$  into the two statements  $\nabla \times (\nabla f) = 0$  and  $\nabla \cdot (\nabla \times \mathbf{v}) = 0$ .

In addition to these differential results, multivariable calculus is centered around a few integral theorems: the fundamental theorem of calculus, Stokes' theorem, and the divergence theorem. In the language of differential forms, we see that all of these become a single theorem, the generalized Stokes' theorem. The integral of a differential  $n$ -form is defined in  $\mathbb{R}^n$  by

$$\int f(x) dx^1 \wedge \cdots \wedge dx^n = \int f(x) dx^1 \cdots dx^n.$$

On a more general manifold, we use the atlas to map to  $\mathbb{R}^n$  and integrate there. In this case, we pick up a Jacobian factor  $\sqrt{|\det g|}$ . To integrate a function  $f$  over a manifold in a coordinate-invariant way, we have to include this factor:

$$\int_M f = \int_M f(x) \sqrt{|\det g|} dx^1 \wedge \cdots \wedge dx^n \equiv \int_M f \omega,$$

where  $\omega = \sqrt{|\det g|} dx^1 \wedge \cdots \wedge dx^n$  is the volume form defined in Example 2.3. Incidentally, this means that we can integrate the inner product of two  $k$ -forms  $a, b$  simply by integrating  $a \wedge \star b$ , since we noted in Example 2.3 that the volume form is included automatically.

**Theorem 2.1** (Stokes' Theorem). For a differential form  $\omega$  on a manifold  $M$  with boundary  $\partial M$ , we have

$$\int_M d\omega = \int_{\partial M} \omega.$$



We will not prove this theorem, but simply list what it says for  $\omega$  a 0-, 1-, or 2-form in three-dimensional space:

$$\begin{aligned}\int_{\gamma} (\nabla f) \cdot d\mathbf{x} &= f(\gamma_f) - f(\gamma_i), \\ \int_A (\nabla \times \mathbf{v}) \cdot d\mathbf{S} &= \int_{\partial A} \mathbf{v} \cdot d\mathbf{x}, \\ \int_V (\nabla \cdot \mathbf{v}) dV &= \int_{\partial V} \mathbf{v} \cdot d\mathbf{S}.\end{aligned}$$

We thus see that the fundamental theorem of calculus, Stokes' theorem, and the divergence theorem are all aspects of the same result.

---

*Mathematical aside:* We know that exact forms are closed; the reverse is often true as well. For example, in  $\mathbb{R}^3$ , closed forms are exact: if  $\nabla \times \mathbf{v} = 0$  then  $\mathbf{v} = \nabla f$  for some  $f$ , and if  $\nabla \cdot \mathbf{v} = 0$  then  $\mathbf{v} = \nabla \times \mathbf{u}$  for some  $\mathbf{u}$ . However, this statement does not hold for all manifolds.

We measure the failure of closed forms to be exact by the de Rham cohomology. Before defining this, we define the cochain complex of differential forms. A cochain complex is a sequence of algebraic objects (more precisely, modules over rings) connected by maps, such that the composition of any two successive maps is zero. Since  $d^2 = 0$ , the spaces of differential forms clearly form a cochain complex:

$$\dots \longrightarrow 0 \longrightarrow \Omega^0(M) \xrightarrow{d_0} \Omega^1(M) \xrightarrow{d_1} \dots \Omega^{p-1}(M) \xrightarrow{d_{p-1}} \Omega^p(M) \longrightarrow 0 \longrightarrow \dots$$

We say a sequence of this form is exact when  $\text{im } d_i = \ker d_{i+1}$ ; in the language of differential forms, this sequence is exact when all closed forms are exact. The failure of closed forms to be exact is measured by the de Rham cohomology modules

$$H_{\text{dR}}^p(M) = \frac{\ker d_p}{\text{im } d_{p-1}}.$$

de Rham's theorem asserts that the de Rham cohomology modules are isomorphic to the singular cohomology modules. Singular cohomology is defined in terms of chains (roughly, polygonal curves on manifolds). Intuitively, de Rham's theorem says that the failure of closed differential forms to be exact is related to the existence of boundariless curves on a manifold which are not themselves boundaries. For example, a torus has such curves, as shown in Figure 4; this means that there are closed forms on the torus which are not exact.

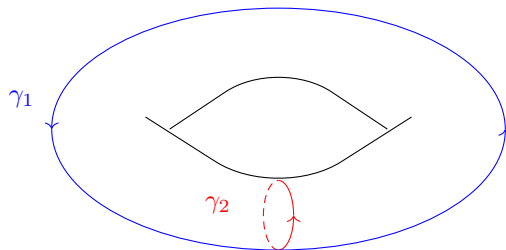


Figure 4: A torus has two types of curves which are boundariless but not themselves boundaries.

---

### 3 Connections on Bundles

We have defined a section of a bundle  $\pi : E \rightarrow B$  as a map  $s : B \rightarrow E$  such that  $\pi \circ s$  is the identity. That is, a section takes points on a manifold and sends them to elements of the fiber of that point. For example, a function  $f : \mathbb{R} \rightarrow \mathbb{R}$  could be thought of as a section of a trivial  $\mathbb{R}$ -bundle over  $\mathbb{R}$ .

However, considering  $f$  as a section, we lose some information about it. We do not currently have the tools to differentiate a section. Even though we know how to evaluate  $\frac{df}{dx}$  when  $f$  is a function of  $\mathbb{R}$ , we cannot do the same when  $f$  is a section of an  $\mathbb{R}$ -bundle. The reason is that a bundle consists of separate fibers at each point; we cannot subtract elements of different fibers, so we cannot take the limit which defines the derivative.

In order to rectify this, we will define a *connection* on a bundle, which gives us a way to link the different fibers together. In this section, we will focus on vector bundles, which are fiber bundles that have vector spaces as fibers. We will develop the idea of a connection, and its curvature, in the context of vector bundles, before moving on to principal bundles in the next section.

#### 3.1 Vector Bundles

A vector bundle is a fiber bundle satisfying two additional properties. First, the fibers of the bundle must be vector spaces, which we will take to be  $\mathbb{R}^n$  (ignoring the case of complex vector bundles with fibers  $\mathbb{C}^n$ ). Second, the local trivialization – the homeomorphism from local pieces of the bundle to a trivial bundle – must be not only a homeomorphism, but a linear isomorphism at each point. This endows the fibers with a linear structure, so we can talk about adding the elements of a fiber and multiplying them by scalars.

Examples of vector bundles include the tangent and cotangent bundles. The tangent bundle for a manifold of dimension  $n$  assigns a vector space  $\mathbb{R}^n$  to each point of the manifold. Indeed, recall that we constructed an element of  $T_pM$  by taking a vector  $v \in \mathbb{R}^n$  and using it to define a directional derivative at  $p$ .

Any section on a vector bundle takes values in the fiber space  $\mathbb{R}^n$ . This means that we can decompose a section in a basis at each point. Of course, it would not be particularly helpful if we picked the basis at each point randomly, leading to wildly discontinuous decompositions. Instead, we use a *frame* on the vector bundle. A frame is a set of sections  $e_i$ ,  $i = 1, \dots, n$ , such that at each point  $p \in B$ , the vectors  $e_i(p)$  form a basis for  $\pi^{-1}(p)$ .

If we have a frame for a vector bundle, then any section can be written in terms of it, via

$$s = s^i e_i,$$

summation notation in effect. We call  $s^i$  the components of the frame  $s$  in the frame  $e$ .

There are various operations on vector spaces which generalize to vector bundles. We have already seen one example of this: we extended the construction of a dual vector space to that of a dual vector bundle, by taking the dual of every fiber in  $TM$  to form the cotangent bundle  $T^*M$ . We can also take two vector bundles and combine them by combining their fibers in a prescribed way. There are two primary ways to combine two vector spaces into a vector space:

- Direct sum: given vector spaces  $V$  and  $W$ , with bases  $\{v_i\}_{i=1}^n$  and  $\{w_j\}_{j=1}^m$  respectively, the vector space  $V \oplus W$  has basis  $\{v_1, \dots, v_n, w_1, \dots, w_m\}$ . We have  $\dim V \oplus W = \dim V + \dim W$ .
- Tensor product: given vector spaces  $V$  and  $W$ , with bases  $\{v_i\}_{i=1}^n$  and  $\{w_j\}_{j=1}^m$  respectively, the vector space  $V \otimes W$  has basis  $\{v_1 \otimes w_1, \dots, v_1 \otimes w_m, \dots, v_n \otimes w_1, \dots, v_n \otimes w_m\}$ . We have  $\dim V \otimes W = \dim V \times \dim W$ .

By using these operations on the fibers of two vector bundles  $E$  and  $F$ , we can form the direct sum (often called the Whitney sum)  $E \oplus F$  and the tensor product  $E \otimes F$ .

**Example 3.1.** Let  $E$  be a vector bundle with fibers  $\mathbb{R}^n$ . Show that the tensor product  $E \otimes E^*$  can be thought of as the bundle of endomorphisms of  $\mathbb{R}^n$ ; that is, the bundle with fibers given by matrices  $\mathbb{R}^{n \times n}$ .

**Solution:** An element of  $\mathbb{R}^n \otimes (\mathbb{R}^n)^*$  is a linear combination of its basis elements,

$$a = a_j^i e_i \otimes e^j,$$

where  $\{e_i\}_{i=1}^n$  is a basis for  $\mathbb{R}^n$ , and  $e^i$  is its dual basis; that is,  $e^i$  is the linear functional which sends  $e_i$  to 1 and all other basis elements to 0. If we act on a vector  $v = v^i e_i$  with  $a$ , we find

$$av = (a_j^i e_i \otimes e^j)(v^k e_k) = a_j^i v^k \delta_k^j e_i = (a_j^i v^j) e_i,$$

which is exactly what we would get by treating  $a$  as a matrix and multiplying by  $v$ .

