Quantum Theory for Topologists

Daniel Dugger

Contents

Preface

For a long time I have wanted to learn about the interactions between mathematical physics and topology. I remember looking at Atiyah's little book *The geometry and physics of knots* [**At**] as a graduate student, and understanding very little. Over the years, as my knowledge of mathematics grew, I periodically came back to that little book—as well as other introductory texts on the subject. Each time I found myself almost as lost as when I was a graduate student. I could understand some bits of the topology going on, but for me it was clouded by all kinds of strange formulas from physics and differential geometry (often with a morass of indices) for which I had no intuition. It was clear that learning this material was going to take a serious time commitment, and as a young professor—eventually one with a family—time is something that is hard to come by. Each year I would declare to myself "I'm going to start learning physics this year," but then each year my time would be sucked away by writing papers in my research program, teaching courses, writing referee reports, and all the other things that go with academic life and are too depressing to make into a list.

Finally I realized that nothing was ever going to change, and I would *never* learn physics—unless I found a way to incorporate that process into the obligations that I already had. So I arranged to teach a course on this subject. In some ways it was a crazy thing to do, and it had some consequences: for example, my research program took a complete nosedive during this time, because all my energy went into learning the material I was putting into my lectures. But it was fun, and exciting, and somehow this process got me over a hump.

The notes from this course have evolved into the present text. During the course, my lectures were put into LaTeX by the attending students. Afterwards I heavily revised what was there, and also added a bunch of additional material. I am grateful to the students for doing the intial typesetting, for asking good questions during the course, and for their patience with my limited understanding: these students were Matthew Arbo, Thomas Bell, Kevin Donahue, John Foster, Jaree Hudson, Liz Henning, Joseph Loubert, Kristy Pelatt, Min Ro, Dylan Rupel, Patrick Schultz, AJ Stewart, Michael Sun, and Jason Winerip.

WARNING: The present document is a work in progress. Certain sections have not been revised, further sections are being added, some things are occasionally moved around, etc. There are plenty of false or misleading statements that are gradually being identified and removed. Use at your own risk!

CHAPTER 1

Introduction

For this first lecture all you need to know about quantum field theory is that it is a mysterious area of modern physics that is built around the “Feynman path integral”. I say it is mysterious because these path integrals are not rigorously defined and often seem to be divergent. Quantum field theory seems to consist of a collection of techniques for getting useful information out of these ill-defined integrals.

Work dating back to the late 1970s shows that QFT techniques can be used to obtain topological invariants of manifolds. Schwartz did this in 1978 for an invariant called the Reidemeister torsion, and Witten did this for Donaldson invariants and the Jones polynomial in several papers dating from the mid 80s. Shortly after Witten’s work Atiyah wrote down a definition of “topological quantum field theory”, and observed that Witten’s constructions factored through this. One has the following picture:

$$\begin{tikzcd}
\text{QFT techniques} \arrow[dd, squiggly] \arrow[dr, squiggly] & \\
& \text{TQFT} \arrow[dl] \\
\text{topological invariants of manifolds} &
\end{tikzcd}$$

Here the squiggly lines indicate non-rigorous material!

Much effort has been spent over the past twenty-years in developing rigorous approaches to constructing TQFTs, avoiding the path integral techniques. The names Reshetikin and Turaev come to mind here, but there are many others as well. However, this course will not focus on this material. Our goal will instead be to understand what we can about Feynman integrals and the squiggly lines above. We will try to fill in the necessary background from physics so that we have some basic idea of what QFT techniques are all about. There’s quite a bit of material to cover:

$$\begin{tikzcd}
\text{Classical mechanics} \arrow[r] \arrow[d] & \text{Classical field theory} \arrow[d] \\
\text{Quantum mechanics} \arrow[r] & \text{Quantum field theory.}
\end{tikzcd}$$

The plan is to spend about a week on each of these areas, giving some kind of survey of the basics.

A good reference for this foundational material is the set of lecture notes by Rabin [**R**]. However, Rabin's lectures get technical pretty quickly and are also short on examples. We will try to remedy that.

1.0.1. Introduction to TQFTs. Let k be a field (feel free to just think about $k = \mathbb{C}$). A "$(d+1)$-dimensional TQFT" consists of

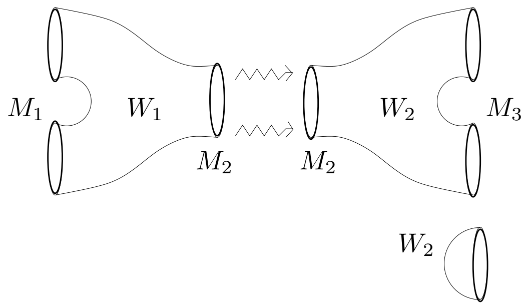
(1) For every closed, oriented d-manifold M a vector space $Z(M)$ over k;

(2) For every oriented $(d+1)$-manifold W together with a homeomorphism $h\colon \partial\Delta^W \to \overline{M}_1 \amalg M_2$, a linear map
$$\phi_{W,h}\colon Z(M_1) \to Z(M_2).$$
Here $\overline{M}_1$ denotes M_1 with the opposite orientation.

(3) Isomorphisms $Z(\emptyset) \cong k$ and $Z(M_1 \amalg M_2) \cong Z(M_1) \otimes Z(M_2)$ for every closed, oriented d-manifolds M_1 and M_2.

This data must satisfy a long list of properties which are somewhat complicated to state in this form. For this introductory lecture let us just give the basic ideas behind a few of the properties:

(i) (Composition) Suppose W_1 is a cobordism from M_1 to M_2, and W_2 is a cobordism from M_2 to M_3. An example of this is depicted here:



Gluing W_1 and W_2 results in a "composite" cobordism W_3 from M_1 to M_3, and we ask that $\phi_{W_3} = \phi_{W_2} \circ \phi_{W_1}$.

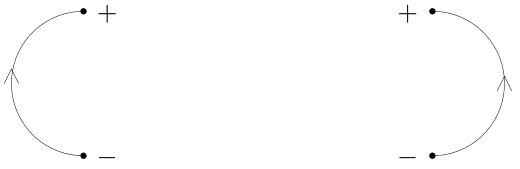
(ii) $\phi_{M\times I}\colon Z(M) \to Z(M)$ is the identity map

(iii) Each $Z(M)$ is finite-dimensional and the map
$$Z(\overline{M}) \otimes Z(M) \cong Z(\overline{M} \amalg M) \xrightarrow{\phi_{M\times I}} Z(\emptyset) \cong k$$
is a perfect pairing. Here $M \times I$ is being thought of as a cobordism between $\overline{M} \amalg M$ and $\emptyset$.

How do we get invariants of manifolds from a TQFT? If W is a closed, oriented $(d+1)$-manifold then it can be thought of as a cobordism from $\emptyset$ to itself, and therefore gives rise to a linear map
$$\phi_W\colon Z(\emptyset) \to Z(\emptyset).$$
Since $Z(\emptyset) \cong k$, this map is just multiplication by an element of k. This element (which we will also denote as ϕ_W) is an invariant of W.

Next let us look at some examples. When $d = 0$, the oriented d-manifolds are just finite sets with the points labelled by plus and minus signs (for the orientation). Let us write pt_+ and pt_- for a point with the two orientations. If $Z(pt_+) = V$, then we know $Z(pt_-) = V^*$ (canonically), and in general Z of any oriented 0-manifold will be a tensor product of V's and V^*'s.

All cobordisms between oriented 0-dimensional manifolds will break up into identity cobordisms and the two types



Applying ϕ to the first gives a map $k \to V \otimes V^*$, and applying ϕ to the second gives $V \otimes V^* \to k$. A little work shows that the second map is just evaluation, namely it sends a tensor $u \otimes \alpha$ to $\alpha(u)$. The first map is "coevaluation", which is slightly harder to describe. If $e_1, \dots, e_n$ is a basis for V, and $f_1, \dots, f_n$ is the dual basis for V^*, then coevaluation sends 1 to $\sum_i e_i \otimes f_i$. One can check that this is independent of the choice of basis for V. (For a basis-independent construction, use the canonical isomorphism $V \otimes V^* \cong \operatorname{End}(V)$, and then take $k \to \operatorname{End}(V)$ to send 1 to the identity map).

What invariant does this TQFT assign to a circle? By breaking up the circle into its two hemispheres, one sees that $\phi_{S^1} : k \to k$ is the composite

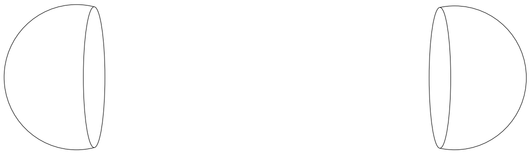
$$k \xrightarrow{coev} V \otimes V^* \xrightarrow{ev} k.$$

This map sends 1 to $\sum_i f_i(e_i) = \sum_i 1 = \dim V$.

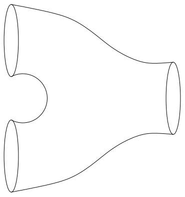
Exercise 1.0.2. What invariant does the TQFT assign to a disjoint union of r circles?

Now let us turn to $d = 1$. Closed 1-manifolds are just disjoint unions of circles, so the $Z(-)$ construction is in some sense determined by $Z(S^1)$. Let $V = Z(S^1)$.

The two corbodisms



give maps $\epsilon \colon k \to V$ and $\operatorname{tr} \colon V \to k$. The "pair of pants" cobordism



gives a map $\mu \colon V \otimes V \to V$. We claim that this product makes V into an associative, commutative algebra with unit ϵ such that the pairing $V \otimes V \to k$ given by $a, b \mapsto \operatorname{tr}(ab)$ is perfect. One sees this by drawing lots of pictures:

????

Exercise 1.0.3. Determine the invariant that this TQFT assigns to each oriented surface Σ_g, in terms of the above structure on V. [Hint: The invariant attached to Σ_0 is $\mathrm{tr}(1)$.]

Finally let us look at $d = 2$. The point of this whole discussion is to observe that while things are very easy for $d = 0$ and $d = 1$, for $d = 2$ the situation is much harder. One must give each vector space $Z(\Sigma_g)$, and a homomorphism ϕ_W for each oriented 3-manifold. This is a lot of information!

The most basic example of a $(2 + 1)$-dimensional TQFT is something called Chern-Simons theory. It was constructed by Witten in his 1989 paper [**W**] using QFT techniques, and much work has since been done to create a rigorous construction. Here is a pathetic outline, only intended to show how daunting the theory is to a newcomer. We will only say how to get the invariants of closed 3-manifolds ϕ_M.

Fix a compact Lie group G, and fix an integer $k \in Z$ (called the "level"). There will be one TQFT for each choice of G and k. Let $\mathfrak{g}$ be the Lie algebra of G.

Let M be a 3-manifold, and let $P \to M$ be the trivial G-bundle over M. If A is a $\mathfrak{g}$-valued connection on P, define the "action"

$$S(A) = \frac{k}{4\pi} \int_M \mathrm{tr}(A \wedge dA + \frac{2}{3} A \wedge A \wedge A).$$

Then

$$\phi_M = \int_A e^{iS(A)} dA$$

where this is a Feynman integral over the (infinite-dimensional) space of all connections A on P. This integral is not really defined! Witten shows that if $C_1, \ldots, C_k$ are knots in M then he could also use this kind of integral formalism to construct invariants

$$\phi_{M,C_1,\ldots,C_k}$$

and in this way he recovered the Jones polynomial for knots in $\mathbb{R}^3$.

This now lets me state the second goal of the course: Understand the basic constructions underlying Chern-Simons theory and Witten's original paper [**W**] on this subject.

We will see how far we get with all this!

Part 1

A first look into modern physics

CHAPTER 2

Classical mechanics

2.1. Lagrangian mechanics

In the next few lectures we will give a very brief introduction to classical mechanics. The classic physics reference is [**Go**], and the classic mathematical reference is [**A**]. I must say, though, that I find both of these books unpleassant to read (in different ways). I have found the two-volume set [**De**] more useful. The Rabin lectures [**R**] also give a quick introduction.

2.1.1. Motivation. To get ourselve started, consider a mass that hangs vertically from a spring, suspended from the ceiling. Let $x(t)$ be the position of the mass at time t, where x is measured vertically with $x = 0$ being the ceiling and the positive direction going down. We will use the physics notation

$$\dot{x} = \dot{x}(t) = \frac{dx}{dt}, \qquad \ddot{x} = \ddot{x}(t) = \frac{d^2x}{dt^2}.$$

Let k be the spring constant, and l its unstretched length. Then Newton's second law states that $x(t)$ satisfies the following second order differential equation:

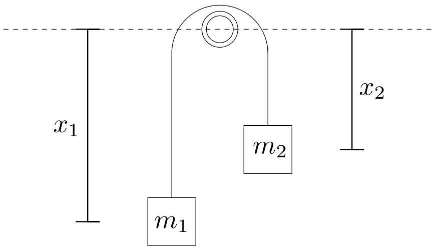
$$m\ddot{x} = F = mg - k(x - l).$$

If we solve this differential equation (subject to whatever initial values are of interest to us), then we have understood the motion of the mass.

In this example the object is moving in one dimension. If we had movement in three dimensions then $\mathbf{x}(t)$ would be a vector in $\mathbb{R}^3$, and we would have the vector equation $m\ddot{\mathbf{x}} = \mathbf{F}$. This consists of three second order differential equations, one in each of the three coordinates.

Newton's law gives a perfectly complete approach to the equations of motion for a system of particles. But if one applies these methods in 100 examples, getting progressively more difficult, one finds that the information contained in Newton's equations is not organized very efficiently, or in a way that interacts well with whatever underlying geometry there may be in the given system. Let us give a simple example of this.

Example 2.1.2. Consider the Atwood machine, consisting of two masses m_1 and m_2 tied together by a rope of length l that hangs over a pulley. Let x_1 and x_2 denote the vertical positions of each mass, measured downward from the baseline of the pulley. Notice that $x_2 = l - x_1$ (in the idea situation where the pulley has zero diameter!)



If we approach this system using Newton's laws then there are two equations, one for each mass:

$$m_1\ddot{x}_1 = m_1 g - T, \qquad m_2\ddot{x}_2 = m_2 g - T.$$

Here T denotes the tension in the rope, which is experienced as an upwards force by both masses. The constraint $x_2 = l - x_1$ gives $\ddot{x}_2 = \ddot{x}_1$, and then subtracting the two equations allows us to eliminate T and obtain

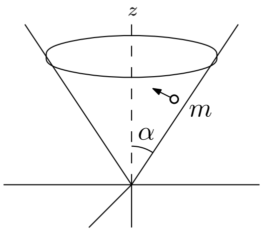
$$(m_1 + m_2)\ddot{x}_1 = (m_1 - m_2)g.$$

Notice that we can now completely solve for x_1, hence also x_2, and thereby determine the motion of the system (given some initial values for position and velocity, of course).

But look what happened here. First of all, it was clear from the beginning that there was really only one independent unknown function. Yet the method of Newton's laws forced us to write down two equations, containing the "constraint force" T, only to then have T eliminated in the end. It would be nice if we could have eliminated T from the very beginning, so that we didn't have to think about it at all!

The above example is somewhat simplistic, as it only contains two objects moving in one dimension and so it is not very difficult to solve. Imagine instead a system of 10 objects moving in three dimensions, with various constraints between them. Now we will have 30 equations, and it might not be so easy to see how to eliminate the constraint forces to reduce this down to a more manageable size. We will see that this is one of the things that Lagrangian mechanics will accomplish for us: it gives us an approach to mechanical systems where these extra "constraint forces" disappear, becoming absorbed into the geometry of the situation.

As a second motivating example, consider a vertical cone with the vertex at the bottom, centered around the z-axis. Let α denote the angle the cone makes with the z-axis (so that the cone is given by the equation $r = z\tan\alpha$). Imagine a bead that sits on the interior edge of the cone and rolls around—we would like to describe its motion.



If this is approached via Newton's laws then we set things up in Cartesian coordinates and we have to introduce the constraint force, which in this case is the normal force that the cone exerts on the bead (that keeps the bead from falling to the ground). We will get three second-order differential equations, one for each of the Cartesian coordinates. But clearly this is really a two-dimensional problem, and we would be much better off using the cylindrical coordinates r and θ to describe the motion: we really should be worrying about only two differential equations.

2.1.3. The Lagrangian approach. Let's return to the Atwood machine. I am going to describe a different method for writing down the equations of motion, and then we will apply the same method for the cone problem. For the moment I am not going to explain why the method works, only demonstrate it.

First we identify a set of independent position coordinates that completely describe the system. For the Atwood machine we can just take x_1 (since x_2 is determined by x_1).

Next we write down a certain function $L(x_1, \dot{x}_1)$, where here x_1 and $\dot{x}_1$ are regarded as formal variables. In all the cases we will nconsider the function L will be the kinetic energy of the system minus the potential energy. For the Atwood machine this is

$$L = K.E. - P.E. = \Big(\frac{1}{2}m_1\dot{x}_1^2 + \frac{1}{2}m_2\dot{x}_2^2\Big) - \Big(-m_1x_1g - m_2x_2g\Big).$$

But we want to write L as a function of x_1 and $\dot{x}_1$, in which case we have

$$L(x_1,\dot{x}_1) = \frac{1}{2}(m_1+m_2)\dot{x}_1^2 + g(m_1-m_2)x_1 + gm_2l.$$

Next we write down the so-called "Euler-Lagrange equations", which are the equations of motion for our system. There is one for each position coordinate, so in this example that means a single equation:

$$\frac{\partial L}{\partial x_1} = \frac{d}{dt}\Big(\frac{\partial L}{\partial \dot{x}_1}\Big).$$

For our Atwood machine this is

$$(m_1-m_2)g = \frac{d}{dt}\Big((m_1+m_2)\dot{x}_1\Big) = (m_1+m_2)\ddot{x}_1.$$

Note that we have obtained the same differential equation as in our first treatment of the problem, but in this case it came to us right away, not as the result of simplifying a system of two equations. Although this example is very simple, this demonstrates the general idea.

So let us next look at the cone problem. We will use the cylindrical coordinates r and θ to describe the position of the bead. In Cartesian coordinates the Lagrangian is

$$L = \frac{1}{2}m(\dot{x}^2 + \dot{y}^2 + \dot{z}^2) - mgz.$$

To change into cylindrical coordinates we recall that

$$x = r\cos\theta, \qquad y = r\sin\theta, \qquad z = r\cot\alpha$$

and so

$$\dot{x} = \dot{r}\cos\theta - r(\sin\theta)\dot{\theta}, \qquad \dot{y} = \dot{r}\sin\theta + r(\cos\theta)\dot{\theta}, \qquad \dot{z} = \dot{r}\cot\alpha.$$

Note that $\dot{x}^2 + \dot{y}^2 = \dot{r}^2 + r^2\dot{\theta}^2$, and so

$$L(r, \theta, \dot{r}, \dot{\theta}) = \frac{1}{2}m\Big((1 + \cot^2\alpha)\dot{r}^2 + r^2\dot{\theta}^2\Big) - mgr\cot\alpha.$$

The Euler-Lagrange equations are

$$\frac{\partial L}{\partial r} = \frac{d}{dt}\Big(\frac{\partial L}{\partial \dot{r}}\Big) \quad \text{and} \quad \frac{\partial L}{\partial \theta} = \frac{d}{dt}\Big(\frac{\partial L}{\partial \dot{\theta}}\Big),$$

which become

$$mr\dot{\theta}^2 - mg\cot\alpha = m(\csc^2\alpha)\ddot{r}, \qquad 0 = \frac{d}{dt}\Big(mr^2\dot{\theta}\Big).$$

The second equation says that the quantity $mr^2\dot{\theta}$ is a **constant of motion**, meaning that it assumes the same value at all points in the bead's path (if you know some physics you might recognize this as the angular momentum about the z-axis). We will write

$$mr^2\dot{\theta} = C.$$

Note that this constant could be determined from whatever initial conditions we set for the bead's motion. We can now eliminate $\dot{\theta}$ from the first differential equation to get a second order differential equation in r. If our goal is to understand the bead's motion we should solve this differential equation. We will not do this here, but let us at least observe the following cute fact. Suppose we want the bead to spin forever in a horizontal circle. For this we would need $\dot{r} = 0$, and so $\ddot{r} = 0$ as well. The first Euler-Lagrange equation then gives

$$\dot{\theta} = \sqrt{\frac{g\cot\alpha}{r}} = \sqrt{\frac{g\cot^2\alpha}{z}}.$$

So for each height on the cone, if we give the bead an initial angular velocity according to the above formula, it will spin around the cone forever staying at the same height (of course in real life there is friction!) This is kind of a cute fact.

2.1.4. Summary of the Lagrangian approach. Extrapolating from the two examples we have considered, here is the general setup for Lagrangian mechanics. For whatever system we are considering we have a smooth manifold M consisting of all possible positions—or configurations—for the objects in the system. This manifold is called **configuration space** in the physics literature. For the Atwood machine M is just the real line (or perhaps an appropriate open interval), whereas for the cone problem M is just the cone itself (but excluding the singular point at the bottom, where certainly the physics we are developing doesn't hold anymore).