## 3.2 Connections

For a vector bundle, a section is a vector-valued function on the base space. For example, if we have an  $\mathbb{R}^3$  vector bundle over  $\mathbb{R}^3$  (that is, the base space and the fiber space are both  $\mathbb{R}^3$ ), then sections correspond to vector fields in  $\mathbb{R}^3$ . When we see a vector field, our first instinct is to do calculus with it. However, we are not yet ready for this. To define a derivative of a section  $s$ , we would need to evaluate a limit of the form

$$\lim_{\epsilon \rightarrow 0} \frac{s(p + \epsilon) - s(p)}{\epsilon}.$$

The notation  $p + \epsilon$  is not precise; it indicates a point near  $p$ , with the distance from  $p$  parameterized by  $\epsilon$ . Regardless of this, there is a bigger problem:  $s$  maps points to their fibers, so  $s(p + \epsilon)$  and  $s(p)$  live in different fibers, i.e., different vector spaces. We do not have a way to subtract these two vectors.

The remedy to this will be the *connection* on the vector bundle. The connection allows us to take a derivative by giving us a way of identifying nearby fibers with each other. However, we will follow this logic in the reverse order, first defining a connection as a way of taking a gradient and then understanding the geometric ideas which result from this, primarily parallel transport and curvature.

Our goal is to take a gradient of a section of a vector bundle  $E$ , with base space  $M$ . We denote the space of sections by  $\Gamma(E)$ . The gradient of a section must tell us how each component of  $s$  changes as we move along each tangent direction, so it contains a matrix worth of information. To make this idea explicit, take a frame  $e_i$  for  $E$ , and express a section  $s \in \Gamma(E)$  as  $s^i e_i$ . Then  $\nabla s$  needs to tell us how each component  $s^i$  changes in each direction of the tangent bundle  $TM$ . Put another way,  $\nabla s$  should act as a function from  $TM$  to the fibers of  $E$ , giving the change of  $s$  in that direction of  $TM$ . This function should be linear if  $\nabla$  is a *bona fide* derivative.

In Example 3.1, we saw that tensoring a vector bundle with its dual gives its endomorphism bundle. We can generalize this logic by saying that tensoring a vector bundle  $E$  with the dual of  $F$ ,  $F^*$ , corresponds to taking the bundle of linear maps from fibers of  $F$  to fibers of  $E$ . In the present case, we are seeking to represent an object which gives us a linear map from fibers of  $TM$  to fibers of  $E$ , so we take  $E \otimes T^*M$ . The connection is then a linear map

$$\nabla : \Gamma(E) \rightarrow \Gamma(E \otimes T^*M).$$

We demand one more property before we call  $\nabla$  a connection. Ordinary derivatives obey the Leibniz rule,

$$\partial(fg) = (\partial f)g + f(\partial g).$$

Connections obey a similar rule. If we take a section  $s$  and multiply it by a scalar function  $f$ , then we must have

$$\nabla(fs) = f\nabla s + s \otimes df,$$

where  $df$  is the differential (or the exterior derivative) of  $f$ .

The connection gives us all the information we need to define a derivative along a direction  $X$ , where  $X \in \Gamma(TM)$  is a vector field on the manifold. Indeed, this is in the definition of the connection: an element of  $\Gamma(E \otimes T^*M)$  stands ready to act on an element of  $\Gamma(TM)$  to give an element of  $\Gamma(E)$ . We thus define  $\nabla_X s = (\nabla s)X$ , and call this the covariant derivative along  $X$ .

The connection takes us from sections of  $E$  to sections of  $E \otimes T^*M = E \otimes \Omega^1(M)$ . It is natural to ask whether we can go one step further, and define an object which takes us from sections of  $E \otimes \Omega^k M$  to sections of  $E \otimes \Omega^{k+1}(M)$ . This is called the exterior connection, and in fact there is a unique exterior connection for a given connection. It satisfies a version of the Leibniz rule,

$$\nabla(v \wedge w) = (\nabla v) \wedge w + (-1)^{\deg v} v \wedge (dw),$$

where  $\deg v$  is the degree of the homogeneous form  $v$  (i.e., if  $v$  is a  $p$ -form,  $\deg v = p$ ). Note that this coincides with the requirement we already have when  $v \in \Gamma(E)$  and  $w$  is a 0-form (i.e., a scalar function).

This is all very abstract; to make it more explicit, we can work in terms of coordinates. Any section can be expressed in terms of a frame as  $s = s^i e_i$ ; in this representation, the Leibniz rule gives

$$\nabla s = e_i \otimes ds^i + s^i (\nabla e_i).$$

Thus, if we know how the connection acts on the frame, we can determine how it acts on any section. Moreover, since  $\nabla e_i$  is a section of  $E \otimes T^*M$ , we can decompose it into elements of the frame weighted by one-forms  $\omega$ :

$$\nabla e_i = e_j \otimes \omega^j_i.$$

We then have an explicit formula for the connection of a section:

$$\nabla s = e_i \otimes ds^i + s^i e_j \otimes \omega^j_i.$$

This is often abbreviated by writing  $\nabla = d + \omega$ ; that is, applying  $\nabla$  to a section is the same as applying  $d$  to its components and then adding the contribution from the frame, which is encoded by the matrix of one-forms  $\omega$ .

**Example 3.2.** In differential geometry, we are chiefly concerned with connections on the tangent bundle  $TM$ . Show that the covariant derivative can be written as

$$\nabla_X v = \partial_X v + \Gamma^i_{jk} v^j X^k,$$

where  $\Gamma^i_{jk}$  are components of the connection one-form.

**Solution:** We first contract  $\nabla s$  with  $X$  to obtain the covariant derivative:

$$\nabla_X v = e_i (dv^i X) + v^i e_j (\omega^j_i X).$$

In each term, we have one-forms acting on vectors. In the first case, we can evaluate this using the definition of the differential:  $dv^i(X) = X(v^i)$ . Recall that  $X$  on the right hand side is acting as a directional derivative on the function  $v^i$ , so we could alternatively write this as  $\partial_X v^i$ . For the second term, we can use the same idea, by expressing both  $\omega^j_i$  and  $X$  in a basis:

$$\omega^j_i X = (\omega^j_{ik} e^k)(X^l e_l) = \omega^j_{ik} X^l \delta^k_l = \omega^j_{ik} X^k.$$

In total, we have obtained

$$\begin{aligned} \nabla_X v &= e_i \partial_X v^i + v^i e_j \omega^j_{ik} X^k \\ &\equiv \partial_X v + \Gamma^i_{jk} v^j X^k, \end{aligned}$$

where  $\Gamma^i_{jk} = \omega^i_{jk}$  is a component of the connection one-form.

Example 3.2 shows that the connection one-form closely associated with the affine connection in differential geometry. An important aspect of this object is its failure to transform as a tensor under coordinate changes. This is also true of the connection one-form, which we can see by changing our frame. Let  $e'_i$  be a new frame, related to the old frame by

$$e'_i = \eta^j_i e_j.$$

To determine the connection one-form of the new frame, we take the connection of both sides, obtaining

$$\begin{aligned} \nabla e'_i &= \nabla(\eta^j_i e_j) \\ &= e_j \otimes d\eta^j_i + \eta^j_i \nabla e_j \\ &= (\eta^{-1})^k_j e'_k \otimes d\eta^j_i + \eta^j_i (e_l \otimes \omega^l_j) \\ &= e'_k \otimes \left( (\eta^{-1})^k_j d\eta^j_i + \eta^j_i \omega^l_j (\eta^{-1})^k_l \right). \end{aligned}$$

The quantity appearing in parentheses is the connection for the new frame. Treating  $\eta$  and  $\omega$  as matrices, we can write this as  $\omega' = \eta^{-1}d\eta + \eta^{-1}\omega\eta$ . The second term is what we expect for a change of basis; the first term is anomalous, since it involves  $d\eta$ .

As promised, we can use the connection to recover a notion of parallel transport between fibers. This is, in fact, relatively simple. In order to have a vector undergo parallel transport along some path  $\gamma$  on the manifold, we wish for it not to change along  $\gamma$ . Thus, we require

$$\nabla_\gamma v = 0,$$

where  $\nabla_\gamma$  denotes the contraction of  $\nabla$  with a vector parallel to  $\gamma$ .

### 3.3 Curvature

Since the connection does not transform nicely, it is explicitly dependent on a choice of frame. Thus, it is not an object of direct geometric interest. However, we can use it to form an object which is, called the curvature. The curvature is simply the covariant derivative of the connection:

$$\Omega = \nabla\omega.$$

Since  $\omega$  is a one-form,  $\Omega$  is a two-form. We can write this more explicitly by expanding the covariant derivative in terms of the connection, giving

$$\Omega = d\omega + \omega \wedge \omega.$$

Still more explicitly, we can write this in terms of the matrix components  $\omega_j^i$  as

$$\Omega_j^i = d\omega_j^i + \omega_k^i \wedge \omega_j^k.$$

Our first task is to verify the claim that this transforms tensorially. If we change to a frame  $e'_i$ , then we have

$$\Omega' = \nabla\omega' = d\omega' + \omega' \wedge \omega'.$$

We already have an expression for  $\omega'$ , namely  $\omega' = \eta^{-1}d\eta + \eta^{-1}\omega\eta$ . Substituting this in, we obtain several simplifications using the identities  $d^2 = 0$  and  $d\eta^{-1} = -\eta^{-1}d\eta\eta^{-1}$ :

$$\begin{aligned} \Omega' &= d(\eta^{-1}d\eta + \eta^{-1}\omega\eta) + (\eta^{-1}d\eta + \eta^{-1}\omega\eta) \wedge (\eta^{-1}d\eta + \eta^{-1}\omega\eta) \\ &= -\eta^{-1}d\eta\eta^{-1} \wedge d\eta - \eta^{-1}d\eta\eta^{-1} \wedge \omega\eta + \eta^{-1}d\omega\eta - \eta^{-1}\omega \wedge d\eta \\ &\quad + \eta^{-1}d\eta \wedge \eta^{-1}d\eta + \eta^{-1}(\omega \wedge \omega)\eta + \eta^{-1}\omega \wedge d\eta + \eta^{-1}d\eta \wedge \eta^{-1}\omega\eta \\ &= \eta^{-1}(d\omega + \omega \wedge \omega)\eta. \end{aligned}$$

Thus,  $\Omega$  transforms appropriately under a change of frame.

**Example 3.3.** Recall from Example 3.2 that the components of the connection form for the tangent bundle are identified with the connection components  $\Gamma_{jk}^i$ . Expand the definition of the curvature two-form to recover the Riemann curvature tensor,

$$R_{jkl}^i = \partial_k\Gamma_{jl}^i - \partial_l\Gamma_{jk}^i + \Gamma_{ka}^i\Gamma_{jl}^a - \Gamma_{la}^i\Gamma_{jk}^a.$$

**Solution:** We express the connection form in components by

$$\omega_j^i = \Gamma_{jk}^i de^k.$$

The first term in the curvature is the exterior derivative of these one-forms, which we write as

$$d\omega_j^i = \partial_l \Gamma_{jk}^i de^l \wedge de^k = \frac{1}{2} (\partial_k \Gamma_{jl}^i - \partial_l \Gamma_{jk}^i) de^k \wedge de^l.$$

The second term is

$$\omega_a^i \wedge \omega_j^a = \Gamma_{ak}^i e^k \wedge \Gamma_{jl}^a e^l = \frac{1}{2} (\Gamma_{ak}^i \Gamma_{jl}^a - \Gamma_{al}^i \Gamma_{jk}^a) e^k \wedge e^l.$$

Putting these together, we find that the components of the curvature two-form are the components of the Riemann tensor, up to a factor of two.

The curvature satisfies a relation called the Bianchi identity, given by

$$\nabla \Omega = 0.$$

This is simple to prove; we simply substitute the definitions of  $\Omega$ , and find

$$d(d\omega + \omega \wedge \omega) = d\omega \wedge \omega - \omega \wedge d\omega = -(\omega \wedge \Omega - \Omega \wedge \omega),$$

and it follows that  $\nabla \Omega = 0$ .

### 3.4 Line Bundles and Electrodynamics

In the next section, we will develop the theory of principal bundles, which have fibers given by Lie groups. This is the formalism required to treat a generic gauge theory. However, electrodynamics is simple enough that we can treat it using vector bundles. Formally, the gauge group of electrodynamics is  $U(1)$ , which has the real line  $\mathbb{R}$  as its universal cover; and so we can replace a  $U(1)$  principal bundle with an  $\mathbb{R}$ -bundle, which is a simple case of a vector bundle.

When a vector bundle has one-dimensional fibers, we call it a line bundle. Complex line bundles are rich and interesting, because a complex line is really the complex plane, and we can define holomorphic structures; but a real line bundle is somewhat trivial. Indeed, the matrices we have been dealing with have only one component over a line bundle, and so they all commute. This simplification results from electrodynamics being an *abelian* gauge theory.

These considerations aside, we can draw a correspondence between the connection and curvature of a line bundle and the potential and field strength in electrodynamics. Since matrices become  $1 \times 1$  on a line bundle,  $\omega$  and  $\Omega$  only carry the indices they have as forms. Thus, we identify the 1-form  $\omega$  with the gauge potential  $A^\mu$ . More precisely, we have

$$\omega = -i \frac{e}{\hbar c} A.$$

The factor of  $\frac{e}{\hbar c}$  is a matter of dimensional analysis; the factor of  $i$  represents a difference between mathematics and physics conventions for Lie algebras. It is introduced so that the gauge potential can be real.

Given this, it follows that

$$\Omega = d\omega + \omega \wedge \omega = -i \frac{e}{\hbar c} dA.$$

Since  $A$  is a 1-form, we can write

$$dA = d(A_\mu dx^\mu) = \partial_\nu A_\mu dx^\nu dx^\mu = \frac{1}{2}(\partial_\mu A_\nu - \partial_\nu A_\mu) dx^\mu dx^\nu = \frac{1}{2} F_{\mu\nu} dx^\mu dx^\nu.$$

Thus, the field strength is proportional to the curvature.

We can directly obtain two physical results from corresponding results on bundles. The first is the phase dependence of the wavefunction as it moves through a Maxwell field. Recall that parallel transport of  $\psi$  requires

$$\nabla_\gamma \psi = 0.$$

We can write this as

$$\frac{d\psi}{dt} + \left( \omega \frac{d\gamma}{dt} \right) \psi = \frac{d\psi}{dt} - i \frac{e}{\hbar c} A_\mu \frac{d\gamma^\mu}{dt} \psi = 0.$$

The solution to this is

$$\psi(t) \propto \exp \left( \frac{ie}{\hbar c} \int_\gamma A_\mu dx^\mu \right),$$

exactly as we obtained before using the classical notion of gauge invariance.

The second physical insight is two of the Maxwell equations. These emerge immediately as a consequence of the Bianchi identity, which over a line bundle reads  $d\Omega = 0$ . Over a contractible space, Poincaré's lemma says that any closed form is exact; so if we have  $d\Omega = 0$ , it must be the case that  $\Omega = dA$  for some potential  $A$ . We have already seen that writing the electric and magnetic fields in terms of their potentials implies Gauss's law for magnetism,  $\nabla \cdot \mathbf{B} = 0$ , and Faraday's law,  $\nabla \times \mathbf{E} = -\frac{d\mathbf{B}}{dt}$ .

## 4 Principal Bundles

We should take a moment to recall our goal in developing this mathematics. A gauge theory is characterized by a local and internal symmetry, and we wish to represent such a symmetry formally. So far, we have seen how to construct a bundle over a manifold using vector spaces as fibers, and how to form a connection on such a bundle. In this section, we will see how to replace the vector spaces with groups, the mathematical objects describing symmetries. A section of such a bundle is given by choosing an element of the symmetry group at each point, in a continuous fashion, which is exactly what we mean by choosing a gauge.

In order to make this construction formal, we will need basic elements of the theory of Lie groups, and also their associated Lie algebras. After this legwork, we will be in a position to define principal bundles, and show how to define a connection on them. At this point, finally, we will be able to extract physics from these formalities.



## 4.1 Lie Groups

A Lie group is a mathematical object which is simultaneously a group and a differentiable manifold, such that the group operations interact nicely with the topology of the manifold. More precisely, a Lie group  $G$  is a manifold, together with an invertible operation  $G \times G \rightarrow G$ , which is continuous with respect to the product topology of  $G \times G$  and has a continuous inverse.

The simplest example of a Lie group is the circle group  $U(1)$ . The notation refers to the group of all unitary  $1 \times 1$  matrices, but these are just the unimodular complex numbers, which form a circle. Clearly the product and inverse operations are continuous, so we have a Lie group.

Instead of thinking of the group structure as a function  $G \times G \rightarrow G$ , we can think about a map  $G \rightarrow \text{End}(G)$ , where  $\text{End}(G)$  denotes the set of endomorphisms of  $G$ . Since  $G$  is a differentiable manifold, its endomorphisms are called *diffeomorphisms*. (This reframing from  $G \times G \rightarrow G$  to  $G \rightarrow \text{End}(G)$  is an example, in spirit at least, of the tensor-hom adjunction in category theory, or currying in computer science). We denote the image of  $g \in G$  under this map by  $L_g$ , and call it the left-translation by  $g$ . (We could dually define a right-translation operator  $R_g$ , but there is no need for both, so we will work only with  $L_g$ ).

Recall that, whenever we have a homomorphism  $G \rightarrow \text{End}(A)$  for some object  $A$ , we say  $G$  acts on  $A$ . For example, an action of  $G$  on a vector space is a representation of  $G$ . A trivial  $G$ -action is a homomorphism which sends every element to the identity of  $\text{End}(A)$ . More complicated representations have elements of  $g$  affecting the structure of  $A$  in some way. We say an action is *free* if  $ga = a$  for any  $a \in A$  implies  $g = e$ , the identity of  $G$ . We say an action is *transitive* if, for any  $a_1, a_2 \in A$ , there exists  $g \in G$  such that  $ga_1 = a_2$ . If an action is both free and transitive, then  $G$  is (non-naturally) isomorphic to  $A$  as sets. To see this, fix some element  $a \in A$ ; then for every  $a' \in A$ , there exists  $g$  such that  $ga = a'$ , by transitivity. If there were another element  $\tilde{g}$  with this property, then we would have  $\tilde{g}^{-1}ga = a$ , and so  $\tilde{g}^{-1}g = e$  by freeness, so  $g = \tilde{g}$ . Thus, each  $a' \in A$  defines a unique element of  $G$ , and clearly each  $g \in G$  defines a unique element  $ga$  of  $A$ .

Clearly, the action  $G \rightarrow \text{End}(G)$  of a Lie group on itself is free and transitive. The conclusion that  $G$  is isomorphic to itself as a set is unsurprising; more interesting is that, if we *only* consider the set structure of the acted-upon copy of  $G$ , this isomorphism is non-natural. We can see this from the construction: we could identify  $e \in G$  with any setwise element of  $G$ . In effect, we have  $G$  acting on a set isomorphic to itself, but without a well-defined identity element; every point is equally well-suited to serve as the identity.

We can formalize these notions with some definitions. A *homogeneous space* for a Lie group  $G$  is a smooth manifold  $X$  on which  $G$  acts transitively. For example, consider the Lie group  $SO(3)$ , the  $3 \times 3$  special orthogonal matrices. As linear transformations, these are the rotations of Euclidean three-dimensional space. The sphere  $S^2$  is a homogeneous space for  $SO(3)$ , since for any two points on the sphere, there exists a rotation which sends one to the other. However, the action is not free, since every rotation has two fixed points along its axis. If we additionally require the  $G$ -action on  $X$  to be free, then  $X$  is said to be a *principal homogeneous space* for  $G$ , or more succinctly, a  $G$ -torsor. We think of a  $G$ -torsor for a Lie group  $G$  as the smooth manifold underlying  $G$ , where any point could equally well be the identity.

The most common examples of Lie groups are matrix groups. The groups  $GL(n, \mathbb{R})$  and  $GL(n, \mathbb{C})$

are the *general linear groups* of dimension  $n$  over  $\mathbb{R}$  and  $\mathbb{C}$ , consisting of all invertible  $n \times n$  matrices under multiplication. A matrix group is a subgroup of one of these groups. Any condition on matrices which is preserved under multiplication can be used to define a matrix group. For example,

$$SL(n, \mathbb{k}) = \{M \in GL(n, \mathbb{k}) \mid \det M = 1\}$$

is the special linear group. It forms a subgroup since  $\det M_1 M_2 = \det M_1 \cdot \det M_2$ . We also have

$$\begin{aligned} SO(n) &= \{M \in SL(n, \mathbb{R}) \mid MM^T = I\}, \\ SU(n) &= \{M \in SL(n, \mathbb{C}) \mid MM^\dagger = I\}. \end{aligned}$$

These are all of the most common matrix groups. There are also the so-called *classical groups*, defined as matrices  $M$  for which  $MAM^T = A$  for some fixed matrix  $A$ . For example, if we pick  $A = \text{diag}(-1, 1, 1, 1)$ , we obtain  $SO(1, 3)$ , the Lorentz group.