Let us write $q_1, \ldots, q_n$ for some choice of local coordinates on M.

Let TM denote the tangent bundle of M. This is also a smooth manifold, called "phase space" in the physics literature. Then $q_1, \ldots, q_n, \dot{q}_1, \ldots, \dot{q}_n$ are local coordinates for TM.

The Lagrangian is a particular smooth function $L\colon TM \to \mathbb{R}$.

Given a smooth path $\gamma\colon I \to M$, we can write γ in local coordinates as

$$\gamma(t) = (q_1(t), \ldots, q_n(t)).$$

The Euler-Lagrange equations for γ are a collection of n second-order differential equations in the functions $q_i(t)$, given by

$$\frac{\partial L}{\partial q_i} = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot q_i}\right).$$

2.1.5. Coordinate invariance of the Euler-Lagrange equations. I am not going to give a derivation of the Euler-Lagrange equations from Newton's law, because you can find this in any book on classical mechanics. But I want us to check that the Euler-Lagrange formulation takes the same form in every coordinate system.

Suppose that $u_1, \dots, u_n$ are another set of local coordinates on M. Then we can write $u_i = u_i(q_1, \dots, q_n)$, where now $u_i(-, \dots, -)$ is a smooth function $\mathbb{R}^n \to \mathbb{R}^n$. Then

$$\dot u_i = \sum_j \frac{\partial u_i}{\partial q_j} \dot q_j.$$

Since we are learning to think like physicists we have to use the Einstein summation convention, which says that we drop summation symbols when we can remember them by a repeated index. So the above equation becomes simply

$$\dot u_i = \frac{\partial u_i}{\partial q_j} \dot q_j.$$

Here are two simple consequences:

$$\frac{\partial \dot u_i}{\partial \dot q_k} = \frac{\partial u_i}{\partial q_k} \tag{2.1.6}$$

and

$$\frac{\partial \dot u_i}{\partial q_k} = \frac{\partial^2 u_i}{\partial q_k q_j} \dot q_j = \frac{d}{dt}\left(\frac{\partial u_i}{\partial q_k}\right). \tag{2.1.7}$$

Now let us suppose that we have a map $\gamma\colon I \to M$ which satisfies the Euler-Lagrange equations with respect to the coordinates q_i. That is to say,

$$\frac{\partial L}{\partial q_i} = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot q_i}\right).$$

When we change into the u-coordinates, we have

$$\frac{\partial L}{\partial q_i} = \frac{\partial L}{\partial u_j}\frac{\partial u_j}{\partial q_i} + \frac{\partial L}{\partial \dot u_j}\frac{\partial \dot u_j}{\partial q_i}$$

and

$$\frac{\partial L}{\partial \dot q_i} = \frac{\partial L}{\partial u_j}\frac{\partial u_j}{\partial \dot q_i} + \frac{\partial L}{\partial \dot u_j}\frac{\partial \dot u_j}{\partial \dot q_i} = \frac{\partial L}{\partial \dot u_j}\frac{\partial \dot u_j}{\partial \dot q_i} = \frac{\partial L}{\partial \dot u_j}\frac{\partial u_j}{\partial q_i}.$$

In the second of the equalities on the previous line we have used that the u_j's only depend on the q's, not the $\dot q$'s, and so $\frac{\partial u_j}{\partial \dot q_i} = 0$. For the third equality on this line we have used (2.1.6). Taking time-derivatives of either side of this equation gives

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot q_i}\right) = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot u_j}\right)\frac{\partial u_j}{\partial q_i} + \frac{\partial L}{\partial \dot u_j}\frac{d}{dt}\left(\frac{\partial u_j}{\partial q_i}\right) = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot u_j}\right)\frac{\partial u_j}{\partial q_i} + \frac{\partial L}{\partial \dot u_j}\frac{\partial \dot u_j}{\partial q_i},$$

where in the last equality we have used (2.1.7).

Comparing the above expressions for $\frac{\partial L}{\partial q_i}$ and $\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right)$, we arrive at

$$\frac{\partial L}{\partial u_j}\frac{\partial u_j}{\partial q_i} = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{u}_j}\right)\frac{\partial u_j}{\partial q_i}.$$

This is best thought of as a matrix equation of the form (matrix)×(column vector) = (same matrix) × (column vector):

$$\left[\frac{\partial u_j}{\partial q_i}\right]_{i,j} \cdot \left[\frac{\partial L}{\partial u_j}\right]_j = \left[\frac{\partial u_j}{\partial q_i}\right]_{i,j} \cdot \left[\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{u}_j}\right)\right]_j.$$

But the matrix $[\frac{\partial u_i}{\partial q_j}]$ is invertible (since it is a coordinate change), and so we must have

$$\frac{\partial L}{\partial u_j} = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{u}_j}\right)$$

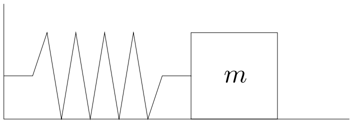
for all j.

2.2. Harmonic oscillators

Harmonic oscillators are pervasive throughout both classical and quantum mechanics, so we need to understand these pretty well. An example to keep in mind is an infinite lattice of masses connected by springs—a mattress, of sorts. If we jump up and down at one point on the mattress, then the motion of the other masses amounts to a kind of "wave" emanating from our location. In this section we will develop the machinery necessary for analyzing this kind of mechanical system (but starting out with much simpler examples).

2.2.1. The easiest oscillator problems.

Example 2.2.2. Consider a spring with one end fixed, and a mass m attached to the other end:



Let k denote the spring-constant (measuring the stiffness of the spring), and let ℓ be the length of the spring at equilibrium It is a law of physics, called Hooke's Law, that if the mass is moved a distance x from equilibrium then the spring exerts a force equal to $F = -kx$. We therefore have that

$$m\ddot{x} = F = -k \cdot x = -\frac{d}{dx}\left(\frac{k}{2}x^2\right).$$

The quantity $\frac{k}{2}x^2$ is the potential energy of the spring-mass system.

Using the Lagrangian approach to classical mechanics, we can write

$$L(x, \dot{x}) = \frac{1}{2}m\dot{x}^2 - \frac{k}{2}x^2.$$

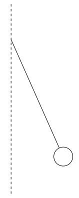
The Euler-Lagrange equation is then $\frac{\partial L}{\partial x} = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}}\right)$, which becomes

$$-kx = \frac{d}{dt}(m\dot{x}) = m\ddot{x}.$$

The solutions of course are $x(t) = A\cos\left(\sqrt{\frac{k}{m}}t + \phi\right)$, where A and ϕ can be arbitrary constants.

Remark 2.2.3. In the above example, suppose that we didn't know Hooke's Law. We would still know from physical experience that $x = 0$ is an equilibrium point of the system, which suggests that it is a local minimum for the potential energy. This would tell us that, at least in a small neighborhood of this equilibrium point, we would have $V(x) \approx Cx^2$ for some constant C. Thus we would still know that locally the potential energy is quadratic. This observation is at the heart of the ubiquity of harmonic oscillators in physical problems—small motions about a stable equilibrium point are approximately harmonic.

Example 2.2.4. For our next example, consider a simple pendulum:



If we use the coordinate θ, then the Lagrangian is

$$L(\theta, \dot{\theta}) = \frac{1}{2}m\ell^2\dot{\theta}^2 - mg\ell\big(1 - \cos(\theta)\big)$$

and the Euler-Lagrange equations become

$$\ddot{\theta} = -\frac{g}{\ell}(\sin\theta).$$

This is a complicated differential equation! But if we only look at small values of θ—in other words, small oscillations of the pendulum—then we can use the second-order approximation $\cos\theta \approx 1 - \frac{\theta^2}{2}$, or

$$L(\theta, \dot{\theta}) \approx \frac{1}{2}m\ell^2\dot{\theta}^2 - mg\ell\frac{\theta^2}{2}.$$

The resulting Euler-Lagrange equations are now $\ddot{\theta} = -\dfrac{g}{\ell}\theta$, and this gives the harmonic oscillatory motion one expects.

2.2.5. Generalization to higher dimensions. Suppose we have a physical system with coordinates $q_1, \dots, q_n$, where the Lagrangian has the form

$$L = \left(\frac{1}{2}\sum_{i,j} m_{ij}\dot{q}_i\dot{q}_j\right) - \left(\frac{1}{2}\sum_{i,j} k_{ij}q_iq_j\right)$$

with the first term being the kinetic energy of the system and the second term the potential energy. We can of course always arrange things so that the matrices

$M = (m_{ij})$ and $K = (k_{ij})$ are symmetric. Note that the Lagrangian can also be written

$$L = \frac{1}{2}\underline{\dot{q}}^T M \underline{\dot{q}} - \frac{1}{2}\underline{q}^T K \underline{q} = Q_m(\underline{\dot{q}}) - Q_k(\underline{q}).$$

where

$$\underline{q} = \begin{bmatrix} q_1 \\ \vdots \\ q_n \end{bmatrix}, \qquad \underline{\dot{q}} = \begin{bmatrix} \dot{q}_1 \\ \vdots \\ \dot{q}_n \end{bmatrix},$$

and Q_m and Q_k are the evident quadratic forms. Since $Q_m(\underline{\dot{q}})$ is the kinetic energy of the system, the quadratic form Q_m will be positive definite.

Since M is positive-definite, we can choose P such that $M = P^T \cdot I \cdot P$. Then

$$\underline{\dot{q}}^T M \underline{\dot{q}} = \big(\underline{\dot{q}}^T P^T\big)\big(P\underline{\dot{q}}\big) = \underline{\dot{r}}^T \underline{\dot{r}}$$

where $\underline{r} = P\underline{q}$, $\underline{\dot{r}} = P\underline{\dot{q}}$. In these new coordinates we have

$$L = \frac{1}{2}\underline{\dot{r}}^T \underline{\dot{r}} - \frac{1}{2}\underline{r}^T\big((P^T)^{-1} \cdot K \cdot P^{-1}\big)\underline{r},$$

so we next need to diagonalize $(P^T)^{-1} \cdot K \cdot P^{-1}$.

Choose an orthogonal matrix Q such that $(P^T)^{-1} \cdot K \cdot P^{-1} = Q^T D Q$ where

$$D = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}.$$

Introduce the new coordinates $\underline{s} = Q\underline{r}$, and note we also have $\underline{\dot{s}} = Q\underline{\dot{r}}$. Then $L(\underline{s}, \underline{\dot{s}}) = \frac{1}{2}\underline{\dot{s}}^T \underline{\dot{s}} - \frac{1}{2}\underline{s}^T D \underline{s}$, and the Euler-Lagranage equations $\frac{\partial L}{\partial s_i} = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{s}_i}\right)$ are just

$$-\lambda_i s_i = \frac{d}{dt}\big(\dot{s}_i\big) = \ddot{s}_i.$$

The solutions to the above differential equations take on different forms depending on whether each λ_i is positive, negative, or zero. When $\lambda_i > 0$ we have that s_i is a sine/cosine wave, when $\lambda_i = 0$ we have that s_i is a first degree polynomial, and when $\lambda_i < 0$ we have that s_i is an exponential function.

Assume now, really just to be specific, that all the λ_i's are positive. Then we have

$$s_i(t) = A_i \cos\big(\sqrt{\lambda_i} t + \phi_i\big)$$

for some constants A_i and ϕ_i. We can write this in vector form as

$$\underline{s}(t) = \sum A_i \cos\big(\sqrt{\lambda_i} t + \phi_i\big) \cdot \mathbf{e}_i,$$

where the $\mathbf{e_i}$'s are the standard basis for $\mathbb{R}^n$. Changing back to our original coordinates, we have that

$$\underline{q}(t) = P^{-1}\underline{r} = P^{-1}Q^{-1}\underline{s} = \sum A_i \cos\big(\sqrt{\lambda_i} t + \phi_i\big)\big(P^{-1}Q^{-1}\mathbf{e}_i\big).$$

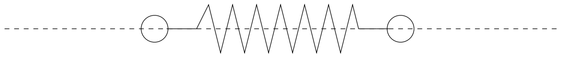
We can make the use of this formula more efficient if we remember two things:

(i) The λ_i's are roots of $\det\big(\lambda I - (P^T)^{-1} K P^{-1}\big)$, which are the same as the roots of $\det\big(\lambda P^T P - K\big) = \det\big(\lambda M - K\big)$.

(ii) The vector $P^{-1}Q^{-1}\mathbf{e}_i$ lies in the kernel of $\big(\lambda_i M - K\big)$. (We leave it as an exercise for the reader to check this.)

In practice, these two facts allow us to write down the solutions $\underline{q}(t)$ without having to find the matrices P and Q (see the examples below). The numbers $\sqrt{\lambda_i}$ are called the **characteristic frequencies** of the system, and the corresponding vectors $P^{-1}Q^{-1}\mathbf{e}_i$ are the corresponding **normal modes of oscillation**.

Example 2.2.6. Consider a system of two identical masses attached by a spring:



Assume that the spring constant is k and the equilibrium length is ℓ, and that the two masses are both equal to m. Let us assume that initially the left mass is at the origin and the right mass is at ℓ. If we tap one of the masses then the system will start to move: let $x_1 = x_1(t)$ and $x_2 = x_2(t)$ denote the positions of each mass at time t.

The Lagrangian for this system is $L = \frac{1}{2}m(\dot{x}_1^2 + \dot{x}_2^2) - \frac{1}{2}k(x_2 - x_1 - \ell)^2$. This is not one of the quadratic Lagrangians we have been considering in this section, since it contains some linear terms—but we can make it quadratic by a simple change of coordinates. Set $s_1 = x_1,\ s_2 = x_2 - \ell$, and note that the Lagrangian is now

$$L = \frac{1}{2}m(\dot{s}_1^2 + \dot{s}_2^2) - \frac{1}{2}k(s_2 - s_1)^2.$$

Note also that s_i represents the deviation of the ith object from its equilibrium position.

In the language of our discussion of quadratic Lagrangians, we now have

$$M = \begin{bmatrix} m & 0 \\ 0 & m \end{bmatrix}, \qquad K = \begin{bmatrix} k & -k \\ -k & k \end{bmatrix}.$$

The squares of the characteristic frequencies are the roots of

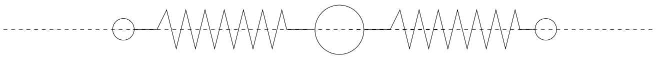
$$\begin{vmatrix} \lambda m - k & k \\ k & \lambda m - k \end{vmatrix} = 0.$$

These roots are $\lambda = 0$ with eigenvector $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\lambda = \frac{2k}{m}$ with eigenvector $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$, so the motion of our system will take the form

$$s(t) = A\cos\left(\sqrt{\frac{2k}{m}}t + \phi\right)\begin{bmatrix} 1 \\ -1 \end{bmatrix} + \big(Bt + C\big)\begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Here A, B, C, and ϕ are undertermined constants. Physically, the first term corresponds to a "vibration" where the center of mass stays fixed and the two objects simultaneously move in and out from that center. The second term is a "translation", where the entire system moves with constant velocity. Physicists have a nice technique for ignoring the translational motion by instead working in coordinates with respect to the center of mass; we will not pursue this, but it is a nice exercise to work it out.

Example 2.2.7. For our next example consider a system with three masses and two springs:



Assume that the two objects on the ends have mass m, and the middle object has mass M. Assume the two springs are identical, with spring constant k and equilibrium length ℓ. Let s_1, s_2, and s_3 be the deviations of the three objects from their equilibrium positions (where the objects are labelled 1, 2, 3 from left to right).

The Lagrangian for this system is $L = \sum \frac{1}{2} m_i \dot{s}_i^2 - \frac{1}{2}k\big((s_2 - s_1)^2 + (s_3 - s_2)^2\big)$, so we have

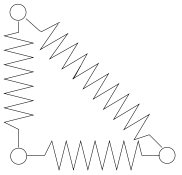
$$M = \begin{bmatrix} m & 0 & 0 \\ 0 & M & 0 \\ 0 & 0 & m \end{bmatrix}, \qquad K = k \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}.$$

Based on our previous example we can guess one of the characteristic frequences: $\lambda = 0$ corresponding to normal mode $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$. Using a little physical intuition we can also guess another normal mode: where the center object remains fixed and the two springs vibrate in sync. Here the characteristic frequency corresponds to $\lambda = \frac{k}{m}$ and the normal mode is $\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$. I don't know an easy way to guess the final characteristic frequency: doing the algebra reveals it to have

$$\lambda = \frac{k}{m}\left(1 + \frac{2m}{M}\right), \qquad \text{normal mode} = \begin{bmatrix} 1 \\ -\frac{2m}{M} \\ 1 \end{bmatrix}.$$

Note that in this vibrational mode the two outer objects move in the same direction, whereas the central object moves in the opposite direction to preserve the center of mass.

2.2.8. Problems in two-dimensional space. The previous two examples dealt with spring-mass systems distributed along a line. The same techniques allow us to deal with spring-mass systems in any number of dimensions; we will be content with analyzing spring-mass systems in the plane. For example, we might consider a system like the following:



Maybe they are three equal masses connected by springs with the same spring constant (but different equilibrium lengths), or maybe two of the masses are the same and the third is different. Physicists use such systems, for instance, to approximate vibrations of molecules.

In the examples we consider all the masses will be equal, and all the spring constants will be equal. Let m denote the mass and k denote the spring constant.

Rather than looking at one geometrical configuration at a time, it turns out its not too hard to just deal with them all at once. Assume that our objects are labelled by indices $a, b, c, \ldots$, and let let $\tilde{X}_a,\ \tilde{Y}_a$ be the positions when the system is at equilibrium. Let d_{ab} be the distance from a to b in the equilibrium state, so that

$$d_{ab} = \sqrt{(\tilde{X}_a - \tilde{X}_b)^2 + (\tilde{Y}_a - \tilde{Y}_b)^2}.$$

While the system is in motion, let $X_a = X_a(t)$ and $Y_a = Y_a(t)$ denote the position of object a. As in our one-dimensional examples, we perform the change of coordinates $s_a = X_a - \tilde{X}_a$, $t_a = Y_a - \tilde{Y}_a$.

The kinetic energy of the system is given by $\frac{1}{2}m\sum_a(\dot{s}_a^2+\dot{t}_a^2)$, while the potential energy equals

$$\sum_{a,b \text{ cnctd}} \frac{1}{2}k\big[\sqrt{(X_b - X_a)^2 + (Y_b - Y_a)^2} - d_{ab}\big]^2$$
$$= \sum_{a,b \text{ cnctd}} \frac{1}{2}k\big[\sqrt{(s_b - s_a + \tilde{X}_b - \tilde{X}_a)^2 + (t_b - t_a + \tilde{Y}_b - \tilde{Y}_a)^2} - d_{ab}\big]^2.$$

The sums here are over all pairs (a, b) that are connected by a spring—said differently, the sum is over all springs. This potential is *not* quadratic in the coordinates, and so we need to analyze its quadratic approximation. Note that our choice of coordinates guarantees that when we take a Taylor expansion of the above expression, all of the linear terms in s and t will vanish. We will make use of this to simplify our analysis.

Setting $A_{ab} = \tilde{X}_b - \tilde{X}_a,\ B_{ab} = \tilde{Y}_b - \tilde{Y}_a$, then up to linear terms the potential energy is

$$\frac{k}{2}\sum_{a,b \text{ cnctd}}\Bigg[(s_b - s_a)^2 + (t_b - t_a)^2$$
$$- 2d_{ab}\sqrt{A_{ab}^2 + 2A_{ab}(s_b - s_a) + (s_b - s_a)^2 + B_{ab}^2 + 2B_{ab}(t_b - t_a) + (t_b - t_a)^2}\Bigg].$$

Since $\sqrt{a+x} \approx \sqrt{a} + \frac{x}{2\sqrt{a}} - \frac{x^2}{8a\sqrt{a}}$, then up to second order, this is equal to

$$\frac{k}{2}\sum_{a,b \text{ cnctd}}\Bigg[(s_b - s_a)^2 + (t_b - t_a)^2$$
$$- 2d_{ab}\left(d_{ab} + \frac{(s_b - s_a)^2 + (t_b - t_a)^2}{2d_{ab}} - \frac{\big(4A_{ab}(s_b - s_a) + 4B_{ab}(t_b - t_a)\big)^2}{8(d_{ab})^3}\right)\Bigg].$$

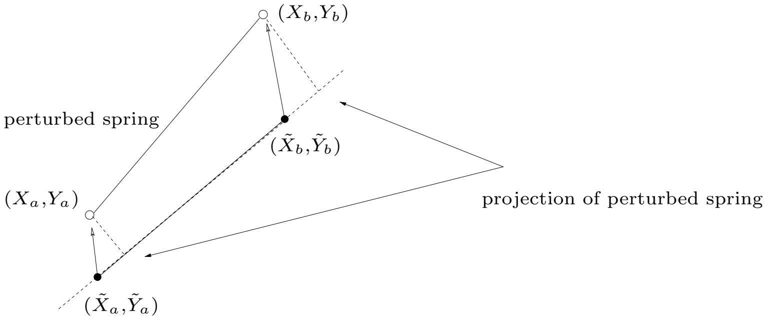
Thus up to second order the potential energy is therefore

$$\frac{k}{2}\sum_{a,b \text{ cnctd}} \frac{1}{(d_{ab})^2}\big[A_{ab}(s_b - s_a) + B_{ab}(t_b - t_a)\big]^2.$$

So we have obtained

$$(2.2.9)\quad L \approx \sum_a \frac{1}{2}m(\dot{s}_a^2 + \dot{t}_a^2) - \frac{k}{2}\sum_{a,b\ cnctd}\left(\frac{A_{ab}(s_b - s_a) + B_{ab}(t_b - t_a)}{d_{ab}}\right)^2.$$

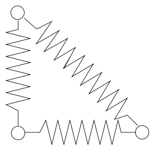
Remark 2.2.10. Formula (2.2.9) admits a nice interpretation. Note that $\frac{1}{d_{ab}}[A_{ab}, B_{ab}]$ is simply the unit vector point from object a to object b in their equilibrium positions. In our potential energy term, the quantity inside the square is simply the dot product of this unit vector with $[s_b, t_b] - [s_a, t_a]$. As shown in the following picture, this is simply the length of the projection of the perturbed spring onto the line determined by its equilibrium position.