For the sake of having a concrete example in mind, we will explore  $SO(3)$  in some detail (and, in course,  $SU(2)$ ). This is the most common Lie group appearing in basic physics, since it describes the symmetry of Euclidean 3-space.

We first need to find its dimension as a manifold. Let  $S(n, \mathbb{R})$  denote symmetric matrices of size  $n \times n$ . These do not form a subgroup of  $GL(n, \mathbb{R})$ , but nonetheless, they form a submanifold of dimension  $\frac{n(n+1)}{2}$ , as can be easily verified by counting the number of independent components of a symmetric matrix. Now consider a map which sends  $M \in GL(n, \mathbb{R})$  to  $MM^T \in S(n, \mathbb{R})$ . The group  $SO(n)$  is the preimage of the single point  $I$  under this map, which means

$$\dim SO(n) = \dim GL(n, \mathbb{R}) - \dim S(n, \mathbb{R}) = \frac{n(n-1)}{2}.$$

Specializing to  $SO(3)$ , we see we are working with a three-dimensional manifold.

The three dimensions of  $SO(3)$  can be thought of in various ways (which correspond to various atlases on the manifold). These ways mostly make reference to the fact that an element of  $SO(3)$  corresponds to a rotation of three-dimensional Euclidean space. One approach is the three Euler angles which define a rotation, which should be familiar from classical mechanics. A similar approach, which we shall use here, is to think of a rotation in terms of an axis and an angle. For any normal vector  $\hat{\mathbf{n}}$  and angle  $\theta$ , we have an element  $R(\hat{\mathbf{n}}, \theta) \in SO(3)$ .

Briefly (before encountering a problem), we will consider a map  $R(\hat{\mathbf{n}}, \theta) \mapsto \frac{\theta}{2\pi} \hat{\mathbf{n}}$ . This represents an element of  $SO(3)$  as a point of the unit ball in three dimensions. This looks like a manifold with boundary  $S^2$ , until we realize that  $R(\hat{\mathbf{n}}, 2\pi)$  is equal to the identity element. When we identify every point on the boundary of the unit ball  $B^3$ , we obtain the three-sphere  $S^3$ . If this is confusing, think about the two-dimensional case: if we take the ball  $B^2$ , and fold it up so that every point on the boundary comes together, we obtain the sphere  $S^2$ . This is shown in Figure 5.

The problem we have is that this is not the only identification we have to make. Clearly,  $R(\hat{\mathbf{n}}, 0)$  is also the identity element, so the north and south poles of our  $S^3$  are identical. Moreover, we have  $R(\hat{\mathbf{n}}, \theta) = R(-\hat{\mathbf{n}}, -\theta)$ . So in fact, any two antipodal points of our  $S^3$  are equivalent. The resulting manifold, essentially  $S^3/\mathbb{Z}^2$  with a  $\mathbb{Z}^2$  action defined by inversion  $x \rightarrow -x$ , is called *real projective space*, and denoted  $RP^3$ . We have shown here that the manifold of  $SO(3)$  is  $RP^3$ .

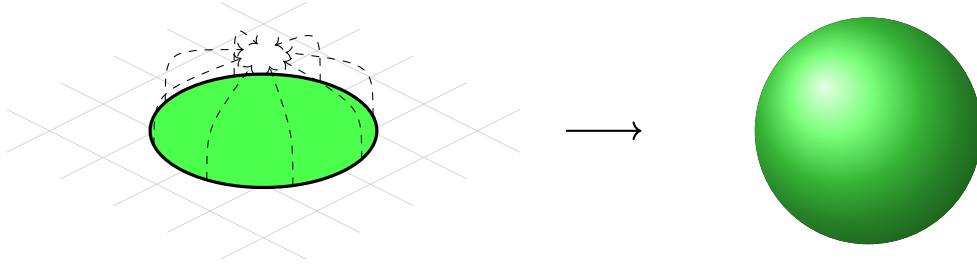


Figure 5: By identifying every point on the boundary of  $B^n$ , we obtain  $S^n$ .

In topology, we are often interested in whether a connected space is *simply connected*. A simply connected space is one for which any path from a point to itself can be continuously deformed to a point. For example,  $S^2$  is simply connected, because any closed path on the sphere can be smoothly retracted to a point. However, the punctured plane  $\mathbb{R}^2 \setminus \{0\}$  is not simply connected, because a circle wrapping around the origin cannot be deformed to a point. The manifold  $RP^n$  is not simply connected, which we can see by taking a path on  $S^n$  from a point to its antipode. This projects to a closed path in  $RP^n$ , but clearly it cannot be deformed to a point, since its endpoints are fixed and are distinct in  $S^n$ . The closed paths in a space can be organized into a group called the *fundamental group* of the space; for  $RP^n$ , the fundamental group is  $\mathbb{Z}_2$  (and incidentally, integer and half-integer spin particles are classified by the representations of this group).

When a space is not simply connected, we can find a *universal cover* for it which is simply connected. A covering space for a space  $X$  is a surjective map  $\pi : Y \rightarrow X$  such that, for any  $x \in X$ , there exists a neighborhood  $U$  of  $x$  for which  $\pi^{-1}(U)$  consists of a union of connected spaces, each of which is homeomorphic to  $U$ . A simple example of a covering space is the plane  $\mathbb{R}^2$  as a cover for the torus  $S^1 \times S^1$ , via the map which projects each copy of  $\mathbb{R}$  onto  $\mathbb{R}/\mathbb{Z} \cong S^1$ . That is, if we have a point  $(x, y) \in \mathbb{R}^2$ , the fractional parts of  $x$  and  $y$  specify the two angles on the torus. If we draw a small neighborhood around any point on the torus, its inverse image is an infinite number of copies of a small neighborhood in  $\mathbb{R}^2$ , arranged on a lattice. A *universal cover* is a covering space which is simply connected.

We have already defined  $RP^n$  via a surjective map from  $S^n$ . It is simple to verify that  $S^n$  is in fact a cover for  $RP^n$ , and since it is simply connected, it is the universal cover. Associated to the idea of a cover in topology is the idea of a covering group for a topological group (in particular, a Lie group). To define a group structure on a covering space  $Y$  for a Lie group  $X$ , we pick an identity  $e^* \in \pi^{-1}(e)$ . For any two elements  $a, b \in Y$ , let  $\gamma_a, \gamma_b : [0, 1] \rightarrow Y$  be paths starting at  $e$  and ending at  $a$  and  $b$  respectively. Then let  $\phi : [0, 1] \rightarrow X$  be given by  $\phi(t) = \pi(\gamma_a(t))\pi(\gamma_b(t))$  (i.e., we project down to  $X$ , and then use the group structure on  $X$ ). By the definition of a covering space  $Y$ , the path  $\phi$  in  $X$  lifts to several paths in  $Y$ , each starting at a different element of  $\pi^{-1}(e)$ ; we pick the one starting at  $e^*$ , and call its terminal point the product  $ab$ .

Using this construction on the group  $SO(n)$ , by lifting the manifold to its double cover  $S^n$ , we obtain groups called  $\text{Spin}(n)$ . For general  $n$ , these are distinct from any of the groups we have mentioned thus far. However, in low dimensions, there can be accidental isomorphisms, and indeed this happens for  $\text{Spin}(3)$ . It turns out that  $\text{Spin}(3) \cong SU(2)$ . To see this, note that unitarity

requires an element of  $SU(2)$  to have the form

$$\begin{pmatrix} \alpha & \beta \\ -\beta^* & \alpha^* \end{pmatrix},$$

and to have determinant one, we must have  $|\alpha|^2 + |\beta|^2 = 1$ . Writing  $\alpha = a + ib$  and  $\beta = c + id$ , this means  $a^2 + b^2 + c^2 + d^2 = 1$ , so we have a point of  $S^3$ . It is not obvious from this alone that  $SU(2) \cong \text{Spin}(3)$  as groups, but in fact this is the case.

## 4.2 Lie Algebras

Since a Lie group is a manifold, we can do everything with it that we could do with manifolds. In this subsection, we will be concerned with the tangent spaces of Lie groups. We will see that the group structure gives a natural isomorphism between all the tangent spaces on  $G$ , so it suffices to consider only the tangent space at the identity; and moreover, that the group structure endows this tangent space with additional structure, making it into an algebra. Before understanding this relationship, we will look at Lie algebras in abstraction.

A Lie algebra  $\mathfrak{g}$  is a vector space, together with a product  $\mathfrak{g} \times \mathfrak{g} \rightarrow \mathfrak{g}$ , denoted by  $[\cdot, \cdot]$ , which satisfies the following properties:

- Antisymmetry:  $[v, w] = -[w, v]$
- Linearity:  $[au + bv, w] = a[u, w] + b[v, w]$
- Jacobi identity:  $[[u, v], w] + [[w, u], v] + [[v, w], u] = 0$

An ideal of a Lie algebra is a linear subspace  $\mathfrak{a} \subset \mathfrak{g}$  such that  $[\mathfrak{g}, \mathfrak{a}] \subset \mathfrak{a}$ , where

$$[\mathfrak{g}, \mathfrak{a}] = \{[u, v] \mid u \in \mathfrak{g}, v \in \mathfrak{a}\}.$$

Every Lie algebra has at least two ideals, namely  $\{0\}$  and itself. Another important ideal (which may coincide with  $\{0\}$  or  $\mathfrak{g}$  in some cases) is the center of  $\mathfrak{g}$ , defined as the maximal subspace  $\mathfrak{a}$  for which  $[\mathfrak{g}, \mathfrak{a}] = \{0\}$ .

If a Lie algebra  $\mathfrak{g}$  has only the two required ideals,  $\{0\}$  and itself, we say  $\mathfrak{g}$  is a simple Lie algebra. We can combine two Lie algebras by taking their direct sum  $\mathfrak{g} \oplus \mathfrak{h}$  as vector spaces, and defining the product by

$$[g_1 + h_1, g_2 + h_2] = [g_1, g_2] + [h_1, h_2].$$

If a Lie algebra is a direct sum of simple Lie algebras, we say it is semisimple.