(X_b,Y_b)
perturbed spring
$(\tilde{X}_b,\tilde{Y}_b)$
(X_a,Y_a)
projection of perturbed spring
$(\tilde{X}_a,\tilde{Y}_a)$

In other words, up to quadratic approximation we can assume that each spring is only being expanded or contracted along its original axis. Unfortunately I don't know an *a priori* explanation of this, it is only an observation coming out of the above calculation. I am grateful to Kevin Donahue for pointing this out to me.

Remark 2.2.11. Notice that it is clear how one generalizes formula (2.2.9) to spring-mass systems in three dimensions, or in any number of dimensions.

Example 2.2.12. Consider again three objects of the same mass m, tied together by three springs in the following configuration:



Assume that the springs all have the same spring constant k, and that the equilibrium positions of the three objects are $(0,0)$, $(0,1)$, and $(1,0)$. In this order label the objects by 1, 2, and 3, and let s_i, t_i give the deviation of object i from its equilibrium position.

According to (2.2.9) the Lagrangian for this system is

$$L = \frac{m}{2}\sum_i(\dot{s}_i^2 + \dot{t}_i^2) - \frac{k}{2}\Big((t_2 - t_1)^2 + (s_3 - s_1)^2 + \frac{1}{2}((s_3 - s_2) - (t_3 - t_2))^2\Big).$$

Our matrices are therefore $M = mI$ and $K = kJ$ where

$$J = \begin{bmatrix} 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ 0 & -1 & \frac{1}{2} & \frac{3}{2} & \frac{1}{2} & -\frac{1}{2} \\ -1 & 0 & -\frac{1}{2} & \frac{1}{2} & \frac{3}{2} & -\frac{1}{2} \\ 0 & 0 & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

The roots of $\det(\lambda M - K) = 0$ are products of $\frac{k}{m}$ with the eigenvalues of J. Feeding this matrix into a computer, one finds that the possible λ's are 0 (with multiplicity three), $\frac{k}{m}$, $\frac{2k}{m}$, and $\frac{3k}{m}$. The corresponding normal modes (the eigenvectors of J) are

$$\begin{bmatrix}1\\0\\1\\0\\1\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\1\\0\\1\end{bmatrix}, \begin{bmatrix}0\\0\\1\\0\\0\\-1\end{bmatrix} (\lambda = 0) \quad \begin{bmatrix}-1\\-1\\1\\0\\0\\1\end{bmatrix} (\lambda = \tfrac{k}{m}), \quad \begin{bmatrix}-1\\1\\0\\-1\\1\\0\end{bmatrix} (\lambda = \tfrac{2k}{m}), \quad \begin{bmatrix}1\\1\\1\\-2\\-2\\1\end{bmatrix} (\lambda = \tfrac{3k}{m}).$$

Let us discuss what all this means geometrically. Based on our previous examples the reader would probably have guessed the $\lambda = 0$ eigenvalue, together with the first two of its normal modes: these correspond to the uniform motion of the system in the x-direction or in the y-direction. But why is there a third normal mode associated to $\lambda = 0$? The reason turns out to be that the group of affine transformation of the plane is 3-dimensional, not 2-dimensional. Recall that the group of affine transformations (or "rigid motions") is a semi-direct product of $SO(2)$ and $\mathbb{R}^2$, corresponding to the subgroups of rotations and translations. The translations are things we have already accounted for, but we also need to account for uniform rotational motion of the system. This is the third normal mode for $\lambda = 0$. We claim that the vector $[0, 0, 1, 0, 0, -1]$ corresponds to an "infinitesimal clockwise rotation" about the point $(0, 0)$. To see this, let R_θ be the clockwise rotation of the plane, centered at the origin, through θ radians. The initial equilibrium configuration has objects at $(0, 0)$, $(0, 1)$, and $(1, 0)$. Applying R_θ puts the three objects at $(0, 0)$, $(\sin\theta, \cos\theta)$, and $(\cos\theta, -\sin\theta)$. Converting this new configuration into s, t-coordinates gives

$$[0, 0, \sin\theta, \cos\theta - 1, \cos\theta - 1, -\sin\theta].$$

We can regard this as a path in θ and take the derivative when $\theta = 0$: this gives the vector

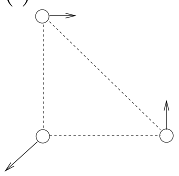
$$[0, 0, 1, 0, 0, -1].$$

It is in this sense that the vector represents an infinitesimal clockwise rotation about $(0, 0)$.

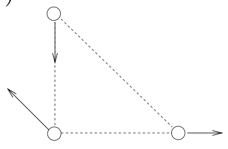
What if we had performed a rotation about another point, rather than the origin? The answer is that we would get some linear combination of our three eigenvectors for $\lambda = 0$. The reader is encouraged to think through this, or to give it a try.

Now let us turn to the other three eigenvalues and their normal modes. These correspond to true "vibrations" of the system, and we can interpret them geometrically as in the following pictures:

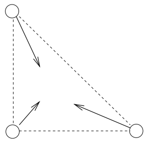
(i)



(ii)



(iii)



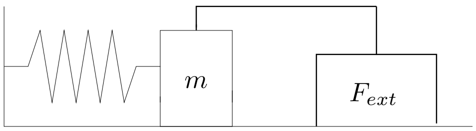
Picture (i) depicts the normal mode for $\lambda = \frac{k}{m}$, the arrows indicating one direction of vibration. All the objects move simultaneously in the direction of the arrows, and then they stop and move in the opposite directions, back and forth forever. This mode and the mode in (iii) (for $\lambda = \frac{3k}{m}$) are the easiest to have an intuitive feeling for. The mode in (ii) is less intuitive for most people—but it is a valid mode nevertheless.

The reader is encouraged to work out the normal modes and characteristic frequencies for some other spring-mass systems.

2.2.13. Concluding remarks about quadratic Lagrangians. One important thing to remember about the case of quadratic Lagrangians is that the associated Euler-Lagrange equations are linear (they look like $\ddot{q}_i = k_i q_i$). Consequently, the solutions form a vector space: different types of motion can combine in a very simple way (i.e., they add). This is usually *not* true for non-quadratic Lagrangians.

For our second remark, suppose we have a spring-mass system and we jiggle one of the masses. For example, maybe we have a large rectangular lattice of masses tied together by springs—something like a two-dimensional mattress. If we go jump up and down at one point of the mattress, we are jiggling one of the masses. What happens? Physical intuition tells us that the jiggling propogates outward to the other masses in some way. How do we analyze this mathematically? We will not go into detail here, but we want to set up the method for solution.

To get us started, let us return to the simplest example of a system with a single mass and a single spring. But this time let us add an external force (the "jiggling"). We picture this as follows:



Here the box on the right labelled F_{ext} is some machine that exerts an external force $F_{ext}(t)$ (varying with time) on the object. We can analyze this physical system with Newton's Second Law, which yields

$$m\ddot{x} = F_{total} = -kx + F_{ext} = -\frac{\partial}{\partial x}\left(\frac{k}{2}x^2 - F_{ext}(t)x\right).$$

So we can treat the external force as giving a new term equal to $-F_{ext}(t)x$ in the potential energy. Our Lagrangian therefore becomes

$$L(x, \dot{x}) = \frac{1}{2}m\dot{x}^2 - \frac{k}{2}x^2 + F_{ext}(t)x.$$

Based on this simple example we can now generalize. Imagine that we have a two-dimensional spring-mass system, with objects indexed by $a, b, c, \ldots$. As usual we use the coordinates $s_a, t_a, s_b, t_b, \ldots$ giving the deviations from the equilibrium positions. We have previously learned how to write down the Lagrangian for such a system. Suppose now that we add an external force, acting on the mass m_a. The lesson of the previous example is that the Lagrangian for this new system will be

$$L_{new} = L_{old} + F_{horiz}(t)s_a + F_{vert}(t)t_a$$

where F_{horiz} and F_{vert} are the two components of the external force. We leave it to the reader to verify this claim. If we have several external forces acting on different masses $m_a, m_b, m_c, \dots$ then that just amounts to adding more analogous terms to the Lagrangian.

2.2.14. A first look at field theory.

Consider an infinite string of masses connected by springs, stretched out along a line:

Let us assume that all the objects have the same mass m, and that all the springs have the same spring constant k and the same equilibrium length ℓ. Let s_i denote the deviation of the ith object from its equilibrium position. Then as in our previous examples, we have

$$L = \frac{1}{2}m\sum_i \dot{s}_i^2 - \frac{k}{2}\sum_i (s_{i+1}-s_i)^2 = \frac{1}{2}m\sum_i \dot{s}_i^2 - \frac{k\ell^2}{2}\sum_i \left(\frac{s_{i+1}-s_i}{\ell}\right)^2.$$

Now imagine passing to the continuum limit, where $\ell \to 0$. In this limit the spring-mass system becomes a flexible rod, with a uniform density σ and a "tensile strength" ????. In the discrete system we had a displacement s_i corresponding to the object originally at position ℓi. In the continuum limit we have a displacement s_x for every $x \in \mathbb{R}$ (corresponding to the point of the rod originally at location x).

In the discrete system the s_i's are actually functions of time $s_i(t)$, and so in the limit we will have s_x also depending on time.

Let us change notation a bit, and write $\varphi(x) = s_x$ and $\varphi(x,t) = s_x(t)$. A "configuration" of the rod system is a (continuous, or perhaps smooth) map $\varphi\colon \mathbb{R} \to \mathbb{R}$, and if we allow it to change over time then it can be regarded as a map $\mathbb{R}^2 \to \mathbb{R}$.

The expression $\dot{s_i}$ in our discrete system becomes $\frac{\partial\varphi}{\partial t}$ in the continuum version, and the expression $\frac{s_{i+1}-s_i}{\ell}$ becomes $\frac{\partial\varphi}{\partial x}$. The Lagrangian becomes

$$\begin{aligned} L(\varphi,\dot{\varphi}) &= \frac{1}{2}\int \left(\frac{m}{dx}\right)\left(\frac{\partial\varphi}{\partial t}\right)^2 dx - \frac{1}{2}\int \left(\frac{k\ell^2}{dx}\right)\left(\frac{\partial\varphi}{\partial x}\right)^2 dx \\ &= \frac{1}{2}\int \sigma\left(\frac{\partial\varphi}{\partial t}\right)^2 dx - \frac{1}{2}\int \rho\left(\frac{\partial\varphi}{\partial x}\right)^2 dx \end{aligned}$$

where σ is density and ρ has units of force. We can also write

$$L(\varphi,\dot{\varphi}) = \frac{\sigma}{2}\int \left[\left(\frac{\partial\varphi}{\partial t}\right)^2 - \left(\frac{\rho}{\sigma}\right)\left(\frac{\partial\varphi}{\partial x}\right)^2\right] dx.$$

The units of $\dfrac{\rho}{\sigma}$ are $\dfrac{\frac{kg\ m}{s^2}}{\frac{kg}{m}} = \dfrac{m^2}{s^2}$, which suggests the square of a velocity.

Two Dimensional Waves

[Picture]

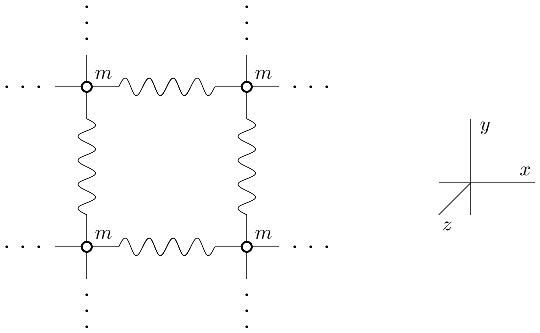
$a \rightsquigarrow \underline{x} \in \mathbb{R}^2 \quad (s_a, t_a) \rightsquigarrow (s_x, t_x) = \varphi(\underline{x}) \in \mathbb{R}^2 \quad \varphi : \mathbb{R}^2 \to \mathbb{R}^2$

$$L = \sum_i \frac{1}{2} m\big(\dot{s}_i^2 + \dot{t}_i^2\big) - \frac{k}{2} \sum_{\text{springs}} \left[\frac{A_{ab}(s_b - s_a) + B_{ab}(t_b - t_a)}{d_{ab}} \right]^2$$

$$L(\varphi, \dot{\varphi}) = \frac{1}{2} \int \sigma \left(\left(\frac{\partial \varphi_1}{\partial t} \right)^2 + \left(\frac{\partial \varphi_2}{\partial t} \right) \right) dxdy - \frac{1}{2} \int \rho \left(\left(\frac{\partial \varphi_1}{\partial x} \right)^2 + \left(\frac{\partial \varphi_2}{\partial x} \right) \right) dxdy$$

<u>Similar Thing for Fields $\varphi : \mathbb{R}^2 \to \mathbb{R}$</u>

Consider an infinite 2-dimensional lattice of springs in $\mathbb{R}^3$ with each mass constrained to move only perpendicular to the plane. Specifically, let s_a, t_a, and u_a be the displacements from rest of the mass a in the x-, y-, and z-directions, respectively, with constraint $s_a = t_a = 0$.



The Lagrangian for this system is

$$L = \sum_a \frac{1}{2} m \dot{u}_a^2 - \frac{k}{2} \sum_{\substack{\text{springs}\\ a \to b}} (u_b - u_a)^2.$$

When we translate to field theory, the indices a become vectors $\mathbf{x} \in \mathbb{R}^2$, and the coordinates u_a become outputs $\varphi(\mathbf{x})$ of a function $\varphi\colon \mathbb{R}^2 \to \mathbb{R}$. In the Lagrangian, we rewrite the potential term as

$$\frac{k\ell^2}{2} \sum_{\substack{\text{springs}\\ a \to b}} \Big(\frac{u_b - u_a}{\ell} \Big)^2$$

so that in the limit as $\ell \to 0$, the Lagrangian becomes

$$L\big(\varphi, \dot{\varphi}\big) = \int_{\mathbb{R}^2} \frac{1}{2} \Big(\cancelto{\sigma}{\frac{m}{dx\,dy}} \Big) \Big(\frac{\partial \varphi}{\partial t} \Big) dx\,dy - \int \frac{\cancelto{\rho}{k\ell^2}}{2} \Bigg[\underbrace{\Big(\frac{\partial \varphi}{\partial x} \Big)^2}_{\text{horiz. springs}} + \underbrace{\Big(\frac{\partial \varphi}{\partial y} \Big)^2}_{\text{vert. springs}} \Bigg]$$

Remark 2.2.15. To add a ”source” term we earlier added $F_i(t)s_i$ or $F_i(t)t_i$ to the Lagrangian. In the field theory this corresponds to adding a term $F(x,t)\varphi(x,t)$, or just $F(x)\varphi(x)$ if we suppress the t dependence as usual.

2.3. Variational approach to the Euler-Lagrange equations

In this section we describe a different perspective on classical mechanics, via the so-called “principle of least action”. The idea is that one considers *all* possible paths that a mechanical system might take as it develops, and to each path one associates a numeric quantity called the *action*. The action does not exactly have a physical interpretation, as it is a quantity one associates to paths that do not actually occur in the real world. I think of the action as something like the "cost to the universe" of going down that path. The principle of least action says that the path the universe actually takes—the one that satisfies the Euler-Lagrange equations—is an *extreme point* for the action, in the sense of the calculus of variations. This means that for all small deviations of our path, the first-order correction to the action is equal to zero.

2.3.1. The basic argument.

M is a smooth manifold and $L\colon TM \to \mathbb{R}$ is a smooth map. Let $a, b \in M$ and let $\gamma\colon I \to M$ be a smooth path from a to b. To such a path we associate a real number $S(\gamma)$ called the "action", defined as

$$S(\gamma) = \int_0^1 L\big(\gamma(t), \dot{\gamma}(t)\big)dt.$$

If we have local coordinates $q_1, \dots, q_n$ on M, and we write $\gamma(t) = (q_1(t), \dots, q_n(t))$, then we will write

$$S(\gamma) = \int_0^1 L(q_i, \dot{q}_i)dt.$$

Now suppose that $\tilde{\gamma}$ is a small variation of γ that has the same endpoints a and b. The path $\tilde{\gamma}$ is given in local coordinates by $\tilde{\gamma}(t) = (\tilde{q_1}(t), \dots, \tilde{q_n}(t))$, and we can write

$$\tilde{q_i}(t) = q_i(t) + \delta q_i(t).$$

We then compute that

$$\begin{aligned} S(\tilde{\gamma}) &= \int_0^1 L(q_i + \delta q_i, \dot{q}_i + \dot{\delta} q_i)dt \\ &\approx \int_0^1 \left(L(q_i, \dot{q}_i) + \frac{\partial L}{\partial q_i}\delta q_i + \frac{\partial L}{\partial \dot{q}_i}\delta \dot{q}_i \right) dt \\ &= S(\gamma) + \int_0^1 \left(\frac{\partial L}{\partial q_i}\delta q_i \right) dt + \frac{\partial L}{\partial \dot{q}_i}\delta q_i \Big]_0^1 - \int_0^1 \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i} \right) \delta q_i dt \\ &= S(\gamma) + \int_0^1 \left(\frac{\partial L}{\partial q_i} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i} \right) \right) \delta q_i dt. \end{aligned}$$

In the third line we have used integration by parts, and in the fourth line we have used that $\delta q_i(1) = \delta q_i(0) = 0$ for all i.

We define

$$\delta S = \int_0^1 \left(\frac{\partial L}{\partial q_i} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i} \right) \right) \delta q_i dt$$

and we think of this as the "first-order variation" in the action. Imagine that γ is a path with the property that this term is zero for all choices of δq_i's. This is equivalent to saying that $\frac{\partial L}{\partial q_i} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i} \right) = 0$ for all choices of i. To see this, just note that if one of these differences were nonzero then choosing δq_i as an appropriate bump function and all other δq_j's to be zero, would give a nonzero δS.

So the moral of our story is as follows:

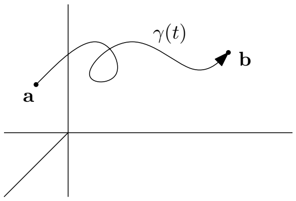
γ satisfies the Euler-Lagrange equations iff γ is an extreme point for S.

Now, on the face of it this is not a very deep observation. In any practical problem the important point is the Euler-Lagrange equations, and this variational approach amounts only to a clever way of encoding them. But in the long run this turns out to be a very good idea. The variational approach makes certain theoretical ideas clearer (we will see an example below in the case of conserved quantities), and more importantly it seems to adapt to a wide variety of different situations. Time and time again in physics one ends up analyzing a situation by writing down some kind of "action" and looking for the critical points.

There is one final remark that is important to make at this stage. In a classical mechanical system, the idea is that only one path is possible given the initial conditions—and the Euler-Lagrange equations are a way of specifying which path that is. In quantum mechanics the situation is very different: in some sense *all* paths are possible, and one can really only ever talk about probabilities of a certain path occurring. It will turn out that the action S becomes more fundamental here, as it is directly related to these probabilities. The is the basis of Feynman path integrals, which we will study further in Sections ???? below. But the short summary is that in classical mechanics it is only the extreme points of the action that are important, while in quantum mechanics it is really the action itself that is fundamental.

2.3.2. Examples from the calculus of variations. The variational arguments we saw above are part of a vast subject called the "calculus of variations", which eventually evolved into modern day functional analysis. Let us look at a few basic examples from the beginnings of the subject.

Example 2.3.3. Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^3$. Find the path of shortest length from $\mathbf{a}$ to $\mathbf{b}$.



If $\gamma(t) = \big(q_1(t), q_2(t), q_3(t)\big)$, then the length of γ is

$$S(\gamma) = \int_0^1 \underbrace{\sqrt{\dot{q}_1(t)^2 + \dot{q}_2(t)^2 + \dot{q}_3(t)^2}}_{\text{the Lagrangian } L(\mathbf{q},\dot{\mathbf{q}})}\, dt$$

It is reasonable to expect that a path of shortest length will be an extreme point for S. But an extreme point for S satisfies the Euler-Lagrange equations $\partial L/\partial q_i = \frac{d}{dt}(\partial L/\partial \dot{q}_i)$. Since L doesn't depend on q_i the left side is zero, giving

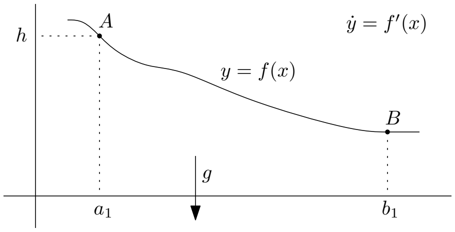
$$0 = \frac{d}{dt}\left(\frac{\cancel{1}}{\cancel{2}} \cdot \frac{1}{|\dot{\mathbf{q}}|} \cdot \cancel{2}\dot{q}_i(t)\right)$$

This says that for each value of i, the quantity $q_i(t)/|\dot{\mathbf{q}(\mathbf{t})}|$ is independent of t. Call this constant C_i, and let $\mathbf{C}$ be the vector with components C_i. Then

$$\dot{\mathbf{q}}(t) = |\dot{\mathbf{q}}(t)| \cdot \mathbf{C},$$

which implies that the velocity of the path γ is always parallel to $\mathbf{C}$, so γ travels along a straight line, as expected. Note that this does not distinguish among the different paths along this line (due to changes in speed), except that the speed can never be 0 and so there can be no change of direction.

Example 2.3.4 (Brachistochrone Problem). This problem was first posed by Johann Bernoulli. Let $A, B \in \mathbb{R}^2$, as shown below. Given a curve γ from A to B, suppose that an object at A, starting at rest, is constrained to move along γ under a constant gravity force with no friction. Find the curve that minimizes the time elapsed until the object reaches B.



The law of conservation of energy gives us a nice formula for the velocity at any point on the curve. The initial energy is purely potential, mgh. At other values of x, the energy is $\frac{1}{2}m \cdot v(x)^2 + mg \cdot f(x)$. Thus

$$\cancel{m}gh = \frac{1}{2}\cancel{m} \cdot v(x)^2 + \cancel{m}g \cdot f(x) \qquad \text{so} \qquad v(x) = \sqrt{2g\big(h - f(x)\big)}\,.$$

We need an expression for the time it takes the object to traverse the path. An infinitesimal piece of arclength near (x, y) is

$$\sqrt{(dx)^2 + (dy)^2} = \sqrt{1 + f'(x)^2} \cdot dx,$$

and the time it takes the object to traverse this length is therefore

$$\frac{\sqrt{1 + f'(x)^2} \cdot dx}{v(x)}.$$

So the total time for the object to move along the path is given by

$$S(f) = \int_{a_1}^{b_1} \frac{\sqrt{1 + f'(x)^2}}{\sqrt{2g\big(h - f(x)\big)}}\, dx.$$

We can think of this as the action for the path $y = y(x)$ corresponding to the Lagrangian

$$L(y, \dot{y}) = \frac{\sqrt{1 + \dot{y}^2}}{\sqrt{2g(h - y)}}\,.$$

This is the setup. The Euler-Lagrange equation is a bit unpleasant to solve, but of course people know how to do it. For the interested reader we give some hints about this in Exercise 2.3.5 below.

Exercise 2.3.5. First notice that the analysis simplifies somewhat by taking $y = h$ to be the origin of the y-axis with positive direction pointing downward (parallel to g), everywhere replacing h with 0, $-f(x)$ with $f(x)$, and $-y$ with y. This gives the Lagrangian

$$L(y, \dot{y}) = \frac{\sqrt{1 + \dot{y}^2}}{\sqrt{2gy}}.$$

In general, the Euler-Lagrange equations imply that $L - \dot{y}\frac{\partial L}{\partial \dot{y}}$ is a constant of motion. This is called the Beltrami Identity, and it is easy to prove by just taking the time derivative and using the chain rule for $\frac{\partial L}{\partial t}$.

Setting $L - \dot{y}\frac{\partial L}{\partial \dot{y}} = C$ and computing the partial derivative, one quickly gets (after rearranging) that

$$\dot{y}^2 = \frac{1}{2C^2gy} - 1 = \frac{(\frac{1}{2C^2g}) - y}{y}.$$

It turns out this is known to be the equation for a cycloid, with a rolling circle of radius $R = \frac{1}{4C^2g}$. It has parametric solution given by

$$x = R(\theta - \sin\theta), \qquad y = R(1 - \cos\theta),$$

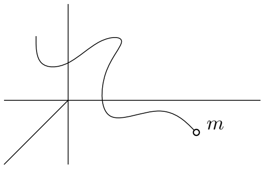
as one may readily check.

2.3.6. Connections between group actions and constants of motion. Noether's Theorem says that for every symmetry of the action there is a corresponding constant of motion. One could even say that any *infinitesimal* symmetry yields a constant of motion. We have seen examples of this already: the Euler-Lagrange equations show that if L is independent of one of the variables q_i, then $\frac{\partial L}{\partial \dot{q}_i}$ is a constant of motion. But being independent of q_i means L is invariant under changes $q_i \mapsto q_i + \delta q_i$, and this is our "infinitesimal symmetry".

We will state the version of Noether's Theorem for global symmetries, or group actions. Suppose we have a configuration space M (a smooth manifold) and the Lagrangian $L\colon TM \to \mathbb{R}$. Suppose we have a group action of a 1-dimensional Lie group G (which will be either $\mathbb{R}$ or S^1) on M, and suppose further that the group action preserves the physical action, i.e.

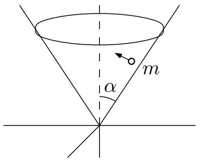
$$S(g \cdot \gamma) = S(\gamma) \quad \text{for all } g \in G \text{ and all paths } \gamma.$$

EXAMPLE 2.3.7. Consider a free particle in $\mathbb{R}^3$ (so there is no potential).



The Lagrangian is $L = \frac{1}{2}m(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)$, and this is independent of x, y, and z. So if we fix any vector $v \in \mathbb{R}^3$ and define an $\mathbb{R}$-action on $\mathbb{R}^3$ by translation in the v-direction, we have $S(g \cdot \gamma) = S(\gamma)$ for any $g \in \mathbb{R}$.

EXAMPLE 2.3.8. Consider the familiar cone example.

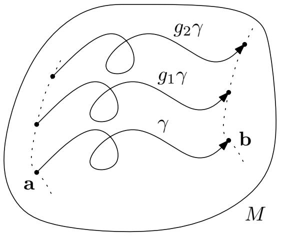


The Lagrangian is $L = \frac{1}{2}m\big((1 + \cot^2\alpha)\dot{r}^2 + r^2\dot{\theta}^2\big) - mgr\cot\alpha$, which is independent of θ. In this example the physical action is preserved by rotation about the vertical axis, which is an S^1-action.

Theorem 2.3.9 (Noether's Theorem). *If G is a one-dimensional Lie group acting on M in a way that preserves S, then there is an associated constant of motion for the mechanical system. (See the proof for a prescription for obtaining this constant of motion).*

Remark 2.3.10. The reader might be wondering why we are only looking at actions of $\mathbb{R}$ and S^1, rather than arbitrary Lie groups. This is only a matter of pedagogical convenience. If we have an action by an arbitrary Lie group G, then the infinitesimal symmetries are determined by a small neighborhood of the identity. Every element of such a neighborhood lies in a 1-parameter subgroup (obtained by exponentiating an appropriate line inside the Lie algebra), and this brings us back down to the 1-dimensional situation. So perhaps the best statement is that every *line* inside the Lie algebra of G gives rise to an associated constant of motion.

Proof. Let γ be a path in M satisfying the Euler-Lagrange equations. Consider the action of G on γ, depicted in the following diagram:



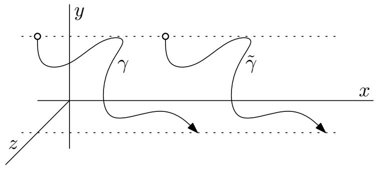
Set $\tilde{\gamma} = g\gamma$ for some "small" g, and write $\tilde{\gamma} = \gamma + \delta\gamma$ in M. (Note that $\delta\gamma$ will typically be nonzero at the endpoints, in contrast to a similar-looking previous example). Then

$$\begin{aligned} S(\tilde{\gamma}) &= S(\gamma + \delta\gamma) \\ &= \int_0^1 L\big(q_i(t) + \delta q_i(t), \dot{q}_i(t) + \delta\dot{q}_i(t)\big)\,dt \\ &\approx \int_0^1 \Big(L(q_i,\dot{q}_i) + \frac{\partial L}{\partial q_i}\delta q_i + \frac{\partial L}{\partial \dot{q}_i}\delta\dot{q}_i\Big)\,dt \qquad \text{(linear approximation)} \\ &= S(\gamma) + 0 + \Bigg[\underbrace{\cancel{\int_0^1 \Big[\Big(\frac{\partial L}{\partial q_i}\Big) - \frac{d}{dt}\Big(\frac{\partial L}{\partial \dot{q}_i}\Big)\Big]\delta q_i dt}}_{\text{0 by E-L equation for } \gamma} + \frac{\partial L}{\partial \dot{q}_i}\delta q_i\Bigg]_0^1 \end{aligned}$$

The sum over i is implied, and the last step is integration by parts. On the other hand, since $\tilde{\gamma}$ is obtained from γ by the Lie group action, $S(\tilde{\gamma}) = S(g\gamma) = S(\gamma)$. Therefore the last term above is zero, so $\Sigma_i(\partial L/\partial\dot{q}_i)\delta q_i$ is a constant of motion. □

Remark 2.3.11. Note that the above proof shows clearly that we did not really need a global action of G on M, only an "infinitesimal" action.

Example 2.3.12. In the example of the free particle, we may consider the $\mathbb{R}$-action which is translation of the path γ in the x-direction.



In this case, $\delta y = \delta z = 0$ and we can take $\delta x = \epsilon$ to be anything we like. So the constant of motion predicted by Noether's Theorem is the single term $(\partial L/\partial \dot{x})\epsilon$. Since the ϵ was arbitrary, we find that $\frac{\partial L}{\partial \dot{x}}$ is a constant of motion. Recalling that $L = \frac{1}{2}m(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)$, we have $\partial L/\partial \dot{x} = m\dot{x}$. The symmetry has shown us that the momentum $m\dot{x}$ in the x-direction is conserved.

Example 2.3.13. Similarly, in the cone problem the S^1-action shows that angular momentum about the z-axis is a constant of motion.

Based on the above two examples, Noether's Theorem does not seem very impressive—we already understood the constants of motion in terms of the Lagrangian being independent of one coordinate. This was because in these examples we had carefully chosen the "right" coordinates to begin with. The importance of Noether's Theorem is really in its coordinate-independence—it tells us where to look for constants of motion just based on the geometry of the mechanical system. And perhaps this demonstrates the importance of the variational approach overall: it gives an approach to mechanics which completely avoids any mention of coordinates.

2.4. Hamiltonian mechanics

The two basic paradigms in classical mechanics are the Lagrangian formalism and the Hamiltonian formalism. The essential difference between the two is something you have probably already seen in an undergraduate differential equations course. Given a second-order differential equation in one varable like $y'' + p(t)y' + q(t) = 0$, there is a silly trick for replacing it by two first-order equations in two variables: one sets $z = y'$ and then has the pair of equations

$$z = y', \quad z' + p(z)z + q(t) = 0.$$

Hamiltonian mechanics is really just a variation on this basic trick. It replaces the second-order differential equations one gets from the Euler-Lagrange equations with pairs of first-order equations, at the expense of introducing a second set of variables. The payoff turns out to be a closer connection with geometry: the solutions to first-order equations are flows of vector fields. The geometry of this situation—called *symplectic geometry* in modern times—is a quite extensive field.

In this section we give a very brief introduction to Hamiltonian mechanics, really just outlining the basic setup. It would be possible to spend quite a bit more time on this subject, but it is not really needed for our present purposes. We will get by with just the basics.

2.4.1. Generalized momenta. Consider a smooth manifold M with a Lagrangian $L\colon TM \to \mathbb{R}$. Let $q_1, \ldots, q_n$ be local coordinates on M, so that we can regard L as a function $L(q_i, \dot{q}_i)$.

Definition 2.4.2. *Let $p_i = \frac{\partial L}{\partial \dot{q}_i}$, and call this the "(generalized) momentum conjugate to q_i."*

Example 2.4.3. For a free particle in $\mathbb{R}^3$ recall that we have $L = \frac{1}{2}m\big(\dot{x}^2 + \dot{y}^2 + \dot{z}^2\big)$. Then $p_x = \frac{\partial L}{\partial \dot{x}} = m\dot{x}$, and similarly for p_y, p_z. So in this simple case the generalized momenta coincide with the notions of momenta one learns about in a freshman physics class.

Example 2.4.4. Consider once again our problem of the particle moving on the cone, where the coordinates are r and θ. Recall that the Lagrangian is

$$L = \frac{1}{2}m\big((1+\cot^2\alpha)\dot{r}^2 + r^2\dot{\theta}^2\big) - mgr\cot\alpha.$$

Then one finds

$$p_r = m(1+\cot^2\alpha)\dot{r} \quad \text{and} \quad p_\theta = mr^2\dot{\theta}.$$

You might recognize the formula for p_θ as just the angular momentum of the object about the z-axis. For p_r, by using that $z = (\cot\alpha)r$ one finds that p_r is a constant multiple of the linear momentum in the z-direction. (Precisely, $p_r = \frac{2}{\sin(2\alpha)}m\dot{z}$, but we will have no need of this equation.)

Remark 2.4.5 (Conservation of Momentum). Recall the Euler-Lagrange equations $\dfrac{d}{dt}\left(\dfrac{\partial L}{\partial \dot{q}_i}\right) = \dfrac{\partial L}{\partial q_i}$. From these we immediately see that if L is independent of q_i then p_i is a conserved quantity for the motion. One sees instances of this in the two examples above: all three linear momenta in the case of a free particle, and the angular momentum about the z-axis in the cone problem.

2.4.6. The Hamiltonian. Now we introduce the expression

$$H = \sum_i \left[p_i\dot{q}_i - L(q_i, \dot{q}_i)\right]$$

called the **Hamiltonian** of our system. For the moment this is really coming out of nowhere, although we will try to give a little more motivation in Remark 2.4.11 below. For now let us just remark that if $q_1, \ldots, q_n$ are Cartesian coordinates on $\mathbb{R}^n$ and we are dealing with a free particle of mass m, then $p_i = m\dot{q}_i$ and $\sum_i p_i\dot{q}_i$ is therefore twice the kinetic energy. Since the Lagrangian is K.E. − P.E., the Hamiltonian is therefore K.E. + P.E.. That is, in this case the Hamiltonian is just the total energy of the system. It turns out the Hamiltonian is almost *always* the total energy.

Coming back to our definition of H, notice that this expression is a function of q_i, $\dot{q}_i$, and p_i. The momemtum p_i is itself a function of the q's and $\dot{q}$'s, and in almost all examples this function can be inverted to solve for $\dot{q}_i$ in terms of the p's and q's. We will assume that this can be done, in which case we regard $H = H(q_i, p_i)$.

Let us now compute the infinitesimal variation of H:

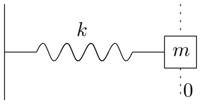
$$dH = (dp_i)\dot{q}_i + \cancel{p_i(d\dot{q}_i)} - \Big(\frac{\partial L}{\partial q_i}dq_i + \cancel{\frac{\partial L}{\partial \dot{q}_i}d\dot{q}_i}\Big).$$

This gives us $\frac{\partial H}{\partial p_i} = \dot{q}_i$ and $\frac{\partial H}{\partial q_i} = -\frac{\partial L}{\partial q_i}$. If we use the Euler-Lagrange equations to rewrite the latter, we get **Hamilton's Equations of Motion**

$$\frac{\partial H}{\partial p_i} = \dot{q}_i \qquad \text{and} \qquad \frac{\partial H}{\partial q_i} = -\frac{\partial L}{\partial q_i} \overset{\text{E-L}}{=} -\frac{d}{dt}\Big(\frac{\partial L}{\partial \dot{q}_i}\Big) = -\dot{p}_i.$$

Note that for each i these are two first-order differential equations, rather than the single second-order equation given by Euler-Lagrange.

Example 2.4.7. Consider the basic spring problem where $x = x(t)$ is the position at time t.



The Lagrangian is $L(x, \dot{x}) = \frac{1}{2}m\dot{x}^2 - \frac{k}{2}x^2$, so

$$p_x = \frac{\partial L}{\partial \dot{x}} = m\dot{x}$$

and we may rewrite the total energy $H = \frac{1}{2}m\dot{x}^2 + \frac{k}{2}x^2$ as $H(x,p) = \frac{p_x^2}{2m} + \frac{k}{2}x^2$, which eliminates the $\dot{x}$. Then Hamilton's equations of motion $\partial H/\partial p = \dot{x}$ and $\partial H/\partial x = -\dot{p}$ in this example become $p/m = \dot{x}$ and $kx = -(m\dot{x})' = -m\ddot{x}$.

From a pragmatic point of view, it may seem to the reader that nothing really useful is going on here. In any specific example, Hamilton's equations of motion are not going to be any easier to solve than the Euler-Lagrange equations—after all, they are equivalent. While this is true, it nevertheless turns out that something has been gained here. By analyzing mechanical systems in turns of the q_i's and p_j's, rather than the q_i's and $\dot{q}_i$'s, one ends with a theory where there are more tools. And having more tools means you can solve more problems.

2.4.8. The general theory of Hamiltonian mechanics. From a mathematical point of view, $q_1, \ldots, q_n$ and $\dot{q}_1, \ldots, \dot{q}_n$ are coordinates on the tangent bundle TM. But $q_1, \ldots, q_n$ and $p_1, \ldots, p_n$ are mostly naturally thought of as coordinates on the *cotangent* bundle T^*M. We will explain this more in Section ???? below, but it comes down to noticing how these quantities behave under a change of coordinates. If $u_1, \ldots, u_n$ is another set of coordinates on M, we can write $q_i = q_i(u_1, \ldots, u_n)$. Then

$$\dot{q}_i = \frac{\partial q_i}{\partial u_j}\dot{u}_j. \tag{2.4.9}$$

Now write $p_i^{(q)}$ and $p_i^{(u)}$ for the generalized momenta conjugate to q_i and u_i. One has

$$p_i^{(u)} = \frac{\partial L}{\partial \dot{u}_i} = \frac{\partial L}{\partial q_j}\cancel{\frac{\partial q_j}{\partial \dot{u}_i}} + \frac{\partial L}{\partial \dot{q}_j}\frac{\partial \dot{q}_j}{\partial \dot{u}_i} = \frac{\partial L}{\partial \dot{q}_j}\frac{\partial \dot{q}_j}{\partial \dot{u}_i} = p_j^{(q)}\frac{\partial q_j}{\partial u_i}. \tag{2.4.10}$$

where in the last equality we have used $\frac{\partial \dot{q}_j}{\partial \dot{u}_i} = \frac{\partial q_j}{\partial u_i}$ from (2.1.6). The difference between the transformation laws (2.4.9) and (2.4.10) is the difference between the coordinate transformations for a tangent vector and a cotangent vector. Again, we will have more to say about this when we talk about tensors in Section ????.

In changing our focus from the Lagrangian to the Hamiltonian, we are replacing $L\colon TM \to \mathbb{R}$ with $H\colon T^*M \to \mathbb{R}$. What's better about T^*M, in modern terminology, is that T^*M carries a natural symplectic structure. That is to say, the cotangent bundle T^*M is a *symplectic manifold*. The study of such manifolds goes under the name "symplectic geometry", and this in some sense this is just the modern name for Hamiltonian mechanics.

This would be a natural place for us to say exactly what a symplectic manifold is, but we don't want to get too far afield. We will leave it to the reader to go look that up. But we do want to point out that for a symplectic manifold W, the collection of smooth functions $C^\infty(W)$ has a Lie bracket called the **Poisson bracket**, denoted with curly braces. For the case $W = T^*M$, with coordinates $\mathbf{p}$ and $\mathbf{q}$, the definition is

$$\{f, g\} = \sum_i \left(\frac{\partial f}{\partial q_i} \frac{\partial g}{\partial p_i} - \frac{\partial f}{\partial p_i} \frac{\partial g}{\partial q_i} \right).$$

Bilinearity and antisymmetry are clear, and as always, the Jacobi identity—while not clear—ends up being just a computation. One may check that

$$\{q_i, q_j\} = 0, \qquad \{p_i, p_j\} = 0, \qquad \{q_i, p_j\} = \delta_{ij}, \qquad \text{and} \quad \{f, H\} = \frac{df}{dt}\,.$$

The Poisson bracket is one of the "extra tools" that makes Hamiltonian mechanics a more robust theory than Lagrangian mechanics.