A homomorphism of Lie algebras  $\phi : \mathfrak{g} \rightarrow \mathfrak{h}$ , is, like any homomorphism, a map which preserves the algebraic structure of its domain. In this case, that means  $\phi$  must be a linear transformation of vector spaces, and also obey the rule

$$\phi([x, y]) = [\phi(x), \phi(y)],$$

where the bracket on the left belongs to  $\mathfrak{g}$  while the bracket on the right belongs to  $\mathfrak{h}$ .

A representation of a Lie algebra is a homomorphism  $\mathfrak{g} \rightarrow \mathfrak{gl}(V)$ , where  $V$  is a vector space and  $\mathfrak{gl}(V)$  is the Lie algebra formed by taking the space of endomorphisms of that vector space, with the commutator as a product. Put another way, a representation is a map sending elements of the algebra to matrices, in such a way that the matrix commutator agrees with the bracket on the original algebra.

Every Lie algebra has a canonical representation called the adjoint representation, which is defined over the algebra itself (though only considering its vector space structure). The adjoint map sends  $x \in \mathfrak{g}$  to  $\text{ad}_x$ , where  $\text{ad}_x : \mathfrak{g} \rightarrow \mathfrak{g}$  is defined by

$$\text{ad}_x y = [x, y].$$

It is clear that  $\text{ad}$  is a linear map, since

$$\text{ad}_{au+bv} w = [au + bv, w] = a[u, w] + b[v, w] = (a \text{ad}_u + b \text{ad}_v)w.$$

Additionally, it respects the bracket, since (using the Jacobi identity)

$$\text{ad}_{[x,y]} z = [[x, y], z] = [x, [y, z]] - [y, [x, z]] = (\text{ad}_x \text{ad}_y - \text{ad}_y \text{ad}_x)z.$$

Therefore, every Lie algebra has a representation with dimension equal to its own dimension, which is an important fact.

We can use the adjoint representation to define a symmetric bilinear form on a Lie algebra, called the Killing form. The Killing form is given by

$$K(x, y) = \text{tr}(\text{ad}_x \text{ad}_y).$$

To understand this definition, remember that  $\text{ad}_x$  and  $\text{ad}_y$  are both members of  $\mathfrak{gl}(\mathfrak{g})$ , the space of endomorphisms of the vector space of  $\mathfrak{g}$  – so in effect, they are matrices. Thus, the trace of their product is well-defined, and is also symmetric in  $x$  and  $y$ .

**Example 4.1.** The algebra  $\mathfrak{sl}(2, \mathbb{C})$  consists of all complex  $2 \times 2$  matrices with trace zero, with the bracket given by the matrix commutator. Using the basis

$$h = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad e = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad f = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

Work out the Lie bracket, the adjoint representation, and the Killing form using this basis.

**Solution:** To determine the values of the Lie bracket, we simply take the matrix commutators. We have

$$\begin{aligned} [h, e] &= he - eh = 2e, \\ [h, f] &= hf - fh = -2f, \\ [e, f] &= ef - fe = h. \end{aligned}$$

Therefore, using the ordered basis  $\{h, e, f\}$ , the adjoint representation is given by

$$\text{ad}_h = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -2 \end{pmatrix} \quad \text{ad}_e = \begin{pmatrix} 0 & 0 & 1 \\ -2 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{ad}_f = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & 0 \\ 2 & 0 & 0 \end{pmatrix}$$

By multiplying these in pairs, we find that the Killing form is

$$\begin{array}{lll} K(h, h) = 8 & K(h, e) = 0 & K(h, f) = 0 \\ K(e, h) = 0 & K(e, e) = 0 & K(e, f) = 4 \\ K(f, h) = 0 & K(f, e) = 4 & K(f, f) = 0 \end{array}$$

More concisely, we can write  $K(x, y) = x^T K y$ , where in the  $\{h, e, f\}$  basis we have

$$K = \begin{pmatrix} 8 & 0 & 0 \\ 0 & 0 & 4 \\ 0 & 4 & 0 \end{pmatrix}.$$

An important fact, which we will not prove, is that a Lie algebra is semisimple if and only if its Killing form is nondegenerate (that is, if  $K(x, x) = 0$  implies  $x = 0$ ). Thus, the previous example shows that  $\mathfrak{sl}(2, \mathbb{C})$  is semisimple. In fact,  $\mathfrak{sl}(2, \mathbb{C})$  is simple.

It is instructive to consider the finite-dimensional representations of  $\mathfrak{sl}(2, \mathbb{C})$ . Consider a representation  $\mathfrak{sl}(2, \mathbb{C}) \rightarrow \mathfrak{gl}(V)$ . Formally we should define a Lie algebra homomorphism  $\phi : \mathfrak{sl}(2, \mathbb{C}) \rightarrow \mathfrak{gl}(V)$ , and denote the action of  $x \in \mathfrak{sl}(2, \mathbb{C})$  on  $v \in V$  by  $\phi(x)v$ . We will abbreviate this by simply writing  $xv$ ; it will be clear from context how this is to be interpreted.

A theorem which we will not prove says that representations of a Lie algebra are completely reducible, meaning in particular that  $\phi(h)$  is a diagonalizable matrix on  $V$ . Thus, we can split  $V$  into eigenspaces of  $h$ :

$$V = V_{\lambda_1} \oplus \cdots \oplus V_{\lambda_n},$$

where  $v \in V_\lambda$  means  $hv = \lambda v$ . Now, since  $[h, e] = 2e$ , we have

$$v \in V_\lambda \implies hev = ([h, e] + eh)v = (\lambda + 2)ev,$$

so  $ev \in V_{\lambda+2}$ . Similarly,  $fv \in V_{\lambda-2}$ . Since  $V$  is finite-dimensional,  $\phi(h)$  can only have a finite number of eigenvalues, this chain of eigenspaces must stop somewhere; that is, we can find  $\lambda$  such that  $eV_\lambda = 0$ . Now, let  $v_0 \in V_\lambda$ , and let  $v_j = f^j v_0$ . Then  $v_j \in V_{\lambda-2j}$ ; the action of  $f$  is moving us down a chain. If we want to move back up, we should act with  $e$ . To determine this, we need  $[e, f^n]$ ; it can be shown by induction that

$$[e, f^n] = n f^{n-1} (h + 1 - n).$$

Using this, we find

$$ev_j = e f^j v_0 = [e, f^j] v_0 = j(\lambda + 1 - j)v_{j-1}.$$

Again, since  $V$  is finite dimensional, there must be some level  $N$  at which  $v_{N+1} = 0$ . Let  $N$  be the lowest possible value, so that  $v_N \neq 0$ . Then

$$0 = ev_{N+1} = (N + 1)(\lambda - N)v_N.$$

Therefore,  $N = \lambda$ . This means  $\lambda$  is an integer, and that the eigenvalue of  $v_N$  is  $\lambda - 2N = -N$ , so there is a symmetry to the tower we have found, shown in Figure 6.

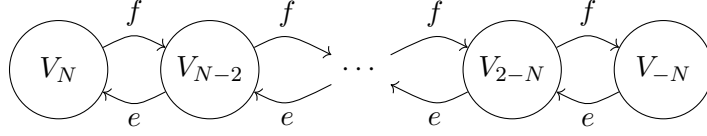


Figure 6: An irreducible representation of  $\mathfrak{sl}(2, \mathbb{C})$ .

Consider the subspace of  $V$  spanned by  $\{v_0, \dots, v_N\}$ . By construction, if we act with  $h$ ,  $e$ , or  $f$ , we remain in this subspace. We call such a subspace an *invariant* subspace. We have just shown that for any representation  $\mathfrak{sl}(2, \mathbb{C}) \rightarrow \mathfrak{gl}(V)$ , the invariant subspaces look like Figure 6.

If a representation has invariant subspaces, it is said to be reducible. If we have a reducible representation, we can decompose the vector space into a direct sum of its invariant subspaces. Thus, the interesting representations are the ones which form the smallest pieces of this decomposition; they are the irreducible representations. An irreducible representation has no invariant subspaces other than  $\{0\}$  and the entire vector space. For  $\mathfrak{sl}(2, \mathbb{C})$ , we have just shown that the irreducible representations are labeled by an integer  $N$ , and have dimension  $2N + 1$ .

We can generalize much of this analysis to arbitrary simple Lie algebras  $\mathfrak{g}$ . Our classification of the  $\mathfrak{sl}(2, \mathbb{C})$  representations was driven by the eigendecomposition of  $h$ ; this is because  $\{h\}$  forms a *Cartan subalgebra* of  $\mathfrak{sl}(2, \mathbb{C})$ . A Cartan subalgebra  $\mathfrak{h} \subset \mathfrak{g}$  is a subalgebra satisfying  $[\mathfrak{h}, \mathfrak{h}] = 0$  (i.e., an abelian subalgebra), such that for any  $H \in \mathfrak{h}$ ,  $\text{ad}_H$  is diagonalizable. A theorem we will not prove states that, for simple Lie algebras, Cartan subalgebras exist and they all have the same dimension. The dimension of the Cartan subalgebra is called the *rank* of  $\mathfrak{g}$ .

Thus, given a simple Lie algebra  $\mathfrak{g}$  with rank  $r$ , let  $\mathfrak{h} \subset \mathfrak{g}$  be a Cartan subalgebra, and let  $\{H_1, \dots, H_r\}$  be a basis for it. By hypothesis, all of  $\text{ad}_{H_i}$  are diagonalizable; moreover, since  $[H_i, H_j] = 0$ ,  $\text{ad}_{H_i}$  commutes with  $\text{ad}_{H_j}$  for any representation  $\phi$ . A simple exercise in linear algebra shows that if two matrices are diagonalizable and they commute, one can find a basis in which they are simultaneously diagonalizable. Thus, we can decompose  $\mathfrak{g}$  into a direct sum of spaces  $\mathfrak{g}_\alpha$ , where  $\alpha \in \mathbb{C}^r$  are vectors such that

$$v \in \mathfrak{g}_\alpha \implies \text{ad}_{H_i} v = \alpha_i v, \quad i = 1, \dots, r.$$

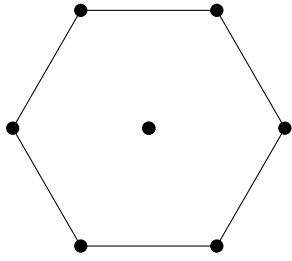
The vectors  $\alpha$  occurring in this decomposition are called the roots of  $\mathfrak{g}$ , and  $\mathfrak{g}_\alpha$  is called the root spaces.

We can verify some simple properties of roots and root spaces. First, note that  $0$  is always a root, and its root space  $\mathfrak{g}_0$  is simply the Cartan subalgebra  $\mathfrak{h}$ . Additionally, if we take  $v \in \mathfrak{g}_\alpha$  and  $w \in \mathfrak{g}_\beta$ , then

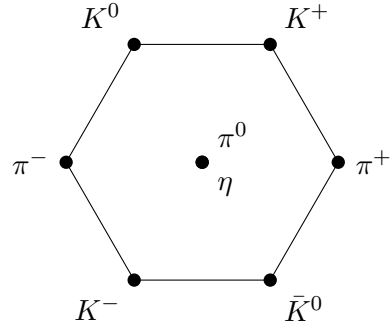
$$\text{ad}_{H_i}[v, w] = [H_i, [v, w]] = [[H_i, v], w] + [v, [H_i, w]] = (\alpha_i + \beta_i)[v, w],$$

so  $[\mathfrak{g}_\alpha, \mathfrak{g}_\beta] \subset \mathfrak{g}_{\alpha+\beta}$ .

As an example, very much non-randomly chosen, we will consider the root system of the simple Lie algebra  $\mathfrak{su}(3)$ . This is an eight-dimensional Lie algebra with rank 2, so its roots can be drawn in the plane. The root  $0$  corresponds to the Cartan subalgebra with two generators; there are six more dimensions of the Lie algebra which must be broken down into root spaces. It turns out that all of the remaining root spaces are one-dimensional, and their roots form the vertices of a hexagon



(a) The root system of the Lie algebra  $\mathfrak{su}(3)$ .



(b) The “eightfold way,” a depiction of meson bound states in quantum chromodynamics.

Figure 7

(with the exact geometry depending on a choice of basis for the Cartan subalgebra). This is shown in Figure 7a.

Some foreshadowing is in order. The Lie group  $SU(3)$ , which is associated with  $\mathfrak{su}(3)$  by a construction which will follow shortly, describes a symmetry of quark flavors in quantum chromodynamics. The strong nuclear force acts in the same way on all quark flavors, so as long as two quarks have small mass compared to the QCD energy scale (which is true of the quarks  $u$ ,  $d$ , and  $s$ ), they can be substituted for one another without significant alterations to the physics. This gives rise to an  $SU(3)$  symmetry of rotations in the Hilbert space spanned by these three flavors. A brief glance at Figure 7b, showing mesons laid out according to their charge and strangeness, should convince you that this  $SU(3)$  symmetry is very much active in determining the spectrum of QCD bound states. For noticing this structure, and for predicting a missing particle in a related baryon structure, Murray Gell-Mann won the Nobel Prize in 1969.

Finally, we will describe how a Lie group gives rise to a Lie algebra. Recall that for a Lie group  $g$ , we have diffeomorphisms  $L_g : G \rightarrow G$  associated to every point  $g \in G$ . This means in particular that for any two points  $g, g' \in G$ , there is a canonical diffeomorphism  $L_{g'g^{-1}}$  which sends  $g$  to  $g'$ .

We can use this diffeomorphism to relate the tangent spaces  $T_g G$  and  $T_{g'} G$ . Recall that the tangent space  $T_p M$  was abstractly defined as the set of all maps

$$C^1(M) \ni \phi \mapsto \mathbf{v} \cdot \nabla(\phi \circ f^{-1})|_p \in \mathbb{R},$$

where  $f$  is a coordinate chart on  $M$  mapping a neighborhood  $U \ni p$  to an open subset of  $\mathbb{R}^n$ . Intuitively, these are the directional derivatives at  $p$ . If we have a diffeomorphism  $\phi : M \rightarrow N$ , with  $\phi(p) = q$ , then we can define a pushforward map  $\phi_* : T_p M \rightarrow T_q N$  by

$$\phi_*(v) = [C^1(N) \ni \psi \mapsto v(\psi \circ \phi)].$$

That is, to evaluate the vector  $\phi_*(v)$  on a function  $\psi$  defined on  $N$ , we first compose  $\psi$  with  $\phi$  so that we have a function on  $M$ , and then evaluate  $v$  on that function.

The important point here is not the exact construction of the pushforward, but what it means: for a Lie group, since we have the family of diffeomorphisms  $L_g$ , we have a linear map between any two



tangent spaces; and moreover, the maps between  $T_g G$  and  $T_{g'} G$  are inverses of each other. This means that all the tangent spaces of a Lie group are naturally isomorphic, so we are free to focus on only one of them. For simplicity, we focus on  $T_e G$ , the tangent space at the identity.

We know that  $T_e G$  has the structure of a vector space; to endow it with the structure of a Lie algebra, all we need to do is define the bracket  $[\cdot, \cdot]$ . Since the elements of  $T_e G$  are tangent directions, there is a simple way to define  $[v, w]$ : take the identity, move it in the direction  $w$ , then the direction  $v$ . Alternatively, move it in the direction  $v$ , then the direction  $w$ . The bracket  $[v, w]$  gives the difference in the results.

To make this idea formal, we first introduce the concept of a left-invariant vector field. A left-invariant vector field on a Lie group  $G$  is a vector field  $X$  for which

$$(L_g)_* X(g') = X(gg').$$

Clearly any left-invariant field defines an element  $X(e) \in T_e G$ ; and likewise, if we fix  $X(e) \in T_e G$ , then  $X(g)$  is fixed by the definition for all  $g$ . Thus,  $T_e G$  is isomorphic to the set of left-invariant vector fields.

Now we can define the Lie bracket. For two vectors  $v, w \in T_e G$ , we build the corresponding left-invariant vector fields  $V, W$ . Given a smooth function  $f$  on the group, a vector field can act on  $f$  to produce another smooth function  $V(f)$ , via  $(V(f))(p) = (V(p))(f)$  – that is, the value of  $V(f)$  at a point  $p$  is given by acting on  $f$  with the vector  $V(p)$ . We tentatively define

$$[v, w] = (VW - WV)(e).$$

In order for this definition to make sense, we need to verify several facts. First, in order to associate a vector in  $T_e G$  with the vector field  $VW - WV$ , we need to check that  $VW - WV$  is left-invariant. Since  $V$  and  $W$  are both left-invariant, this is immediate. Additionally, antisymmetry and linearity are immediate. To check the Jacobi identity, we simply compute:

$$\begin{aligned} [u, [v, w]] + [w, [u, v]] + [v, [w, u]] &= U(VW - WV) - (VW - WV)U + W(UV - VU) \\ &\quad - (UV - VU)W + V(WU - UW) - (WU - UW)V \\ &= 0. \end{aligned}$$

Thus, our bracket satisfies the conditions for making  $T_e G$  into a Lie algebra.

Now that we have seen how to go from a Lie group to a Lie algebra, we might wonder how to go from an algebra to a group. There is a relatively clear answer: starting from a vector  $v \in T_e G$ , we build the left-invariant vector field  $V$  on the manifold  $G$ . From here, we can define a one-parameter subgroup of  $G$  by solving the differential equation

$$\frac{d\phi_V}{dt} = V(\phi_V(t));$$

the solution  $\phi : \mathbb{R} \rightarrow G$  will be a group homomorphism. Finally, we define the exponential map  $\exp : T_e G \rightarrow G$  by  $v \mapsto \phi_V(1)$ . Intuitively, all this means is that we take a vector  $v \in T_e G$  and “move in the  $v$  direction” for a bit to reach  $\exp(v) \in G$ .

When we have a matrix group, the exponential map can be made much more concrete: it reduces to the standard exponential map of matrices. This gives a quick and dirty way of understanding

what the Lie algebra associated to a matrix Lie group ought to be. For example, for the special linear group  $SL(n, \mathbb{C})$ , defined to be  $n \times n$  complex matrices with determinant 1, a matrix  $A$  in the associated Lie algebra must satisfy  $\det e^A = 1$ . But for any matrix  $A$ ,  $\det e^A = e^{\operatorname{tr} A}$ , so in fact we need  $\operatorname{tr} A = 0$ . This is why we said the algebra  $\mathfrak{sl}(2, \mathbb{C})$  consists of complex  $2 \times 2$  traceless matrices.

It is important to note that, while we can associate Lie algebras to Lie groups and vice-versa, the correspondence is not one-to-one. The Lie algebra captures only local information about a group, so different groups can have the same Lie algebra. In particular, recall from the previous subsection that the group  $SU(2)$  is a double cover of  $SO(3)$ , so locally they should look the same. This is made manifest in their algebras.

**Example 4.2.** Determine the Lie algebras  $\mathfrak{su}(2)$  and  $\mathfrak{so}(3)$  (over the field  $\mathbb{R}$ ), and show that they are isomorphic.

**Solution:** If  $A \in \mathfrak{su}(2)$ , then  $e^A \in SU(2)$ , so

$$(e^A)(e^A)^\dagger = e^{A+A^\dagger} = I.$$

Therefore,  $A$  is anti-Hermitian. Additionally,  $\det e^A = 1$ , so  $\operatorname{tr} A = 0$ . It follows that  $A = iH$ , where  $H$  is a trace-free Hermitian  $2 \times 2$  complex matrix. A basis for these is given by

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

so a basis for  $\mathfrak{su}(2)$  is given by  $i\sigma_j$  for  $j = 1, 2, 3$ . By taking commutators, we can verify that the bracket is given by

$$[i\sigma_a, i\sigma_b] = -2\epsilon_{abc}(i\sigma_c).$$

If  $A \in \mathfrak{so}(3)$ , then  $e^A \in SO(3)$ , so

$$(e^A)(e^A)^\top = e^{A+A^\top} = I.$$

Therefore,  $A$  is antisymmetric. This also guarantees that  $\operatorname{tr} A = 0$ , so  $\det e^A = 1$ . A basis for  $3 \times 3$  antisymmetric matrices is given by

$$T_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}, \quad T_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \quad T_3 = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

By taking commutators, we find

$$[T_i, T_j] = \epsilon_{ijk}T_k.$$

It is immediate to show that the isomorphism between the two algebras is given by  $i\sigma_i \mapsto -2T_i$ .