We will not say more about the Poisson bracket, but let us point out that in quantum mechanics there is also an important Lie bracket—the commutator bracket for linear operators on a Hilbert space. Dirac realized that making analogies between the Poisson bracket and the commutator bracket was an important tool in passing from classical to quantum mechanics. To some extent people are still trying to fully understand this, nowadays under the banner of "deformation quantization". Before Feynman, this was essentially the *only* way to do quantum mechanics—the "Hamiltonian" way. Feynman's viewpoint re-emphasized the importance of the Lagrangian. We will start to understand this better in the next part of these notes.

REMARK 2.4.11 (Significance of the Hamiltonian). ????

CHAPTER 3

Quantum mechanics

3.1. Preamble to quantum mechanics

In the introduction we gave a commutative diagram of modern physics:

$$\begin{tikzcd}
\text{Classical mechanics} \arrow[r] \arrow[d] & \text{Classical field theory} \arrow[d] \\
\text{Quantum mechanics} \arrow[r] & \text{Quantum field theory.}
\end{tikzcd}$$

Having tackled classical mechanics, we could at this point proceed in the direction of either of the two arrows. I've decided to talk about quantum mechanics next. For the sensibilities of mathematicians, [**G1**] is an excellent reference on this subject. And by the way, the heading "Quantum mechanics" here really means *non-relativistic* quantum mechanics; relativistic quantum mechanics is quantum field theory.

3.1.1. A motivating example. Quantum mechanics seems to be inherently confusing, so I want to say a few things upfront to get us oriented. Most of us have it drilled into our heads that science is about making predictions about what will happen in an experiment. Based on this, it might be surprising to find that quantum mechanics will not predict the result of any single experiment you ever do! Quantum mechanics is only a theory of *probabilities*. So if you do an experiment a thousand times and make a histogram of the results, quantum mechanics will have something to say about what that histogram looks like. But if you do an experiment *once*, quantum mechanics tells you basically nothing.

One could argue about whether this is an inherent limitation of quantum mechanics, or an inherent limitation of the universe. Physicists are mostly of the opinion that it is the latter. Maybe one day a new theory will come along to prove otherwise, but until then quantum mechanics is what we're stuck with: we can only predict probabilities, not specific outcomes.

So probability theory and quantum mechanics are inextricably woven together. And although I'm not an expert, at this point it mostly seems to me that the confusing things about quantum mechanics are actually confusing things about probability theory.

Let us now consider a silly example, really just a basic probability question. Consider a coin toss, where coming up heads (H) has probability p and coming up tails (T) has probability $1-p$. [Total aside: When I was an undergraduate I took a course on biophysics, and at some point I was having a disagreement with the professor about his use of probability theory. To illustrate my point I began by saying, "Imagine a coin where the probability of coming up heads is $\frac{1}{3}$." He looked at me like I was a crazy person, explaining—as if talking to a five-year-old—that for a coin the probability of course always had to be $\frac{1}{2}$. I couldn't exactly disagree with that! So I ended up up in the awkward position of having to explain that I wasn't a crazy person, just a mathematician.]

Anyway, let's come back to our coin. If we flip the coin 10 times, what is the probability that it comes up heads exactly n times? We all know how to solve this problem: if an "outcome" is an ordered sequence of H's and T's, there are $\binom{10}{n}$ outcomes with exactly n H's. Each such outcome has probability $p^n(1-p)^{10-n}$ of occurring, so we have

$$P(n \text{ heads}) = \binom{10}{n} p^n (1-p)^{10-n}.$$

Now I am going to give you a completely different method for solving the same problem, one using a somewhat unusual formalism. Define V, the "vector space of states," to be $V = \mathbb{C}\langle H, T\rangle$. That is, V is the two-dimensional complex vector space with basis elements H and T. We represent our single coin toss by the vector $\underline{c} = pH + (1-p)T$.

For 10 distinguishable coin tosses, we would consider the space of states to be $V^{\otimes 10}$. For indistinguishable tosses (where the order doesn't affect the outcome), we instead use $V^{\otimes 10}/\Sigma_{10}$. We compute the vector $c^{\otimes 10}$ in this state space:

$$\begin{aligned}
\mathbf{c}^{\otimes 10} &= (pH + (1-p)T)^{\otimes 10} \\
&= \sum_{\text{ordered outcomes}} p^{\#H}(1-p)^{10-\#H}(H \otimes T \otimes T \otimes H \dots) \\
&= \sum_{n=0}^{10} p^n (1-p)^{10-n} \binom{10}{n} \underbrace{(H \otimes H \otimes \dots)}_{n} \otimes \underbrace{(T \otimes T \otimes \dots)}_{10-n}
\end{aligned}$$

The first two equalities are effectively happening in the space $V^{\otimes 10}$, whereas for the last equality we really use $V^{\otimes 10}/\Sigma_{10}$.

We can now make an interesting observation: the probability of getting n heads can be simply read off as the coefficient of the appropriate term in $c^{\otimes 10}$.

The techniques of the above example are very typical of quantum mechanics. One has a vector space of "states", and probabilites are encoded as the coefficients of vectors with respect to a certain basis. ??????

3.1.2. Dirac notation for linear algebra. Let V be a finite-dimensional complex vector space, equipped with a Hermitian inner product $\langle -, - \rangle$. Recall that this pairing is additive in both variables, and the following hold:

(1) $\langle \lambda \mathbf{v}, \mathbf{w} \rangle = \overline{\lambda} \langle \mathbf{v}, \mathbf{w} \rangle$
(2) $\langle \mathbf{v}, \lambda \mathbf{w} \rangle = \lambda \langle \mathbf{v}, \mathbf{w} \rangle$
(3) $\langle \mathbf{v}, \mathbf{w} \rangle = \overline{\langle \mathbf{w}, \mathbf{v} \rangle}$
(4) $\langle \mathbf{v}, \mathbf{v} \rangle \geq 0$, with equality if and only if $\mathbf{v} = 0$;
(5) If $\langle \mathbf{v}, \mathbf{w} \rangle = 0$ for all $\mathbf{w} \in V$, then $\mathbf{v} = 0$.

This inner product gives us an abelian group isomorphism between V and V^*, given by $\mathbf{v} \longrightarrow \langle \mathbf{v}, - \rangle$. So every vector in V also determines a dual vector in V^*.

In Dirac's notation, elements in V are denoted $|\mathbf{v}\rangle$ and the dual vector in V^* is denoted $\langle\mathbf{v}|$. The elements $\langle\mathbf{w}|$ are called "bras" and the elements $|\mathbf{v}\rangle$ are called "kets" (from "bra-c-ket" of course).

We can pair a bra and a ket together via the combined notation $\langle\mathbf{w}\mid\mathbf{v}\rangle$. This can only mean one thing, which is to feed the vector $|\mathbf{v}\rangle$ into the functional $\langle\mathbf{w}|$. So we have

$$\langle\mathbf{w}\mid\mathbf{v}\rangle=\langle\mathbf{w}|\Big(|\mathbf{v}\rangle\Big)=\langle\mathbf{w},\mathbf{v}\rangle.$$

In other words, $\langle\mathbf{w}\mid\mathbf{v}\rangle$ is just another notation for the inner product. In the above equation we are mixing physicists's notation with mathematicians' notation, which we won't often do—but occasionally it is useful to accentuate a point.

Physicists also like to pair a ket with a bra in the opposite order, writing things like $|\mathbf{v}\rangle\langle\mathbf{w}|$. What could this mean? For physicists the more relevant question is more like, "What can we do with this?" For instance, we can put a ket $|\mathbf{u}\rangle$ on the right and get

$$|\mathbf{v}\rangle\langle\mathbf{w}\mid\mathbf{u}\rangle=\langle\mathbf{w}\mid\mathbf{u}\rangle\cdot|\mathbf{v}\rangle.$$

So $|\mathbf{v}\rangle\langle\mathbf{w}|$ is a device which takes a ket $|\mathbf{u}\rangle$ and outputs a multiple of $|\mathbf{v}\rangle$. We can regard it as a linear transformation $V\to V$, given by

$$|\mathbf{u}\rangle\longrightarrow|\mathbf{v}\rangle\langle\mathbf{w}\mid\mathbf{u}\rangle=\langle\mathbf{w}\mid\mathbf{u}\rangle\cdot|\mathbf{v}\rangle.$$

Notice that this linear transformation has rank 1; its image is the line containing $|\mathbf{v}\rangle$.

Now suppose that we are given a linear map $A\colon V\to V$, and an orthonormal basis $\{\mathbf{e}_i\}$ for V. Then we have $A(\mathbf{e_i})=\sum a_{ij}\mathbf{e_j}$ for some $a_{ij}\in\mathbb{C}$. We claim that we may then write

$$A=\sum a_{ij}|\mathbf{e}_j\rangle\langle\mathbf{e}_i|.$$

Indeed, if we pair the right-hand-side with $|\mathbf{e_r}\rangle$ then we get

$$\sum a_{ij}|\mathbf{e}_j\rangle\langle\mathbf{e}_i\mid\mathbf{e}_r\rangle=\sum a_{ij}|\mathbf{e}_j\rangle\delta_{ir}=\sum a_{rj}|\mathbf{e}_j\rangle=A(\mathbf{e}_r).$$

Note in particular that in the case $A=\mathrm{id}$ we have $a_{ij}=\delta_{ij}$, and so we may write

$$\mathrm{id}=\sum|\mathbf{e}_i\rangle\langle\mathbf{e}_i|.$$

This is a useful relation that is often used by physicists.

Continuing to introduce notation, physicists will write $A|\mathbf{v}\rangle$ for the mathematicians' $A(|\mathbf{v}\rangle)$ or $A\mathbf{v}$. But they also write $\langle\mathbf{w}|A$, so what can this mean? Clearly this must be the composite of our linear transformation with the functional $\langle\mathbf{w}|$:

$$V\xrightarrow{A}V\xrightarrow{\langle\mathbf{w}|}\mathbb{C}.$$

That is to say, if we pair $\langle\mathbf{w}|A$ with a ket $|\mathbf{v}\rangle$ then we get

$$\langle\mathbf{w}|A|\mathbf{v}\rangle=\langle\mathbf{w}\mid A\mathbf{v}\rangle.$$

Finally, let us briefly review the theory of adjoints in this context. Fix $\mathbf{w}\in V$ and consider the functional $V\to\mathbb{C}$ given by $\mathbf{v}\to\langle\mathbf{w},A\mathbf{v}\rangle$. Using the isomorphism $V\to V^*$ given by the inner product, this functional equals $\langle\phi(\mathbf{w}),-\rangle$ for a unique $\phi(\mathbf{w})\in V$. In this way we obtain a map $\phi\colon V\to V$ that is readily checked to be linear. This map is called the **adjoint** of A, and we will denote it by $A^\dagger$. We have the key formula $\langle A^\dagger\mathbf{w},\mathbf{v}\rangle=\langle\mathbf{w},A\mathbf{v}\rangle$ for all $\mathbf{w},\mathbf{v}\in V$.

Mixing the above with the physicists' notation, we get

$$\langle \mathbf{w}|A|\mathbf{v}\rangle = \langle \mathbf{w}|A\mathbf{v}\rangle = \langle A^\dagger \mathbf{w}|\mathbf{v}\rangle.$$

Since this holds for all $\mathbf{v}$, we can in fact write the identity of bras

$$\langle \mathbf{w}|A = \langle A^\dagger \mathbf{w}|.$$

Example 3.1.3. Putting our notation together, note that we can write

$$\langle \mathbf{v} \mid \mathbf{w}\rangle = \langle \mathbf{v} \mid \mathrm{id} \mid \mathbf{w}\rangle = \sum_i \langle \mathbf{v} \mid \mathbf{e}_i\rangle\langle \mathbf{e}_i \mid \mathbf{w}\rangle = \sum_i \overline{\langle \mathbf{e}_i \mid \mathbf{v}\rangle}\langle \mathbf{e}_i \mid \mathbf{w}\rangle.$$

The transistion from the first to the third term is called "inserting a complete set of orthonormal states".

Remark 3.1.4. When I first encountered the Dirac notation for linear algebra, I was somewhat appalled. It seemed less *careful* than mathematicians' notation (although maybe this was more a function of how physicists used it than the notation scheme itself). I also couldn't see the point of introducing all this new notation in the first place, when the mathematical notation seemed perfectly good. Just as with any new language, though, over time I have become more accustomed to it. I find myself liking the Dirac notation more and more. My point is to say, "Don't be scared of it!" Once you start forcing yourself to use the notation, you will quickly become comfortable with it. It's just linear algebra, after all, no matter how one chooses to write it.

3.1.5. Delta functions. The idea for delta functions is, remarkably, also due to Dirac. Technically speaking they are not really "functions", but physicists regard them as functions and use them to great effect. We will follow their example, but see Remark 3.1.7 below for something about the real mathematics underlying these ideas.

Consider $V = L^2(\mathbb{R})$ with the usual inner product $\langle f, g\rangle = \int_{\mathbb{R}} f(x)g(x)dx$. We again have the map $V \to V^*$, $f \mapsto \langle f, -\rangle$, but because V is infinite-dimensional this is not an isomorphism: it is injective but not surjective. As a specific example, the element of V^* sending $f \mapsto f(0)$ is not in the image. We know this because if $\delta(x)$ *were* a preimage then by integrating it against bump functions concentrated away from zero we would find that $\delta(x) = 0$ for all $x \neq 0$; and yet this would imply that $\langle \delta, g\rangle = 0$ for all g, in which case we would have $\delta = 0$.

Despite the fact that there is no function giving a preimage for $f \mapsto f(0)$, physicists like to pretend that there is. And this is what they mean by the "delta function" $\delta(x)$. At first this might seem very crazy—pretending that something exists when one has proven that it doesn't. But it's worth remembering that lots of important mathematics has developed in this way. There is no real number such that $x^2 = -1$, but once upon a time someone decided to pretend there was. You have to give up something (in this case, you need to give up having a partial order $\leq$) but in return you get a new idea about what 'numbers' are. The story of the delta function is the same way. On the one hand, you have to give up some properties that you are used to for ordinary functions; on the other hand, when all is said and done you have broadened the objects of study from functions into something greater.

The delta function and its "derivatives" satisfy a number of useful formulae, which are easily derived using integration by parts and basic substitutions (we do not justify that these tricks are valid):

Proposition 3.1.6. *For any $f \in L^2(\mathbb{R})$ one has*

(i) $\int f(x)\delta(a-x)\,dx = f(a)$

(ii) $\int f(x)\delta^{(n)}(x)\,dx = (-1)^n f^{(n)}(0)$

(iii) $\int f(x)\delta^{(n)}(a-x)\,dx = f^{(n)}(a)$

(iv) $x\delta^{(n)}(x) = -n\delta^{(n-1)}(x)$

(v) $\delta^{(n)}(\lambda x) = \frac{1}{\lambda^{n+1}}\delta^{(n)}(x)$, *for* $\lambda > 0$.

Proof. Most of these we leave to the reader, but we will do two of them to make the idea clear. For any $f \in L^2(\mathbb{R})$,

$$\int f(x)\cdot\delta'(x)\,dx = f(x)\delta(x)\Big]_{x=-\infty}^{x=\infty} - \int f'(x)\cdot\delta(x)\,dx = -\int f'(x)\delta(x)\,dx = -f'(0).$$

The first step is integration by parts, the second using that both $f(x)$ and $\delta(x)$ vanish at $x=\infty$ and $-\infty$. This proves (ii) in the case $n=1$, and the general case follows from the same argument by induction.

Likewise we compute that

$$\begin{aligned}\int f(x)\cdot x\delta'(x)\,dx = -(xf(x))'\big|_{x=0} &= -(xf'(x)+f(x))\big|_{x=0}\\ &= -f(0)\\ &= -\int f(x)\cdot\delta(x)\,dx.\end{aligned}$$

Since this holds for every $f \in L^2(\mathbb{R})$ we conclude that $x\delta'(x) = \delta(x)$, and this proves (iv) in the case $n=1$. □

Remark 3.1.7. Although we will follow the physicists in treating delta functions as if they were actually functions, it is worthwhile to say a little about how mathematicians make sense of all this.

Consider a vector space $T \subseteq L^2(\mathbb{R})$, which we will call the space of "test-functions." For example, T could be chosen to be the space of smooth functions with compact support. This space can be topologized based on notions of functional convergence. We next define the space of **distributions** $\mathcal{D}(\mathbb{R}) \subseteq T^*$ by

$$\mathcal{D}(\mathbb{R}) = \Bigg\{\phi\colon T \to \mathbb{R}\,\Bigg|\,\phi \text{ is linear and } \lim_{n\to\infty}\phi(f_n) = \phi(\lim_{n\to\infty} f_n)$$
$$\text{for any convergent sequence } \{f_n\} \text{ in } T\Bigg\}.$$

Note that $L^2(\mathbb{R})$ includes into $\mathcal{D}(\mathbb{R})$ by the map $f \mapsto \langle f, -\rangle$, and so one may regard distributions as being "generalized functions." In this context the delta function *is* the linear functions $f \mapsto f(0)$, just thought of as an element in $\mathcal{D}(\mathbb{R})$.

The theory of distributions is due to Schwartz. One may read about it in many basic modern textbooks on analysis, for instance ?????

3.1.8. Delta Functions and Fourier Transforms. It will be convenient in this section to imagine that we have two copies of the real numbers, denoted $\mathbb{R}_x$ and $\mathbb{R}_k$. Functions $f\colon \mathbb{R}_x \to S$ (for any set S) will be denoted $f(x)$, and functions $f\colon \mathbb{R}_k \to S$ will be denoted $f(k)$.

Consider a map $f\colon \mathbb{R}_x \to \mathbb{C}$. The Fourier transform of f (if it exists) is defined to be the function $\hat{f}\colon \mathbb{R}_k \to \mathbb{C}$ given by

$$\hat{f}(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-ikx} f(x)\, dx.$$

Similarly, for a function $g\colon \mathbb{R}_k \to \mathbb{C}$ its *inverse* Fourier transform (if it exists) is defined to be the function $\check{g}\colon \mathbb{R}_x \to \mathbb{C}$ given by

$$\check{g}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{ixk} g(k)\, dk.$$

For nice enough functions these operations are indeed inverses.

Now we are going to be physicists and not worry about when things are "nice enough" to make sense, but just follow the manipulations formally. Here we go:

$$\begin{aligned}\check{\hat{f}}(x) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{ixk} \hat{f}(x)\, dk \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathrm{e}^{ikx} \mathrm{e}^{-ikx'} f(x')\, dx'\, dk \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathrm{e}^{ik(x-x')} f(x')\, dx'\, dk.\end{aligned}$$

Now plug in $x = 0$ to get

$$\begin{aligned}f(0) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathrm{e}^{-ikx} f(x')\, dx\, dk \\ &= \int_{-\infty}^{\infty} f(x) \Big[\frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{-ikx}\, dk \Big]\, dx.\end{aligned}$$

This seems to give us a formula for $\delta(x)$, namely

$$\delta(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{-ikx}\, dk. \tag{3.1.9}$$

Up to scalar multiple this is the Fourier transform of the constant function 1—except for the fact that this function doesn't *have* a Fourier transform, because the above integral doesn't converge. But okay, let's not ask too many questions!

The integral in (3.1.9) is a so-called "oscillatory integral." The values of e^{-ikx} rotate around the unit circle, so that adding them up through one complete revolution gives zero. On this intuitive level it makes sense that $\int_{-\infty}^{\infty} e^{-ikx}\, dk$ is zero when $x \neq 0$, whereas when $x = 0$ this is the integral of the constant function 1 and therefore infinite. Quantum field theory is filled with these kinds of oscillatory integrals.

Before leaving this topic we prove one result that will be needed later, and also nicely demonstrates some of the above techniques for manipulating delta functions.

Lemma 3.1.10. *For $\lambda > 0$ and $k, l \in \mathbb{Z}_{\geq 0}$, $\iint p^k u^l e^{ipu/\lambda} dp\, du = \delta_{k,l} \cdot 2\pi \lambda^{k+1} \cdot k! \cdot i^k$.*

PROOF. We just compute

$$\begin{aligned}\iint p^k u^l e^{ipu/\lambda} dp\, du &= \int_{-\infty}^{\infty} u^l \left[\int_{-\infty}^{\infty} p^k e^{ipu/\lambda}\, dp\right] du \\ &= \int_{-\infty}^{\infty} u^l \cdot \frac{2\pi}{i^k} \cdot \delta^{(k)}\Big(\frac{u}{\lambda}\Big)\, du \\ &= \frac{2\pi}{i^k} \cdot \lambda^{k+1} \cdot \int_{-\infty}^{\infty} u^l \delta^{(k)}(u)\, du \\ &= \frac{2\pi}{i^k}\lambda^{k+1} \cdot (-1)^k \cdot \frac{d^k}{du^k}\Big(u^l\Big)\Big|_{u=0} \\ &= \frac{2\pi}{i^k}\lambda^{k+1} \cdot (-1)^k \cdot \delta_{k,l} k! \\ &= 2\pi\lambda^{k+1} \cdot i^k \cdot k! \cdot \delta_{k,l}.\end{aligned}$$

□

3.1.11. Gaussian integrals. Because quantum mechanics is intimately tied to probability theory, Gaussian functions end up playing a crucial role. The following proposition lists a number of important integrals that will be needed later.

PROPOSITION 3.1.12.

(a) $\displaystyle\int_{-\infty}^{\infty} \mathrm{e}^{-ax^2} dx = \sqrt{\frac{\pi}{a}}$

(b) $\displaystyle\int_{-\infty}^{\infty} \mathrm{e}^{-ax^2+bx} dx = \mathrm{e}^{\frac{b^2}{4a}}\sqrt{\frac{\pi}{a}}$

(c) If Q is a symmetric, positive definite $n\times n$ matrix, then $\displaystyle\int_{\mathbb{R}^n} \mathrm{e}^{-\mathbf{x}^T Q\mathbf{x}} = \sqrt{\frac{\pi^n}{\det Q}}$.

(d) With Q as above and J any $1 \times n$ matrix, $\displaystyle\int_{\mathbb{R}^n} \mathrm{e}^{-\mathbf{x}^T Q\mathbf{x}+J\mathbf{x}} = \mathrm{e}^{\frac{JQ^{-1}J^T}{4}}\sqrt{\frac{\pi^n}{\det Q}}$.

(e) $\displaystyle\int_{-\infty}^{\infty} \mathrm{e}^{-ax^2} x^n dx = \begin{cases} 0 & \textit{if } n \textit{ is odd} \\ \sqrt{\dfrac{\pi}{a^{n+1}}} \cdot \dfrac{1\cdot 3\cdot 5\cdot\dots\cdot(n-1)}{2^{n/2}} & \textit{if } n \textit{ is even}\end{cases}$

PROOF. Part (a) is usually proven by a clever trick using double integrals and polar coordinates:

$$\begin{aligned}\left(\int_{\mathbb{R}} e^{-ax^2}\, dx\right)^2 = \int_{\mathbb{R}} e^{-ax^2}\, dx \cdot \int_{\mathbb{R}} e^{-ay^2}\, dy &= \int_{\mathbb{R}^2} e^{-a(x^2+y^2)}\, dx\, dy \\ &= \int_0^{\infty}\int_0^{2\pi} e^{-ar^2} r\, dr\, d\theta \\ &= 2\pi \cdot \frac{1}{2a} = \frac{\pi}{a}.\end{aligned}$$

Part (b) follows from (a) by completing the square.

Part (c) follows from (a) by diagonalization. Precisely, we know that $Q = P^T DP$ for a diagonal matrix $D = \operatorname{diag}\{d_1, d_2, \dots, d_n\}$. Then using the change of variables $\mathbf{y} = P\mathbf{x}$ we get

$$\int_{\mathbb{R}^n} e^{-\mathbf{x}^T Q\mathbf{x}} = \int_{\mathbb{R}^n} \mathrm{e}^{-\mathbf{y}^T D\mathbf{y}} \frac{d\mathbf{y}}{|\det(P)|} = \frac{1}{|\det(P)|}\sqrt{\frac{\pi}{d_1}}\cdots\sqrt{\frac{\pi}{d_n}}$$
$$= \sqrt{\frac{\pi^n}{(\det P)^2 \det D}} = \sqrt{\frac{\pi^n}{\det Q}}.$$

Part (d) again follows from part (c) by completing the square. Specifically, let $\mathbf{y} = \mathbf{x} - (2Q)^{-1}J^T\mathbf{x}$.

Part (e) is clear when n is odd because we are integrating an odd function, and when n is even it follows by differentiating the equation in (a) with respect to a (and moving the differential operator under the integral sign). □

3.2. Quantum mechanics

In classical mechanics, a mechanical system gives us a manifold M called "configuration space". Observables of the system—like position, or momentum, or energy—are smooth functions $T^*M \to \mathbb{R}$. In quantum mechanics, we have instead a complex Hilbert space $\mathcal{H}$ and observables correspond to certain linear operators $K\colon \mathcal{H} \to \mathcal{H}$. The eigenvalues of K correspond to what we would classically think of as the possible values for the observable quantity corresponding to K.

Write $\langle -, - \rangle$ for the Hermitian inner product on $\mathcal{H}$, and write $|\psi| = \sqrt{\langle \psi, \psi \rangle}$. Elements of $\mathcal{H}$ are thought of as "states" of our quantum-mechanical system.

In many cases $\mathcal{H}$ is simply the Hilbert space $L^2(M)$ of complex-valued functions on M, with inner product given by $\langle f, g \rangle = \int_M \overline{f(x)}g(x)dx$.

Our goal in this section is to list some of the basic principles underlying quantum mechanics, and then to explore a few examples. Below we call these principles "axioms", although that suggests a greater level of rigor than we actually intend.

Remark 3.2.1. Before proceeding further we need an explanation and an assumption. Given an observable operator $K\colon \mathcal{H} \to \mathcal{H}$, the eigenvalues of K will sometimes be discreet and sometimes form a continuum. In both these cases, we will make the general assumption that every eigenspace of K is one-dimensional. This is not entirely a reasonable assumption, but it will cover most of our examples for a while.

If λ is an eigenvalue of K, we will write $|K = \lambda\rangle$ for some chosen unit vector in the corresponding eigenspace. This notation is hard to parse at first: we are using the Dirac ket notation $|v\rangle$, but with the symbols "$K = \lambda$" replacing v. In the case that the spectrum of K is continuous, we assume things set up so that $\lambda \mapsto |K = \lambda\rangle$ is a continuous map. As is common in quantum mechanics texts, if the operator K is understood then we will abbreviate $|K = \lambda\rangle$ to just $|\lambda\rangle$.

We come now to our first axiom, stating how our mathematical setup encodes something measurable about the universe.

Axiom 3.2.2. *Let $K\colon \mathcal{H} \to \mathcal{H}$ be an observable of the system, and let $v \in \mathcal{H}$ be an eigenvector of K with eigenvalue λ. Then for any state $\psi \in \mathcal{H} - \{0\}$, the probability of a system in state ψ being measured to have K-observable equal to λ is*

$$\frac{|\langle v|\psi\rangle|^2}{|v|^2 \cdot |\psi|^2}.$$

Note that this number can also be written as

$$\frac{|\langle \lambda|\psi\rangle|^2}{|\psi|^2}.$$

Several remarks are in order about the above axiom:

(1) Note that if $u \in \mathbb{C} - \{0\}$ then the states ψ and $u\psi$ lead to the same probabilities for all observables. Based on this, sometimes one says that the physical states of a system correspond to complex *lines* in the Hilbert space $\mathcal{H}$. One can get in trouble taking this too far, however.

(2) In a sneaky way, the above axiom conceals the most important thing about quantum mechanics. All the probabilities that quantum mechanics predicts come about as norm-squares of complex numbers; these complex numbers are usually called **probability amplitudes**. For instance, if ψ is a unit vector in $\mathcal{H}$ then the probability amplitude for a system in state ψ being measured to have K-observable equal to λ is just $\langle \lambda|\psi\rangle$.

The previous axiom, when taken far enough, leads to an important conclusion. Suppose that μ and λ are distinct eigenvalues of an operator $K \colon \mathcal{H} \to \mathcal{H}$ that represents a certain observable. Then clearly we should expect

$$\langle \mu|\lambda\rangle = 0;$$

for a state in which one knows that $K = \lambda$ should have zero probability of being measured to have $K = \mu$.

This in turn leads to another observation. An observable linear operator $K \colon \mathcal{H} \to \mathcal{H}$ must have orthogonal eigenvectors, and it must have real eigenvalues (because the eigenvalues correspond to the classical values of the observable). This is equivalent to saying that K is Hermitian (also known as self-adjoint). We thus have

Axiom 3.2.3. *Observable quantities of the mechanical system correspond to* ***Hermitian*** *operators* $K \colon \mathcal{H} \to \mathcal{H}$.

Remark 3.2.4. Given a collection of vectors $|\lambda\rangle \in \mathcal{H}$, for λ in some indexing set S, we wish to say what it means for this collection to be orthonormal. In the case that S is discreet this means the usual thing, namely

$$\langle \lambda|\lambda'\rangle = \delta_{\lambda,\lambda'}.$$

In the case where S is a continuum, such as $\mathbb{R}$, a different statement is needed. One usually writes

$$\langle \lambda|\lambda'\rangle = \delta(\lambda - \lambda')$$

and interprets this as an identity of functions $\mathbb{R} \to \mathbb{R}$ for any fixed value of λ (or λ'). Here the right-hand-side is a suitable Dirac delta function, and in fact these kinds of statements were Dirac's motivation for introducing delta functions in the first place.

Now, the statment is confusing at first because $\langle \lambda|\lambda\rangle$ seems to be undefined (because $\delta(0)$ is undefined). To make sense of this we need to consider a brief example.

Let $\mathcal{H} = L^2(\mathbb{R})$ with its usual Hermitian inner product. The operator $K \colon \mathcal{H} \to \mathcal{H}$ given by $K(f) = -if'$ is Hermitian, and its eigenvectors are clearly the functions $f(x) = e^{ikx}$ for $k \in \mathbb{R}$. But note that these functions are not in the original Hilbert space $\mathcal{H}$; in fact K has no eigenvectors in $\mathcal{H}$, although this is somehow not what one wants to happen. It turns out to be a good idea to expand $\mathcal{H}$ somewhat, via an inclusion $\mathcal{H} \subseteq \mathcal{H}'$ where $\mathcal{H}'$ is another Hilbert space that contains $\mathcal{H}$ as a dense subspace. In our case one might imagine throwing in the functions e^{ikx} and whatever else one might need along with them. But if we do this, the inner product gets screwed up, as $\int_x \overline{f(x)} f(x) dx$ doesn't make sense when $f(x) = e^{ikx}$. In fact we have

$$\langle e^{ikx} | e^{ik'x} \rangle = 2\pi \cdot \delta(k - k').$$

The mathematical setting for all of this is something called a **rigged Hilbert space**, which is a pair $\mathcal{H} \subseteq \mathcal{H}'$ of Hilbert spaces having certain properties. We will not go into this definition in any detail, but it is something introduced by analysts to help with the spectral theory of operators. For us what will be important is that anytime we have a continuous family of eigenvalues for an operator, our eigenvectors are really living in the $\mathcal{H}'$ rather than the $\mathcal{H}$ and so their inner products don't always make sense under the usual rules.

3.2.5. Expectation values and uncertainties. Because quantum mechanics only allows us to predict probabilities of certain measurements, we need to talk about expectation values and deviations (or "uncertainties"). If $K \colon \mathcal{H} \to \mathcal{H}$ is an observable and $\psi \in \mathcal{H}$, then the expectation value for the observable K in state ψ is

$$\langle K \rangle_\psi = \int_\lambda \lambda \cdot \frac{|\langle \lambda | \psi \rangle|^2}{|\psi|^2} \, d\lambda.$$

This is the usual formula one is used to for an expectation value: the sum over all possible outcomes of the outcome times the probability of its occurrence. When the state ψ is understood, it is common to just write $\langle K \rangle$.

Here is another formula for the expectation value:

PROPOSITION 3.2.6. *We have* $\langle K \rangle_\psi = \dfrac{\langle \psi | K | \psi \rangle}{\langle \psi | \psi \rangle}$.

PROOF. Let $|\lambda\rangle$ be the eigenvectors of K, and write

$$\psi = \int \psi(\lambda) |\lambda\rangle \, d\lambda.$$

Then $K\psi = \int \lambda \psi(\lambda) \cdot |\lambda\rangle$, and thus

$$\begin{aligned} \langle \psi | K \psi \rangle = \iint_{\lambda, \lambda'} \overline{\psi(\lambda')} \lambda \psi(\lambda) \langle \lambda' | \lambda \rangle &= \iint_{\lambda, \lambda'} \overline{\psi(\lambda')} \lambda \psi(\lambda) \delta(\lambda - \lambda') \\ &= \int_\lambda \overline{\psi(\lambda)} \lambda \psi(\lambda) \\ &= \int_\lambda \lambda |\psi(\lambda)|^2. \end{aligned}$$

But $\psi(\lambda)$ is just $\langle \lambda | \psi \rangle$ (using that $\langle \lambda | \lambda' \rangle = \delta(\lambda - \lambda')$ again). The desired identity follows immediately. □

A second quantity that statistics tells us to consider is the **deviation** of an observable, usually called the **uncertainty** in the context of quantum mechanics. Gven a state $\psi \in \mathcal{H}$, this is defined by

$$(\Delta K)_\psi = \sqrt{\langle K^2 - \langle K \rangle_\psi^2 \rangle_\psi}.$$

Again, we will usually drop the subscripts when the state is understood.

In the examples of the next section we will look carefully at the expectation values and uncertainties of various observables in different states. For now, these are just two definitions to get used to.

3.2.7. The time evolution of a quantum mechanical system. Finally, we come to the last of our general axioms. This one concerns how a quantum-mechanical system develops over time. For this we need the constant $\hbar$, which is called the **reduced Planck constant** or the **Dirac constant** in the physics literation. All we need to know about it is that it is a physical constant, like the speed of light, which is built into the universe somehow. Perhaps it is also worth saying that the SI units are Joules × seconds, and that in these units $\hbar$ is extremely small (on the order of 10^{-34}).

AXIOM 3.2.8 (Time evolution axiom). *Let $H\colon \mathcal{H} \to \mathcal{H}$ be the observable operator corresponding to energy (usually called the **Hamiltonian operator**). Then if the quantum-mechanical system is in state ψ at time t', it will be in state*

$$e^{-i(t-t')H/\hbar}\psi$$

at time t. Here

$$e^{sH} = \operatorname{Id} + sH + \frac{s^2}{2}H^2 + \cdots$$

is the usual exponential of a linear operator. If U is a unitary operator such that $H = UDU^{-1}$ where D is diagonal and real, then

$$e^{sH} = U(e^{sD})U^{-1},$$

*whcih makes it clear that the above series converges. The operator $e^{-isH/\hbar}$ is called the **time evolution operator** (for time intervals of length s).*

An eigenvector of the Hamiltonian operator H is called a **stationary state** of the system. The reason can be explained as follows. Suppose that ψ is a stationary state with H-eigenvalue equal to λ (where $\lambda \in \mathbb{R}$, of course). Then

$$e^{-iHt\hbar}\psi = e^{-i\lambda t\hbar}\psi.$$

For any observable operator $K\colon \mathcal{H} \to \mathcal{H}$ and possible eigenvalue μ, we now compute that

$$|\langle \mu | e^{-iHt/\hbar}\psi\rangle| = |\langle \mu | e^{-i\lambda t/\hbar}\psi\rangle| = |e^{-i\lambda t/\hbar} \cdot \langle \mu, \psi\rangle| = |\langle \mu, \psi\rangle|.$$

It follows that at all times t, the probability of our system having K-observable equal to μ remains constant—that is, the probability is independent of t. This is the sense in which the state is "stationary". Notice that the state vector itself is not staying constant: as time progresses the vector changes its "phase". It is only the probabilities for measurements about the system that remain constant.

3.2.9. Generalities about one-dimensional systems. For the remainder of this section we will study one-dimensional quantum mechanical systems (that is, systems where the underlying classical manifold M would be one-dimensional). The classical concept of "position" is replaced by a certain operator $X\colon \mathcal{H} \to \mathcal{H}$, and the classical concept of "momentum" is replaced by an operator $P\colon \mathcal{H} \to \mathcal{H}$.

The eigenvectors of X are denoted $|x\rangle$ (for $x \in \mathbb{R}$), and the eigenvectors of P are denoted $|p\rangle$. This notation is confusing to mathematicians, as it is unclear what $|5\rangle$ should mean—is it an eigenvalue of X with eigenvalue 5, or an eigenvector of P

with eigenvalue 5? This is why $|X = 5\rangle$ and $|P = 5\rangle$ are a bit better notation, but we will continue to follow the physicists' notation as much as we can.

We postulate that the vectors $|x\rangle$, for $x \in \mathbb{R}$, together form an orthonormal basis for $\mathcal{H}$. Recall that this means

$$\langle x|x'\rangle = \delta(x - x')$$

as functions $\mathbb{R} \to \mathbb{R}$, where either x or x' is considered fixed. This is called the "X-basis" for $\mathcal{H}$. Since we have a basis, any state $\psi \in \mathcal{H}$ may be written in this basis, which would give us an expression like

$$\psi = \int_x \psi(x) \cdot |x\rangle.$$

Notice that this allows us to identify any vector $\psi \in \mathcal{H}$ with a function $\psi\colon \mathbb{R} \to \mathbb{C}$. This function is called the **wave function** for the state, in the X-basis.

The vectors $|p\rangle$ likewise form the "P-basis", and we can write any vector ψ as

$$\psi = \int_p \psi(p) \cdot |p\rangle.$$

This gives us another kind of "wave function" associated to ψ. Of course there are lots of other bases for $\mathcal{H}$ as well, but these are the most common ones (particularly the X-basis).

Now we come to a crucial axiom of our theory.

Axiom 3.2.10 (Schrödinger axiom). $\langle x|p\rangle = \frac{1}{\sqrt{2\pi\hbar}} e^{ixp/\hbar}$.

Strangely, you will not find this axiom in any book on quantum mechanics (that I know of, anyway). It is equivalent to something you *do* find in the textbooks, however:

Proposition 3.2.11. *In the X-basis we have $P\psi = -i\hbar\frac{d\psi}{dx}$ (or we can just write $P = -i\hbar\frac{d}{dx}$ as an identity of operators).*

What the proposition says is that if $\psi = \int_x \psi(x)|x\rangle$, then $P\psi = \int_x(-i\hbar\psi'(x))|x\rangle$. In other words, when modelling vectors of $\mathcal{H}$ by wave functions using the X-basis, the operator P becomes $-i\hbar\frac{d}{dx}$.

Before proving the above proposition, let us be clear that we understand the notation. Certainly $X(|x\rangle) = x \cdot |x\rangle$ and $P(|p\rangle) = p \cdot |p\rangle$, just by virtue of $|x\rangle$ and $|p\rangle$ being eigenvectors with the indicated eigenvalue.

Proof. This is just a computation:

$$\begin{aligned} P\psi = P\Big(\int \psi(x)|x\rangle\Big) &= P\Big(\iint \psi(x) \cdot |p\rangle\langle p|x\rangle\Big) \qquad \Big(\text{using Id} = \int_p |p\rangle\langle p|\Big) \\ &= \iint \psi(x) \cdot \langle p|x\rangle \cdot P(|p\rangle\Big) \\ &= \iint \psi(x) \cdot \langle p|x\rangle \cdot p \cdot |p\rangle \\ &= \iint_{x,p}\int_{x'} \psi(x) \cdot \langle p|x\rangle \cdot p \cdot |x'\rangle\langle x'|p\rangle \\ &= \int_x \Big(\iint_{x',p} \psi(x')\langle p|x'\rangle p\langle x|p\rangle\Big) \cdot |x\rangle \end{aligned}$$

(changing the role of x and x' in the final line). Using the Schrödinger axiom now, the term inside the parentheses in the final line is

$$\begin{aligned}\frac{1}{2\pi\hbar}\iint_{x',p}\psi(x')e^{ixp/\hbar}\cdot p\cdot e^{-ix'p/\hbar} &= \frac{1}{2\pi\hbar}\iint_{x',p}\psi(x')pe^{i(x-x')p/\hbar}\\ &= \frac{1}{2\pi\hbar}\iint_{u,p}\psi(x-u)pe^{iup/\hbar} \qquad \text{(setting } u=x-x'\text{)}\\ &= \frac{1}{2\pi\hbar}\iint_{u,p}\big[\psi(x)-u\psi'(x)+\frac{u^2}{2}\psi''(x)-\cdots\big]pe^{iup/\hbar}.\end{aligned}$$

Using Lemma 3.1.10, the only term in the Taylor expansion which will result in a nonzero integral is the term $u\psi'(x)$. So by that lemma the above expression is equal to

$$\frac{1}{2\pi\hbar}\cdot -\psi'(x)\cdot 2\pi\hbar^2\cdot i = -i\hbar\psi'(x).$$

We have thus shown that

$$P(\psi)=\int_x -i\hbar\psi'(x)|x\rangle.$$

That is, $P(\psi)=-i\hbar\psi'$. □

Here is one more example of the above kinds of manipulations:

Proposition 3.2.12. $XP-PX=i\hbar\cdot\mathrm{Id}$.

Proof. The fastest way to prove this, given what we have done above, is just to work in the X-basis. In this basis the operators X and P have the form

$$X\colon \psi(x)\mapsto x\psi(x) \qquad P\colon \psi(x)\mapsto -i\hbar\psi'(x).$$

It takes one line to deduce that $[X,P]=i\hbar$.