### 4.3 Principal Bundles

At last, we are ready to the mathematical object which naturally underlies any gauge theory. A principal  $G$ -bundle is a fiber bundle for which the fibers are all copies of some Lie group  $G$  (or more precisely, they are all  $G$ -torsors). More formally, a fiber bundle  $\pi : E \rightarrow B$  is called a principal  $G$ -bundle if it can be equipped with a  $G$ -action which is free<sup>1</sup>, transitive on each fiber, and leaves the fibers invariant (that is, the projection  $\pi$  commutes with the action of any  $g \in G$ ).

As we have alluded to previously, a principal  $G$ -bundle captures exactly what we mean by a gauge symmetry. Recall that a gauge symmetry is an internal symmetry of the theory which is local, i.e., it can act at each spacetime point. If we take the base  $B$  to be spacetime, then a principal  $G$ -bundle  $E \rightarrow B$  assigns a copy of the symmetry group  $G$  to each point of spacetime. When we “gauge” a symmetry – that is, take a global symmetry and make it local – what we are doing mathematically is taking the group  $G$  and replacing it with a principal  $G$ -bundle over spacetime.

A common example of a principal bundle is the frame bundle for a manifold. The frame bundle of a manifold  $M$  assigns to each point  $p \in M$  the fiber consisting of all possible frames at  $p$ ; that is, all possible bases for  $T_pM$ . Clearly any two frames can be related by an invertible linear transformation, and only the identity transformation can leave a frame invariant, so the fibers are indeed  $GL(n)$ -torsors. It is clear from this example that, although each fiber is homeomorphic to  $GL(n)$ , there is no canonical isomorphism matching a fiber with  $GL(n)$ , since there is no frame which naturally corresponds to the identity.

---

*Mathematical aside:* The frame bundle has many subbundles corresponding to the subgroups of  $GL(n)$ . If  $G \subset GL(n)$ , a  $G$ -subbundle of the frame bundle is called a  $G$ -structure on a manifold. The notion of a  $G$ -structure helps to unify many of the possible structures we can define on a manifold.

For example, if we have an orientable manifold, then there is a well-defined way of deciding which frames have positive determinant and which have negative determinant. Thus, on an orientable manifold, there is a  $GL^+(n)$ -subbundle of the frame bundle, where  $GL^+(n) \subset GL(n)$  is the subgroup of invertible matrices with positive determinant. Conversely, if we can define a  $GL^+(n)$ -structure on a manifold, then it is orientable.

A more familiar example might be Riemannian manifolds. A Riemannian manifold has a metric, which tells us the angles between vectors in each tangent space. Using the metric, we have a well-defined way of telling which frames are orthogonal. This means we can define a  $O(n)$ -subbundle of the frame bundle. In this way, the notion of Riemannian structure on a manifold is presented as an  $O(n)$ -structure.

---

In order to study a principal bundle, we will need to define a connection on it in order to take derivatives. However, we will not be able to directly use the tools we developed for vector bundles. Recall that for vector bundles, we explicitly defined a frame and then decomposed the connection into this frame, writing  $\nabla e_i = e_j \otimes \omega_i^j$ . In this case, the fibers need not have the structure of a vector space, so this approach will not work.

Instead, we have to think more abstractly about the meaning of the connection, and adapt it to

---

<sup>1</sup>The “principal” in principal  $G$ -bundle is the same as in principal homogeneous space; the fibers are principal homogeneous spaces for  $G$ . If we drop the requirement that the action be free, then we have only a  $G$ -bundle.

an arbitrary fiber bundle. In a vector bundle, the connection gives a notion of parallel transport, via the differential equation  $\nabla v = 0$ . Similarly, in a fiber bundle, we would like to have an idea of what it means for a section to be “parallel.” That is, when we move through the total space  $E$  in a fiber bundle  $\pi : E \rightarrow B$ , when are we changing position in the fiber and when are we not?

We can think about this question concretely. When we move away from a point  $p \in E$ , we are moving in a direction contained in  $T_p E$ . There should be some subset of these directions which are “horizontal,” that is, directions which move us over the base space  $B$  without changing position in the fiber. Moreover, if we add two of these directions together we should obtain another horizontal direction. This means we can define a vector subspace  $H_p \subset T_p E$  of horizontal directions in the tangent space of the bundle.

In the case of a vector bundle, a point of the bundle is the pair of a base point  $b \in B$  and a vector  $v$  in the fiber. The horizontal directions are then defined by pairs of directions  $X \in T_b B$  and solutions to  $\nabla_X v = 0$ . That is, as soon as we define a linear connection, we know the horizontal subspace  $H_{(b,v)} \subset T_{(b,v)} E$ . In a general fiber bundle, we will take the horizontal subspace to be the fundamental object. The *Ehresmann connection* on a fiber bundle  $\pi : E \rightarrow B$  is simply a choice of a smooth subbundle  $H \subset TE$ , called the horizontal bundle, which defines a horizontal subspace  $H_p \subset T_p E$  at each point  $p \in E$ .

There are, of course, some conditions on the horizontal bundle in order for it to be sensible. Of course the horizontal subspace  $H_p$  should be a smooth function of  $p$ . In addition, it needs to be compatible with the projection  $\pi : E \rightarrow B$  in the following sense. The differential  $d\pi$  defines a map  $TE \rightarrow TB$ , which projects a tangent vector of the total space down to a tangent vector of the base space. The kernel of this subspace at  $p \in E$  is called the vertical subspace  $V_p$ , since these vectors have no component parallel to the base space. Naturally, we would want the horizontal subspace to be orthogonal to the vertical subspace, but we do not have an inner product structure on  $T_p E$ , so this is too much to ask. Instead, we ask for a direct sum decomposition

$$T_p E = H_p \oplus V_p.$$

Recall this means that  $H_p \cap V_p = \{0\}$ , and any vector in  $T_p E$  can be written as a sum of a vector from  $H_p$  and a vector from  $V_p$ .

The Ehresmann connection gives a very direct way of defining parallel transport. Given a path  $\gamma : [0, t] \rightarrow B$ , and an Ehresmann connection  $H \subset TE$  on the fiber bundle  $\pi : E \rightarrow B$ , we can parallel transport a point  $p \in E$  (with  $\pi(p) = \gamma(0)$ ) by moving in the horizontal direction. That is, we define a lift  $\tilde{\gamma} : [0, t] \rightarrow E$  such that  $\pi \circ \tilde{\gamma} = \gamma$  and  $d\pi(\tilde{\gamma}') \in H_{\gamma(t)}$ .

**Example 4.3.** Prove that the horizontal lift  $\tilde{\gamma}$  of a path  $\gamma : [0, t] \rightarrow B$  is unique.

**Solution:** Consider a point  $b \in B$  and a vector  $v \in T_b B$ . If we are given a point  $p \in \pi^{-1}(b)$ , then there is a unique horizontal lift of  $v$  to a vector  $V \in T_p E$ . Indeed, assume we have two different vectors  $V_1, V_2 \in H_p E$  such that  $d\pi(V_1) = d\pi(V_2) = v$ . Then  $d\pi(V_1 - V_2) = 0$ , which implies  $V_1 - V_2 \in V_p E$ . But since

$$V_1 = V_2 + (V_1 - V_2),$$

we must have  $V_1 - V_2 = 0$  (since  $T_p E = H_p E \oplus V_p E$ ). Therefore,  $V_1 = V_2$ .

This result implies that there is a unique horizontal lift of the tangent vector  $\gamma'(x)$ , for each  $x \in [0, t]$ . Integrating these unique vectors gives the unique lift  $\tilde{\gamma}$ .

An Ehresmann connection can be defined on any fiber bundle. Naturally, we should expect some restrictions when we work with a principal  $G$ -bundle. Since the action of the Lie group  $G$  gives a canonical diffeomorphism between any two points of the same fiber, we require that the horizontal subspaces within some fiber are carried to each other by the pushforward of these diffeomorphisms. More precisely, let  $g \in G$  and  $p \in E$ . We have a left-translation map  $L_g : E \rightarrow E$  which is a diffeomorphism, and we require that

$$H_{L_gp}E = (L_g)_*(H_pE).$$

This condition (along with other conditions involving compatibility with Lie group structure) is sometimes referred to as  $G$ -equivariance.

An Ehresmann connection on a principal bundle which satisfies this  $G$ -equivariance condition can also be thought of in terms of the Lie algebra  $\mathfrak{g}$ . At each point  $p \in E$ , the vertical subspace  $V_pE$  consists of all the directions we can move through the fiber. Since the fiber is a principal homogeneous space for the Lie group  $G$ ,  $V_pE$  is isomorphic to  $\mathfrak{g}$ . If we choose a linear map  $\phi : T_pE \rightarrow \mathfrak{g}$  which exhibits this isomorphism when restricted to  $V_pE$ , then  $\ker \phi$  is naturally identified with  $H_pE$ . Conversely, if we have an Ehresmann connection which tells us  $H_pE$  at each point  $p$ , then we can uniquely construct a linear map  $\phi_p : T_pE \rightarrow \mathfrak{g}$  which has  $H_pE$  as its kernel and which becomes the isomorphism  $V_pE \cong \mathfrak{g}$  when restricted to  $V_pE$ . The  $G$ -equivariance condition of  $H_pE$  guarantees that the maps  $\phi_p$  will be  $G$ -equivariant, in the sense that

$$\phi_{L_gp} = (L_g)_*\phi_p.$$

Thus, instead of thinking about an Ehresmann connection, we can think about linear maps from  $T_pE$  to  $\mathfrak{g}$  at each point  $p \in E$ , all of which respect the isomorphism  $V_pE \cong \mathfrak{g}$ . The space of linear maps from  $T_pE$  to  $\mathfrak{g}$  can be thought of as the fiber of the tensor product bundle  $\mathfrak{g} \otimes T_p^*E$ , so a connection on a principal  $G$ -bundle is a member of the space of sections  $\Gamma(\mathfrak{g} \otimes T^*E)$  (where  $\mathfrak{g}$  in this context is understood to mean the trivial bundle  $E \times \mathfrak{g}$ ).

A section of  $\mathfrak{g} \otimes T^*E$  is a  $\mathfrak{g}$ -valued one-form on  $E$ . This is akin to the connection form on a vector bundle, which is a matrix-valued one-form, except on the base space  $B$ . It is strange that a connection on a principal bundle would require a one-form over the entire bundle, while the connection on a vector bundle only requires a one-form over the base space. Indeed, this distinction is illusory; the  $G$ -equivariance condition allows us to specialize to only a  $\mathfrak{g}$ -valued one-form on the base space. To see this, let  $\omega \in \Gamma(\mathfrak{g} \otimes T^*E)$  be the connection, and let  $a \in \Gamma(E)$  be any section. We can define a  $\mathfrak{g}$ -valued one-form over  $B$ , which we will denote  $\tilde{\omega}$ , using the pullback through  $a$ :

$$\tilde{\omega}(b) = \omega(a(b)).$$

The one-form  $\tilde{\omega}$  uniquely specifies a  $G$ -equivariant one-form over the entire bundle  $E$ . To find a value of  $\omega$  at  $p \in E$ , we evaluate  $a(\pi(p))$ , and then find the unique group element  $g$  such that  $g \cdot a(\pi(p)) = p$ . It follows that

$$\omega(p) = (L_g)_*\tilde{\omega}(b).$$

In summary: a connection on a fiber bundle tells us which tangent directions in the bundle are horizontal, i.e., which directions correspond to parallel transport. On a principal  $G$ -bundle, we require that the choice of horizontal subspace is  $G$ -equivariant. Since the vertical subspaces of a principal  $G$ -bundle are isomorphic to the algebra  $\mathfrak{g}$ , we can alternatively view a connection on a principal  $G$ -bundle as a  $G$ -equivariant  $\mathfrak{g}$ -valued one-form over the whole bundle. We can be more economical by instead representing the connection as a  $\mathfrak{g}$ -valued one-form over only the base space, using  $G$ -equivariance to recover the full connection.

Our final task is to understand the curvature of a principal bundle. Recall that a connection is in some sense basis or frame dependent (or in more physical language, gauge dependent), while the curvature carries robust geometric information about a bundle. For a vector bundle, we defined the curvature as the covariant derivative of the connection form, which took the form

$$\Omega = d\omega + \omega \wedge \omega.$$

We will find essentially the same definition holds in the case of a principal bundle. However, since we are now discussing the Ehresmann connection, it is worth defining the covariant derivative in this more abstract setting, and seeing how it reduces down to more familiar algebraic expressions. Associated with any Ehresmann connection is a horizontal projection  $h : TE \rightarrow TE$  which acts as the identity on each  $H_pE$  and annihilates each  $V_pE$ . We define the covariant derivative on  $X$ -valued differential forms by

$$\nabla : \Omega^k(E, X) \rightarrow \Omega^{k+1}(E, X), \quad \nabla \alpha = h^* d\alpha = d\alpha(hv_1, \dots, hv_{k+1}).$$

We define the curvature on a principal bundle to be the covariant derivative of the connection form, so  $\Omega = h^* d\omega$ . To understand how this expression works, we can try evaluating the curvature form on vectors. Since  $T_pE = H_pE \oplus V_pE$ , it is enough to consider three cases: both vectors horizontal, both vectors vertical, and one horizontal, one vertical. If we evaluate  $\Omega$  on two horizontal vectors  $u$  and  $v$ , then we are done quickly:

$$\Omega(u, v) = h^* d\omega(u, v) = d\omega(hu, hv) = d\omega(u, v).$$

If at least one of  $u$  and  $v$  is vertical, we are done even more quickly:  $\Omega(u, v) = 0$ .

If we are to write  $\Omega$  without using  $h$ , then it should be a sum of  $d\omega$  and some term which vanishes when both  $u$  and  $v$  are horizontal. Since  $\omega$  vanishes on horizontal vectors, it is a reasonable candidate. Consider the expression

$$d\omega(u, v) + [\omega(u), \omega(v)].$$

This agrees with  $\Omega$  when  $u$  and  $v$  are horizontal. If  $u$  and  $v$  are both vertical, then the expression ought to vanish. Indeed, evaluating the exterior derivative gives

$$d\omega(u, v) = u(\omega(v)) - v(\omega(u)) - \omega([u, v]).$$

The last term cancels  $[\omega(u), \omega(v)]$ , and the first two terms cancel each other since the derivatives commute, so our expression vanishes. Similarly, if  $u$  is horizontal and  $v$  is vertical, then  $\omega(u)$  vanishes and we are left with

$$d\omega(u, v) = -\omega([u, v]).$$

The Lie bracket of a horizontal and a vertical vector field is still horizontal (this is a consequence of the  $G$ -equivariance of the connection), so this vanishes as well. It follows that

$$\Omega(u, v) = d\omega(u, v) + [\omega(u), \omega(v)]$$

for any vector fields  $u, v \in \Gamma(TE)$ .

Just as in the case of vector bundles, the covariant derivative of  $\Omega$  vanishes (this is called the Bianchi identity). One advantage of our abstract definition of the covariant derivative is that it makes the proof of this fact simple:

$$\begin{aligned} \nabla\Omega(u, v, w) &= d\Omega(hu, hv, hw) \\ &= (d^2\omega + d([\omega, \omega]))(hu, hv, hw) \\ &= ([d\omega(hu, hv), \omega(hw)] - [\omega(hu), d\omega(hv, hw)]) \\ &= 0, \end{aligned}$$

since  $\omega$  annihilates any horizontal vector field.

The details of this story are important to gauge theory, but the main message is simpler: if we have a principal  $G$ -bundle, then a connection on this bundle and its curvature are naturally expressed in terms of the Lie algebra  $\mathfrak{g}$ . As we have already suggested, the curvature form will be associated with the gauge field, so the structure of the Lie algebra of the gauge group is tied to the nature of the particles which carry that field. As a preview of this fact, recall that the gauge group for the Standard Model is  $U(1) \times SU(2) \times SU(3)$ , corresponding to an electric, weak, and strong sector respectively (or more precisely, an electroweak and a strong sector). These groups have dimension 1, 3, and 8 respectively. The force-carrying particles are the photon (1 particle) for electromagnetism; the  $W^\pm$  and  $Z$  bosons (3 particles) for the weak force; and 8 flavors of gluons for the strong force.

## 5 Electrodynamics as a Gauge Theory

We are at last ready to understand on a deep level what it means for electrodynamics to be a  $U(1)$  gauge theory. We have seen some aspects of this in Section 3.4, but here we will develop the physics more fully.

The key idea in mapping the mathematics of a principal bundle to a physical gauge theory is to associate the connection with the gauge potential, and the curvature with the field strength. This means that both the gauge potential and the gauge field are Lie algebra valued.

In the case of electrodynamics, our principal bundle is a  $U(1)$ -bundle over spacetime. This is not a unique geometric object; there are many possible  $U(1)$ -bundles. The question of which  $U(1)$ -bundle we are working with is the same as the question of what configuration of the electromagnetic field we are working with.

To make this more precise, recall the correspondence defined in Section 3.4,

$$\omega = -i \frac{e}{hc} A.$$

Now that we are working with a  $U(1)$ -bundle, we can give a more satisfying explanation for the factor of  $i$ . Since  $\omega$  is a connection on a  $U(1)$ -bundle, it lives in the Lie algebra of  $U(1)$ , which consists of pure imaginary numbers. In order to keep the gauge field  $A$  real-valued, we must include this factor of  $i$ .

The curvature is defined in general by the formula given in the previous section,

$$\Omega(u, v) = d\omega(u, v) + [\omega(u), \omega(v)].$$

Since  $U(1)$  is Abelian, the Lie bracket in the second term on the right vanishes identically, so we can drop it. It follows that

$$\Omega = d\omega = -i\frac{e}{\hbar c}dA = -i\frac{e}{\hbar c}F,$$

where  $F = \frac{1}{2}F_{\mu\nu}dx^\mu dx^\nu$ . The Bianchi identity tells us that  $\nabla\Omega = 0$ , which implies  $\nabla F = 0$ . The covariant derivative is the sum of the exterior derivative and a Lie bracket term, as we saw with the curvature; since  $U(1)$  is Abelian, the Lie bracket will vanish, and so  $dF = 0$ . This equation contains Gauss's law for magnetism and Faraday's law, as we saw in Section 3.4.

To understand the other two Maxwell equations, we need to write down a Lagrangian for the electromagnetic field. We know, of course, that the Lagrangian density is  $-\frac{1}{4}F_{\mu\nu}F^{\mu\nu}$ . However, in preparation for the next section, we will express this in a more general context, and then specialize back to the electromagnetic field.

Recall from Example 2.3 that taking the inner product of two  $k$ -forms  $a, b$  amounts to taking  $a \wedge \star b$ . This will be the  $n$ -form  $\langle a, b \rangle \omega$ , where  $\omega$  is the volume-form and  $\langle a, b \rangle = a_{\mu_1 \dots \mu_k} b^{\mu_1 \dots \mu_k}$ . Thus, instead of  $-\frac{1}{4}F_{\mu\nu}F^{\mu\nu}$ , we can write  $-F \wedge \star F$ . This is a quantity that we can integrate, since it is proportional to the volume form.

However, in a general gauge theory,  $F$  is Lie algebra valued. In order to form a Lagrangian density, we need a real number. Thus, we will need a real symmetric bilinear form on the Lie algebra. An obvious choice would be the Killing form, but this vanishes in an Abelian algebra, so we would not be able to form a  $U(1)$  gauge theory. Instead of the Killing form, we will use the trace form,  $\text{tr}(xy)$ . You can show that in a simple Lie algebra, the trace form and the Killing form are multiples of one another.

We could be even more direct by writing the Lagrangian density in terms of the curvature  $\Omega$ . If we set  $\hbar = c = 1$ , then we have

$$\frac{1}{e^2} \text{tr}(\Omega \wedge \star \Omega) = -\text{tr}(F \wedge \star F).$$

This convention allows us to avoid the matter of choosing a constant of proportionality between the connection and the gauge field. Instead, we take the gauge field to be  $A = i\omega$ , and the field strength to be  $F = \nabla A$ . A physical constant only enters when we write the Lagrangian density in terms of a gauge coupling  $g$  as

$$\mathcal{L} = \frac{1}{g^2} \text{tr}(F \wedge \star F).$$

In the case of electromagnetism, the gauge coupling is given by  $e$ .



**6 Yang-Mills Theories**

**7 Reduction of Symmetry**

**8 Renormalization of Gauge Couplings**

**9 Wilson Loops**

**10 Lattice Gauge Theory**

## **References**

- [1] K. Young. “Foreign exchange market as a lattice gauge theory”. In: *American Journal of Physics* 67 (Oct. 1999), pp. 862–868. DOI: 10.1119/1.19139.