But just for practice, let's work in the P-basis and do the computation from scratch, only using the Schrödinger Axiom. We have $XP(|p\rangle)=X(p|p\rangle)=pX(|p\rangle)$ and so

$$(XP-PX)(|p\rangle)=(p\,\mathrm{Id}-P)(X(|p\rangle)).$$

But since $|p\rangle=\int_x\langle x|p\rangle|x\rangle$, we get $X(|p\rangle)=\int_x\langle x|p\rangle X(|x\rangle)=\int_x\langle x|p\rangle x\cdot|x\rangle$. We can change this back into the P-basis by

$$X(|p\rangle)=\int_x\langle x|p\rangle x\cdot|x\rangle=\iint_{x,p'}\langle x|p\rangle x\cdot\langle p'|x\rangle|p'\rangle.$$

Therefore

$$\begin{aligned}
(XP-PX)(|p\rangle) &= \iint_{x,p'} \langle x|p\rangle x\langle p'|x\rangle \cdot (p\,\mathrm{Id}-P)(|p'\rangle)\\
&= \iint_{x,p'} \langle x|p\rangle x\langle p'|x\rangle \cdot (p-p')|p'\rangle\\
&= \int_{p'} \Bigl(\frac{1}{2\pi\hbar i}\int_x ixe^{ix(p-p')/\hbar}(p-p')\Bigr)|p'\rangle\\
&= \int_{p'} \Bigl(\frac{1}{i\hbar}\delta'\Bigl(\frac{p-p'}{\hbar}\Bigr)(p-p')\Bigr)|p'\rangle\\
&= -\int_{p'} \Bigl(\frac{1}{i\hbar}\cdot\hbar^2\delta(p-p')\Bigr)|p'\rangle\\
&= i\hbar|p\rangle.
\end{aligned}$$

In the second-to-last equality we have used $\delta'(x/a)=a^2\delta'(x)$ and $x\delta'(x)=-\delta(x)$. □

Proposition 3.2.13 (Heisenberg uncertainty principle). *For any state $\psi\in\mathcal{H}$ one has $(\Delta X)(\Delta P)\geq\frac{\hbar}{2}$.*

We will not give the proof of the above result right now, although one can find it in [**G1**]. It is a purely mathematical consequence of the commutation relation $[X,P]=i\hbar$.

3.3. Examples of one-dimensional quantum systems

We consider a particle moving (or maybe just "existing") along the real line, with a potential energy $V(x)$ depending only on its location. In Hamiltonian mechanics we would write down the Hamiltonian function

$$H(x,p)=\frac{p^2}{2m}+V(x)$$

representing the total energy of the system. Below we will consider the quantum mechanical systems corresponding to three potentials $V(x)$:

(1) The free particle, where $V(x)=0$ for all x.
(2) The infinite square well, with $V(x)=0$ for $-L<x<L$ and $V(x)=\infty$ for $x<-L$ or $x>L$ (here L is some fixed positive real number).
(3) The simple harmonic oscillator, where $V(x)=\frac{k}{2}x^2$.

These are the standard first examples one sees in any quantum mechanics course.

3.3.1. The free particle.

This example is in some ways the simplest, and in others ways the most confusing. We will work in the X-basis, so that any state ψ gets identified with a wave function $\psi\colon\mathbb{R}\to\mathbb{C}$ via the equation $\psi=\int_x\psi(x)|x\rangle$. Under this correspondence the state $|x'\rangle$ (a state where we know with absolute certainty that a particle is at location x') has wave function $\delta(x'-x)$.

Now suppose we have a particle whose location is unknown to us, but where we know with certainty that the momentum is equal to p. The state is then the eigenvector $|p\rangle$ of the momentum operator, and the corresponding wave function is

$$\psi_p(x)=\langle x|p\rangle=\frac{e^{ixp/\hbar}}{\sqrt{2\pi\hbar}}$$

by the Schrödinger Axiom. Notice that ψ_p is not normalizable: $\int_x \overline{\psi_p(x)}\psi_p(x)dx$ is not finite. This is because ψ_p is not in our original Hilbert space $L^2(\mathbb{R})$, rather it is in the "larger" rigged Hilbert space. This is spectral theory again: the momentum operator P didn't have enough eigenvectors in $L^2(\mathbb{R})$.

The fact that neither of the wave functions for $|x\rangle$ or $|p\rangle$ are in $L^2(\mathbb{R})$ has an interesting physical interpretation. Really, it is not possible to have a physical particle in which we exactly know the position or momentum—for then the uncertainty principle would imply that we have infinite uncertainty in the other observable, and this is unrealistic. One does have states where the level of uncertainty (of either position or momentum) is incredibly small, and in these cases the wave functions *approximate* the ones given above, but one never has exact equality.

So one sees that the expansion of $L^2(\mathbb{R})$ to include functions such as $\delta(x)$ and $e^{ipx/\hbar}$ (for fixed p) corresponds to introducing certain "idealized physical states" which are limits of actual physical states. Hopefully this seems reasonably intuitive.

One should think of our quantum-mechanical free particle as having an existence which stretches out over the entire real line, where at each position there is a complex number specifying the probability amplitude for the particle to be detected at that spot. It is reasonable to expect these complex numbers to vary continuously, so as we look up and down the real line (at any given moment) we see these complex numbers rotating and growing/shrinking. These complex numbers are just the values $\psi(x)$ of our wave function.

Keeping this in mind, let us look again at the wave function $\psi_p(x) = \frac{1}{\sqrt{2\pi\hbar}} \cdot e^{ixp/\hbar}$. Here all of the values have norm 1, and so our particle is equally likely to be detected at all positions (the uncertainity in position is infinite, as we will see in a moment). Our probability amplitudes rotate as we look up and down the real axis, and the "speed" at which they are rotating is determined by p (note that "speed" is a bad word here, since we are looking at a single moment in time). The larger the momentum, the faster our unit complex numbers are rotating as we look up and down the line. This is a good thing to keep in mind in all examples: in quantum mechanics having "momentum" has to do with spatial phase changes of probability amplitudes, with large momentum meaning that the phase changes very rapidly throughout space. Note that this connects directly with the interpretation of the momentum operator as $-i\hbar\frac{d}{dx}$.

Up until now in this example we have only talked about the state of our particle at one moment in time. Now let us role the film and let time fly. According to our time evolution axiom, this amounts to applying the operator $e^{-itH/\hbar}$ to the original state of the system. The Hamiltonian for a free particle is $H = \frac{p^2}{2m}$, where m is the mass. Applying the operator $e^{-itH/\hbar}$ is easiest if we apply it to an eigenvector of H, and so in general we can best understand $e^{-itH/\hbar}\psi$ if we express ψ as a sum of eigenvectors for H. This will be our approach to understanding time evolution in almost all examples.

In the present example the eigenvalues of H will be nonnegative, since up to the factor $\frac{1}{2m}$ they are the squares of the eigenvalues for p. For a certain energy value E, the eigenstates of H will be $\psi_p(x)$ and $\psi_{-p}(x)$ where $p = \sqrt{2mE}$. Physically, these eigenvectors correspond to a particle moving in either direction along the x-axis, but with a given speed that is determined by E.

If one regards $\psi_p(x)$ as a collection of probability amplitudes at each point x on the real line, then $e^{-itH/\hbar}\psi_p(x)$ amounts to rotating all of these amplitudes by the same amount—that is, to multiplying them by $e^{-itp^2/(2m\hbar)}$. So as time marches on our probability amplitudes spin around the complex plane, and the rate at which they are spinning around is controlled by the energy of the system (the eigenvalue of H).

This is so important it is worth saying again. A particle in quantum mechanics should be thought of as having an existence that stretches out over the whole real line (or $\mathbb{R}^2$, or $\mathbb{R}^3$, or wherever the particle is living). This "existence" is governed by a probability amplitude at each point in space, and having momentum has to do with the rotation of these probability amplitudes as one looks from one point in space to another. If the particle is in an eigenstate of the Hamiltonian H, with corresponding eigenvalue E (the energy of the system), then as time goes by the probability amplitudes all rotate uniformly, at a speed directly proportional to E. Particles of high energy have probability amplitudes that rotate very fast, particles of low energy have amplitudes that rotate very slowly.

For quantum states that do not have a well-defined energy—states that are linear combinations of eigenvectors of H—the situation is of course more complicated. One can only break the state up into a sum of pure energy states, and then think of the amplitudes in each of the pure states as rotating—but at different frequencies for the different eigenvalues. Picturing the time evolution when we combine these states is not so easy.

3.3.2. Statistics of a free particle. For a particle that is known absolutely to be located at x', its wave function is $\psi(x) = \delta(x - x')$. Likewise, for a particle whose momentum is known absolutely to equal p, the wave function is $\psi(x) = \frac{1}{\sqrt{2\pi\hbar}} e^{ixp/\hbar}$. In the first case there is infinite uncertainty in momentum, and in the second case there is infinite uncertainty in position. Both of these are ideal examples, not exactly representative of a real situation.

So let us next consider a more realistic wave function, where there is some finite uncertainty in both position and momentum. There are all kinds of scenarios one could look at, but let us focus on

$$\psi(x) = Ne^{-ax^2/2+ipx/\hbar}.$$

where $a > 0$ and p are fixed real numbers. Here N is to be taken so that $|\psi|^2 = 1$, which gives $N = (a/\pi)^{1/4}$. Now we compute that

$$\langle X\rangle_\psi = \langle\psi|X|\psi\rangle = \int_{x\in\mathbb{R}} \overline{\psi}(x)x\psi(x)dx = N^2 \int_{x\in\mathbb{R}} xe^{-ax^2}dx = 0$$

and

$$\langle X^2\rangle = \int_{x\in\mathbb{R}} \overline{\psi}(x)x^2\psi(x)dx = N^2 \int_{x\in\mathbb{R}} x^2e^{-ax^2}dx = \sqrt{\frac{a}{\pi}}\frac{\sqrt{\pi}}{2}a^{-3/2} = \frac{1}{2a}.$$

Therefore we find that

$$\Delta X = \sqrt{\langle X^2\rangle - \langle X\rangle^2} = \sqrt{\langle X^2\rangle} = \sqrt{\frac{1}{2a}}.$$

When we do the corresponding calculations for momentum we find that

$$\begin{aligned}\langle P\rangle = \int_{x\in\mathbb{R}} \overline{\psi}(x)\left(-i\hbar\frac{\partial}{\partial x}\right)\psi(x)dx &= -i\hbar N^2\int_{x\in\mathbb{R}}(-ax+ip/\hbar)e^{-ax^2}dx\\ &= -i\hbar\sqrt{\frac{a}{\pi}}\frac{ip}{\hbar}\sqrt{\frac{\pi}{a}}\\ &= p.\end{aligned}$$

Likewise

$$\begin{aligned}\langle P^2\rangle = \int \overline{\psi}(x)\Big[-\hbar^2\frac{\partial^2}{\partial x^2}\psi(x)\Big]dx &= -N^2\hbar^2\int[(-ax+ip/\hbar)^2-a]e^{-ax^2}dx\\ &= -N^2\hbar^2\int[a^2x^2-p^2/\hbar^2-a]e^{-ax^2}dx\\ &= p^2+\frac{a\hbar^2}{2}.\end{aligned}$$

So

$$\Delta P = \sqrt{\langle P^2\rangle - \langle P\rangle^2} = \sqrt{p^2+\frac{a\hbar^2}{2}-p^2} = \sqrt{\frac{a}{2}}\cdot\hbar.$$

Note in this case that we have

$$\Delta X\cdot\Delta P = \frac{1}{\sqrt{2a}}\cdot\sqrt{\frac{a}{2}}\hbar = \frac{\hbar}{2}$$

which agrees with the Heisenberg Uncertaintly Principle.

These calculations become more intuitive if we look a little deeper. The probability of detecting our particle at position x' is

$$\Big|\Big\langle\delta(x-x')|\psi(x)\Big\rangle\Big|^2 = \left|\int\psi(x)\delta(x-x')\,dx\right|^2 = |\psi(x')|^2 = Ne^{-a(x')^2}.$$

This is precisely a Gaussian distribution centered at 0 whose uncertainty is $\sqrt{\frac{1}{2a}}$. Likewise, the probability of detecting our particle to have momentum p' is

$$\begin{aligned}\left|\left\langle\frac{1}{\sqrt{2\pi\hbar}}e^{ixp'/\hbar}\middle|\psi(x)\right\rangle\right|^2 &= \left|\frac{1}{\sqrt{2\pi\hbar}}N\cdot\int e^{-ax^2/2+ipx/\hbar}e^{-ixp'/\hbar}\,dx\right|^2\\ &= \left|\frac{1}{\sqrt{2\pi\hbar}}N\cdot\int e^{-ax^2/2+i(p-p')x/\hbar}\,dx\right|^2\\ &= \left|\frac{1}{\sqrt{2\pi\hbar}}N\cdot\sqrt{\frac{2\pi}{a}}e^{-(p-p')^2/2a\hbar^2}\right|^2\\ &= \frac{1}{\hbar\cdot(\pi a)^{1/2}}\cdot e^{-(p-p')^2/a\hbar^2}.\end{aligned}$$

As a function of p', this is a Gaussian distribution centered at p whose uncertainly is $\sqrt{\frac{1}{2(1/a\hbar^2)}} = \sqrt{\frac{a}{2}}\hbar$.

Next let us consider the time evolution of this particle. Recall $H = \frac{P^2}{2m}$ for a free particle. Noting that $P|p\rangle = p|p\rangle$ and $\langle x|p\rangle = \frac{e^{ixp/\hbar}}{\sqrt{2\pi\hbar}}$, we have the following:

$$\begin{aligned}
\psi_t = e^{\frac{-iH}{\hbar}t}\psi &= e^{\frac{-iH}{\hbar}t}\int_x \psi(x)|x\rangle\,dx \\
&= e^{\frac{-iH}{\hbar}t}\int_x \psi(x)\Big(\int_p |p\rangle\langle p|\,dp\Big)|x\rangle\,dx \\
&= e^{\frac{-i}{\hbar}\frac{P^2}{2m}t}\int_{x,p} \psi(x)\langle p|x\rangle\,|p\rangle \\
&= \int_{x,p} \psi(x)\langle p|x\rangle e^{\frac{-i}{\hbar}\frac{p^2}{2m}t}|p\rangle \\
&= \int_{x,p,x'} \psi(x)\langle p|x\rangle\, e^{\frac{-i}{\hbar}\frac{p^2}{2m}t}|x'\rangle\langle x'|p\rangle \\
&= \int_{x,p,x'} \psi(x)e^{\frac{-i}{\hbar}\frac{p^2}{2m}t}\cdot e^{\frac{i(x'-x)p}{\hbar}}\frac{1}{2\pi\hbar}|x'\rangle \\
&= \int_{x'}\left[\frac{1}{2\pi\hbar}\int_{x,p}\psi(x)e^{\frac{i}{\hbar}\left(-\frac{p^2}{2m}t+(x'-x)p\right)}\right]|x'\rangle.
\end{aligned} \tag{3.3.3}$$

The expression in square brackets must be $\psi_t(x')$. The integral with respect to p is just a Gaussian, so we can evaluate it using Proposition 3.1.12. We get

$$\psi_t(x') = \sqrt{\frac{m}{2\pi\hbar it}}\int e^{\frac{im(x'-x)^2}{2t\hbar}}\psi(x)\,dx.$$

At this point we recall that $\psi(x) = \left(\frac{a}{\pi}\right)^{\frac14}e^{-\frac{a}{2}x^2+\frac{ipx}{\hbar}}$. Plugging this in, we find that the integral with respect to x is *another* Gaussian:

Things become slightly horrible at this point. We get

$$\begin{aligned}
\psi_t(x') &= \sqrt{\frac{m}{2\pi\hbar it}}\Big(\frac{a}{\pi}\Big)^{\frac14}\int e^{-(\frac{a}{2}-\frac{im}{2t\hbar})x^2+x(\frac{ip}{\hbar}-\frac{imx'}{t\hbar})+\frac{im(x')^2}{2t\hbar}}\,dx \\
&= \sqrt{\frac{m}{2\pi\hbar it}}\Big(\frac{a}{\pi}\Big)^{\frac14}\cdot\sqrt{\frac{\pi}{\left(\frac{a}{2}-\frac{im}{2t\hbar}\right)}}\cdot e^{\frac{im(x')^2}{2t\hbar}}\cdot e^{-\frac{(\frac{p}{\hbar}-\frac{mx'}{t\hbar})^2}{2(a-\frac{im}{t\hbar})}} \\
&= \left[\sqrt{\frac{\pi}{a}}\Big(1+\frac{iat\hbar}{m}\Big)\right]^{-\frac12}\cdot e^{-\frac{a(\frac{pt}{m}-x')^2}{2\Delta}}\cdot e^{\frac{i}{\Delta}(-\frac{p^2t}{2m\hbar}+\frac{px'}{\hbar}+\frac{(x')^2a^2t\hbar}{2m})}
\end{aligned} \tag{3.3.4}$$

where $\Delta = 1+\frac{a^2t^2\hbar^2}{m^2}$. As a reality check, one can plug in $t=0$ and verify that the expression coincides with $\psi(x)$.

At this point we can dispense with x' and just write x. Also, we will just write N for the normalization constant in front.

Note that

$$|\psi_t(x)|^2 = \overline{\psi}_t(x)\psi_t(x) = N\overline{N}\cdot e^{-\frac{a}{\Delta}(x-\frac{pt}{m})^2}.$$

This is a Gaussian centered at $\frac{pt}{m}$ with deviation $\sqrt{\frac{1}{2(\frac{a}{\Delta})}}$. So we have

$$\langle X\rangle_t = \int \bar{\psi}_t(x)\cdot x\cdot\psi_t(x)\,dx = \Big(\frac{p}{m}\Big)t$$

and

$$(\Delta X)_t = \frac{1}{\sqrt{2\left(\frac{a}{1+a^2t^2\hbar^2/m^2}\right)}} = \frac{1}{\sqrt{2a}} \cdot \sqrt{1+\frac{a^2t^2\hbar^2}{m^2}} = (\Delta X)_0 \cdot \sqrt{1+\frac{a^2t^2\hbar^2}{m^2}}.$$

The first equation says that the average position of the particle is moving with velocity $\frac{p}{m}$. The second equation shows that for t near 0 we have $(\Delta X)_t \approx (\Delta X)_0$ to first order in t, while for very large t we have

$$(\Delta X)_t \approx \frac{1}{\sqrt{2a}} \cdot \frac{at\hbar}{m} = \left(\frac{\hbar}{m}\sqrt{\frac{a}{2}}\right)t = \frac{(\Delta P)_0}{m}t$$

That is, for small times the uncertainty in the particle's position is entirely due to the uncertainty $(\Delta X)_0$ in initial *position*, while for large times it is entirely due to the uncertainty $(\Delta P)_0/m$ in initial *velocity*.

We will also find $\langle P\rangle_t$ and $(\Delta P)_t$. A hard way to go about this is to write down the necessary integrals and compute them directly. This is possible, but it is quite a bit of work. Instead we simply observe that $H = P^2/2m$ and therefore P commutes with H; hence by ???? both $\langle P\rangle_t$ and $(\Delta P)_t$ are independent of t. So we have $\langle P\rangle_t = p$ and $(\Delta P)_t = (\Delta P)_0 = \hbar\sqrt{a/2}$.

Before leaving this example let us make one final point that will pave the way for our treatment of Fenyman path integrals in Section 3.4. In (3.3.3) we derived a quite general formula for the time evolution of a free particle:

$$\psi_t(x') = \int_x \left(\sqrt{\frac{m}{2\pi it\hbar}} \cdot e^{i(x'-x)^2m/(2t\hbar)}\right) \cdot \psi(x). \tag{3.3.5}$$

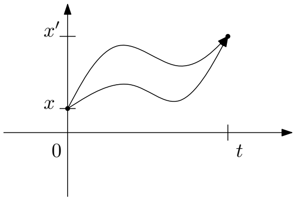
Recall that we did this starting with $\psi_t = e^{-\frac{iHt}{\hbar}}\psi$, expanding ψ in terms of eigenvectors for H (which were also eigenvectors for P), and then changing back into the X-basis to get the new wave function. The process was mostly formal, except for the fact that the simple form of H gave us a particularly easy integral to compute (this was the Gaussian integral in p that appears at the end of (3.3.3)).

The expression in parentheses in (3.3.5) is called the **propagator**, in this case from x at time 0 to x' at time t. We will write this as $U(x', t; x, 0)$. It should be regarded as the probability amplitude for a particle at position x when $t = 0$ to be detected as position x' at time t. This interpretation makes the equation $\psi_t(x') = \int_x U(x', t; x, 0)\psi(x)$ seem obvious: $\psi_t(x')$ is the probability amplitude for the particle to be detected at position x' at time t, $\psi(x)$ is the probability amplitude for the particle to be detected at position x at time 0, and this integral equation is just adding up all possibilities for how a particle *somewhere* at time 0 could end up being detected at x' at time t.

This perspective works for any time evolution problem in quantum mechanics. One can always obtain ψ_t by integrating ψ against a propagator. The challenge, which can be quite difficult, is understanding what the propagator looks like and computing these integrals. For a free particle we computed that

$$U(x', t; x, 0) = \sqrt{\frac{m}{2\pi it\hbar}} \cdot e^{i(x'-x)^2m/(2t\hbar)}. \tag{3.3.6}$$

Now here's an interesting observation about the above formula. Recall the least action principle in Lagrangian mechanics, where we consider all paths γ from $(x, 0)$ to (x', t) and choose the one which minimizes the action $S(\gamma)$:



In the absence of forces, we found that this gave us a straight line. Here we used

$$L = \frac{1}{2}m\dot{x}^2 \qquad \text{and} \qquad S(\gamma) = \int_0^t \frac{1}{2}m\dot{x}(t)^2$$

and found that the path of least action was the constant velocity path

$$\gamma_{\text{least}}(T) = \left(\frac{x' - x}{t}\right) \cdot T + x$$

The action of this path is

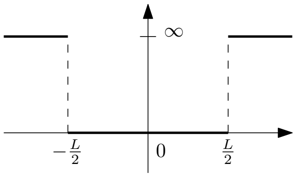
$$S(\gamma_{\text{least}}) = \frac{1}{2}m\left(\frac{x' - x}{t}\right)^2 \cdot t = \frac{m(x' - x)^2}{2t}.$$

Notice that this exact expression appears inside the exponent in our propagator; precisely, we have

$$U(x', t; x, 0) = Ce^{\frac{iS(\gamma_{\text{least}})}{\hbar}}$$

where C is some constant. This observation is important! We will see it again when we come to Feynman path integrals.

3.3.7. The infinite well. (This example is sometimes referred to as the particle in a one-dimensional box.) We imagine a particle existing on the real line, but this time with a nonzero potential function V given by



$$V(x) = \begin{cases} 0 & \text{if } -\frac{L}{2} < x < \frac{L}{2} \\ \infty & \text{otherwise.} \end{cases}$$

Classically, there are two possible behaviors. If the particle's energy is zero, then the particle just sits in the well, not moving. If the energy $E = \frac{1}{2}mv^2$ is positive, then the particle bounces between the walls, reversing direction instantly upon contact with either wall to maintain constant velocity magnitude. In the latter case, the particle is equally likely to be detected anywhere in the well $(-\frac{L}{2}, \frac{L}{2})$.

We now consider the quantum mechanical version. We recognize that $\psi(x) = 0$ if $x < -L/2$ or $x > L/2$, and that inside the well, the particle is free as in the previous example, with $H = P^2/2m$. We must analyze the eigenvectors of this operator.

Let ψ be an eigenvector of H with eigenvalue $E \geq 0$ (positive because it is the square of an eigenvalue of P). Then $p = \sqrt{2mE}$ and thus ψ is some linear combination of wave functions having eigenvalues p and $-p$, i.e.

$$\psi(x) = Ae^{ipx/\hbar} + Be^{-ipx/\hbar}$$

for some A and B, when $-L/2 < x < L/2$. We force continuity at the endpoints $-L/2$ and $L/2$:

$$\psi\left(\tfrac{L}{2}\right) = Ae^{+ipL/2\hbar} + Be^{-ipL/2\hbar} = 0$$
$$\psi\left(-\tfrac{L}{2}\right) = Ae^{-ipL/2\hbar} + Be^{+ipL/2\hbar} = 0.$$

This system corresponds to a 2×2 matrix equation in indeterminates A and B, which has nontrivial solution if and only if the determinant is nonzero, i.e.

$$e^{ipL/\hbar} - e^{-ipL/\hbar} = 2i\sin\left(\tfrac{pL}{\hbar}\right) = 0$$

which implies that $pL/\hbar = n\pi$ for some integer n. Squaring this gives $p^2L^2/\hbar^2 = n^2\pi^2$, and from $p^2 = 2mE$ we have

$$E = \frac{p^2}{2m} = \frac{n^2\pi^2\hbar^2}{2mL^2}.$$

So the set of possible eigenvalues of the Hamiltonian operator—the possible energies—is discrete. This is the first difference between the classical and quantum models of this example.

We have arranged it so that we have nontrivial solutions for A and B, and the space of such solutions is a one-dimensional. So we may as well assume $A = 1$, which gives $B = -e^{ipL/\hbar}$. Therefore

$$\psi(x) = N(e^{ipx/\hbar} - e^{ipL/\hbar}e^{-ipx/\hbar})$$

for some normalization constant N. We make this expression look more symmetrical by adjusting the constant to get

$$\psi(x) = \tilde{N}\Big(e^{\frac{ip}{\hbar}\left(x-\frac{L}{2}\right)} - e^{-\frac{ip}{\hbar}\left(x-\frac{L}{2}\right)}\Big) = \tilde{\tilde{N}}\sin\Big(\tfrac{p}{\hbar}\left(x - \tfrac{L}{2}\right)\Big).$$

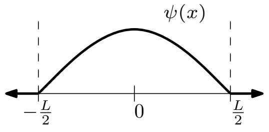
Since $pL/\hbar = n\pi$ we can write this as

$$\psi(x) = N\sin\left(\tfrac{p}{\hbar}x - \tfrac{n\pi}{2}\right) = \begin{cases} N\sin\left(\tfrac{p}{\hbar}x\right) & \text{if } n \text{ is even} \\ N\cos\left(\tfrac{p}{\hbar}x\right) & \text{if } n \text{ is odd} \end{cases}$$

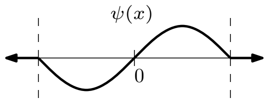
where in each case N now just represents some appropriate normalization constant (not always the same one!).

Let's look at the above wave functions for a few different values of n (recalling that $E = (n\pi\hbar/L)^2/2m$ and $p = n\pi\hbar/L$). If $n = 0$, then $\psi(x) = 0$, so the particle is non-existent since there is no probability of detecting it anywhere. We will disallow this solution.

If $n = 1$ (or -1) then $\psi(x) = N\cos(\frac{\pi}{L}x)$. Because the probability of detecting the particle at x is $|\psi(x)|^2$, we see that the particle is usually in the middle of the well, and only sometimes near the edges:



If $n = 2$ (or -2) then $\psi(x) = N\sin(\frac{2\pi}{L}x)$:



Note in particular that the particle will never be detected at the origin; or more precisely, the probability of being detected there is zero. The particle spends most of its time at the 1/4 and 3/4 mark inside the well, with little time spent at the edges or in the middle.

It should be clear enough what happens for increasing n. When $n = 3$, for instance, one gets a wave function with three "humps" (two above the x-axis and one below). The particle spends most of its time near $x = -L/3$, $x = 0$, and $x = L/3$.

Let us now proceed from these examples to a more general analysis of expectation values and uncertainties. Write $\psi_n(x)$ for the state corresponding to the integer n. One readily checks that the normalization constant N is $\sqrt{\frac{2}{L}}$ for all n. Because $\overline{\psi_n}(x) \cdot \psi_n(x)$ is always even (as is evident in the pictures),

$$\langle X \rangle_n = \int \overline{\psi_n}(x) \cdot x \cdot \psi_n(x)\, dx = 0$$

and

$$(\Delta X)_n = \langle X^2 \rangle_n^{1/2} = \left(\int x^2 \cdot \psi_n(x)^2 \right)^{1/2} = \frac{L}{2}\sqrt{\frac{1}{3} \pm \frac{2}{n^2\pi^2}}$$

where the ± is plus when n is even and minus when n is odd (in the third equality we simply looked up the integral).

Recalling that the wave functions ψ_n are real-valued, we find that

$$\langle P \rangle_n = \int \psi_n(x) \cdot -i\hbar\psi_n'(x)\, dx = -i\hbar \cdot \frac{\psi_n(x)^2}{2}\Big]_{-\infty}^{\infty} = 0.$$

Noting that $\psi_n'' = -(\frac{n\pi}{L})^2\psi_n$ and $\int \psi_n(x)^2\, dx = 1$, we have that $\int \psi_n(x) \cdot \psi_n''(x) = -(\frac{n\pi}{L})^2$. So

$$(\Delta P)_n = \left(\int \psi_n(x) \cdot -\hbar^2\psi_n''(x)\, dx \right)^{1/2} = \frac{n\pi\hbar}{L}.$$

Note that

$$\Delta X \cdot \Delta P = \frac{n\pi\hbar}{L} \cdot \frac{L}{2} \cdot \sqrt{\frac{1}{3} \pm \frac{2}{n^2\pi^2}} = \frac{\hbar}{2} \cdot \sqrt{\frac{n^2\pi^2}{3} \pm 2}.$$

The minimum value for the square root occurs when $n = 1$, and is about 1.13. In particular it is greater than 1, so we once again verify the Heisenberg Uncertainty Principle $(\Delta X)(\Delta P) \geq \hbar/2$.

Let us do one more calculation relevant to this example. For any given momentum p', let us compute the probability amplitude for a particle in state ψ_n to be detected to have momentum p'. Before embarking on the calculation, let us also take a moment and see if we can guess the answer. A particle in state ψ_n has energy $E = \frac{n^2\pi^2\hbar^2}{2mL^2}$. Since the particle is free inside the box $H = P^2/(2m)$, which suggests that the momentum of the particle is $p = \pm\sqrt{2mE} = \pm\frac{n\pi\hbar}{L}$. This suggests that $\langle p' | \psi_n \rangle$ will be zero unless p' is one of these two values.

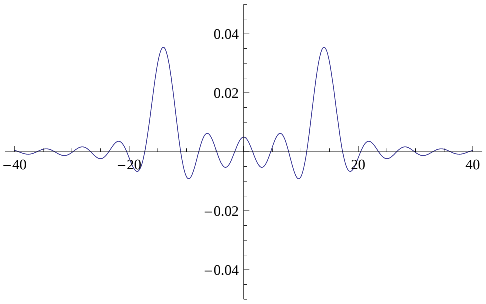
For convenience we will only do the calculation when n is odd, so that $\psi_n(x) = \sqrt{\frac{2}{L}}\cos(\frac{n\pi}{L}x)$. Then

$$\begin{aligned}
\langle p'|\psi_n\rangle &= \int_{-\frac{L}{2}}^{\frac{L}{2}} \frac{e^{-ip'x/\hbar}}{\sqrt{2\pi\hbar}} \cdot \sqrt{\frac{2}{L}}\cos\left(\frac{n\pi}{L}x\right) \\
&= \frac{1}{\sqrt{\pi\hbar L}} \cdot \frac{1}{(\frac{n\pi}{L})^2 - (\frac{p'}{\hbar})^2} e^{-ip'x/\hbar}\left[\frac{n\pi}{L}\sin\Big(\frac{n\pi}{L}x\Big) - \frac{ip'}{\hbar}\cancel{\cos\Big(\frac{n\pi}{L}x\Big)}\right]_{-L/2}^{L/2} \\
&= \frac{1}{\sqrt{\pi\hbar L}} \cdot \frac{1}{(\frac{n\pi}{L})^2 - (\frac{p'}{\hbar})^2} \cdot \frac{n\pi}{L} \cdot \left[e^{-\frac{ip'L}{2\hbar}}\sin\Big(\frac{n\pi}{2}\Big) - e^{\frac{ip'L}{2\hbar}}\sin\Big(-\frac{n\pi}{2}\Big)\right] \\
&= \frac{1}{\sqrt{\pi\hbar L}} \cdot \frac{1}{(\frac{n\pi}{L})^2 - (\frac{p'}{\hbar})^2} \cdot \frac{n\pi}{L} \cdot \pm 1 \cdot 2\cos\Big(\frac{p'L}{2\hbar}\Big) \\
&= \pm\frac{n}{2}\sqrt{\frac{\pi L}{\hbar}} \cdot \left[\frac{1}{(\frac{n\pi}{2})^2 - (\frac{p'L}{2\hbar})^2} \cdot \cos\Big(\frac{p'L}{2\hbar}\Big)\right].
\end{aligned}$$

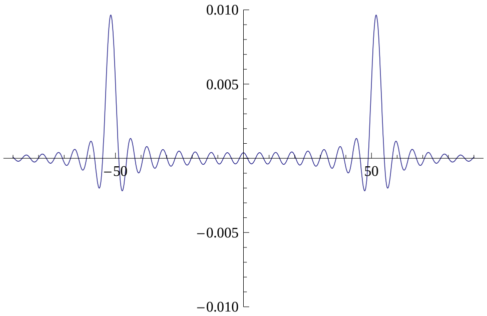
For our purposes we can ignore the constant out front and just look inside the brackets. Let us consider the function

$$f_n(u) = \frac{1}{(\frac{n\pi}{2})^2 - u^2} \cdot \cos(u)$$

(where n is odd). If one thinks of the product development for cosine from complex analysis, $\cos(u) = \prod_{k \text{ odd}}(1 - \frac{4u^2}{k^2\pi^2})$, then up to constants $f_n(u)$ is the result of removing one of the factors. Here is a graph of $f_n(u)$ for $n = 9$:



Note that the function is sharply peaked in two places, and aside from this there are some minor peaks—which rapidly move to zero as u tends to infinity. This is the behavior for every n, although as n grows the difference between the sharp peaks and the minor peaks becomes much more dramatic. For instance, here is $n = 33$ on a similar scale:



The sharp peaks are located approximately—but not exactly—at $\pm\frac{n\pi}{2}$. As n grows, the location of the peaks converges to these values.

Returning to our study of $|\langle p'|\psi_n\rangle|^2$, up to constants this is $f_n(\frac{p'L}{2\hbar})$. So there are two values of p' which are by far the most likely to occur, and these are very near (but not quite equal to) $\pm\frac{n\pi\hbar}{L}$. Other values of p' have some small likelihood of being detected, with the probability rapidly falling off (but still nonzero) for p' outside of the interval $[-\frac{n\pi\hbar}{L}, \frac{n\pi\hbar}{L}]$.

Notice that the result of this analysis is quite different from our initial guess. In particular, though only one energy can be detected for particles in the ψ_n state, nothing similar to this can be said for momentum. A whole continuous range of momenta can be detected, although two particular values are the most likely.

3.3.8. The quantum harmonic oscillator. This will be the last one-dimensional system we examine in detail. Recall that for a spring-mass system the Hamiltonian is

$$H = \frac{p^2}{2m} + V(x) = \frac{p^2}{2m} + kx^2.$$

We introduce the constant $\omega = \sqrt{\frac{k}{m}}$, which is the frequency of the classical system. Using ω instead of k we can write

$$H = \frac{p^2}{2m} + \frac{1}{2}m\omega^2x^2.$$

For the quantum harmonic oscillator we simply replace this with the operator $H = \frac{P^2}{2m} + \frac{1}{2}m\omega^2X^2$. Our goal in this section will be to analyze the eigenvectors and eigenvalues of this operator.

Remark 3.3.9. When I first learned this subject in an undergraduate quantum mechanics course, I found the whole point of this endeavor somewhat mystifying. Since I couldn't imagine what a "quantum spring" was, I didn't see the benefit in pursuing this example at all. But perhaps this is clearer to the reader, based on our work in Part 1. The real goal is to understand small perturbations from a stable equilibrium—a very common situation—and to second order approximation such things always result in positive-definite quadratic potentials. The simple harmonic oscillator is important because it is the one-dimensional case of this.

One way of proceeding at this point is to write down the eigenvector equation $H\psi = E\psi$, for $E \in \mathbb{R}$, and try to solve this. In terms of our wave functions, this is the Schrödinger differential equation

$$-\frac{\hbar^2}{2m}\psi''(x)+\frac{1}{2}m\omega^2x^2\psi(x)=E\psi(x).$$

This is not a terribly hard differential equation to solve, although doing so is a bit of work—it's certainly not a trivial problem. Dirac found a clever method for doing most of the hard work algebraically, and we will take this route. It essentially amounts to finding a Lie algebra action on the solution space of the above differential equation.

Before doing anything else, let us note that E must be nonnegative for the eigenvector equation to have a solution. For if ψ is such an eigenvector, then

$$E\langle\psi|\psi\rangle=\langle\psi|H|\psi\rangle=\frac{1}{2m}\langle\psi|P^2|\psi\rangle+\frac{1}{2}m\omega^2\langle\psi|X^2|\psi\rangle.$$

The expectation values for both P^2 and X^2 must be nonnegative, so it follows at once that E is also nonnegative.

We now embark on Dirac's method. Write the operator H as

$$H=\frac{P^2}{2m}+\frac{1}{2}m\omega^2X^2=\frac{\hbar\omega}{2}\left(\frac{P^2}{\hbar m\omega}+\frac{m\omega}{\hbar}X^2\right). \tag{3.3.10}$$

Dirac's observation is that we can almost factor the operator in parentheses: if P and X commuted we could exactly factor it, but this doesn't quite work here. Introduce the operators

$$a=\alpha X+i\beta P,\quad a^\dagger=\alpha X-i\beta P.$$

where α and β are certain real numbers we will specify later. We now compute

$$aa^\dagger=\alpha^2X^2+\beta^2P^2-\alpha\beta i[X,P]=\alpha^2X^2+\beta^2P^2+\alpha\beta\hbar$$

using that $[X,P]=i\hbar$. We likewise compute that

$$a^\dagger a=\alpha^2X^2+\beta^2P^2-\alpha\beta\hbar.$$

So $aa^\dagger+a^\dagger a=2\alpha^2X+2\beta^2P$. Comparing this with the operator in parentheses in (3.3.10), it will convenient $2\alpha^2=\frac{m\omega}{\hbar}$ and $2\beta^2=\frac{1}{m\omega\hbar}$. So define

$$\alpha=\sqrt{\frac{m\omega}{2\hbar}}\qquad\text{and}\qquad\beta=\sqrt{\frac{1}{2m\omega\hbar}}.$$

We then have

$$H=\frac{\hbar\omega}{2}(aa^\dagger+a^\dagger a).$$

We can also compute that

$$[a,a^\dagger]=2\alpha\beta\hbar=1.$$

It will be convenient to make one more defintion. Define

$$N=a^\dagger a=\frac{m\omega}{2\hbar}X^2+\frac{1}{2m\omega\hbar}P^2-\frac{1}{2}.$$

Then $H=\hbar\omega(N+\frac{1}{2})$ and one checks that

$$[N,a]=a^\dagger a^2=aa^\dagger a=[a^\dagger,a]a=-a$$
$$[N,a^\dagger]=a^\dagger aa^\dagger-a^\dagger a^\dagger a=a^\dagger[a,a^\dagger]=a^\dagger.$$

Let $\mathfrak{h}=\mathbb{C}\langle 1,N,a,a^\dagger\rangle$. This is a 4-dimensional Lie algebra under the commutator bracket, called the **Heisenberg algebra**.

Suppose ψ is an eigenvector for N, say $N\psi = e\psi$. Then

$$N(a\psi) = (aN - a)\psi = a(N-1)\psi = (e-1)a\psi$$

and likewise

$$N(a^\dagger\psi) = (a^\dagger N + a^\dagger)\psi = a^\dagger(N+1)\psi = (e+1)a^\dagger\psi.$$

Based on this, we get the following infinite ladder:

$$\begin{tikzcd}
(e\text{-eigenspace of } N) \arrow[d, shift right, "a"'] \\
((e-1)\text{-eigenspace of } N) \arrow[u, shift right, "a^\dagger"'] \arrow[d, shift right, "a"'] \\
((e-2)\text{-eigenspace of } N) \arrow[u, shift right, "a^\dagger"']
\end{tikzcd}$$

In Lie theory a and $a^\dagger$ are usually called lowering and raising operators; in physics they are usually called annihilation and creation operators.

Since the eigenvalues of H are all nonnegative, the eigenvalues of N must all be at least $-\frac{1}{2}$ (since $H = \hbar\omega(N + \frac{1}{2})$). So no matter what e is, if one goes low enough in the ladder then the eigenspaces must be zero. It's useful to consider the lowest nonzero eigenspace, which will necessarily be killed by a. But note that if $a\psi = 0$ then $N\psi = 0$ since $N = a^\dagger a$, and conversely if $N\psi = 0$ then $a\psi$ lies in the -1-eigenspace for N and is therefore zero. So $a\psi = 0 \iff N\psi = 0$, which tells us that the lowest nonzero eigenspace in the ladder (if such a thing exists) must be the 0-eigenspace. And from this it follows that the only nonzero eigenspaces satisfy $e \in \mathbb{Z}_{\geq 0}$.

At any spot of the ladder other than the 0-eigenspace, the operator N is an isomorphism. But $N = a^\dagger a$, and so the a leaving this spot is injective and the $a^\dagger$ coming into this spot is surjective. Likewise $aa^\dagger = 1 + a^\dagger a = 1 + N$, and this is an isomorphism at every spot; so the a leaving each spot is surjective, and the $a^\dagger$ coming into each spot is injective. This verifies that from the 0-eigenspace up the ladder, all the maps a and $a^\dagger$ are isomorphisms.

The above arguments show that everything reduces to understanding the 0-eigenspace of N, which is the same as the 0-eigenspace of a. The operator a is a first-order differential operator, and so it is easy to find this eigenspace. The differential equation is

$$0 = a\psi = \alpha x\psi(x) + i\beta(-i\hbar)\psi'(x), \qquad \text{or} \qquad \psi'(x) = -\frac{\alpha}{\beta\hbar}x\psi(x).$$

This is a separable equation that is readily solved to give $\psi(x) = Ce^{-\frac{\alpha}{2\beta\hbar}x^2} = Ce^{-\frac{m\omega}{2\hbar}x^2}$.

We conclude that for $n \in \mathbb{Z}_{\geq 0}$ the n-eigenspace of N is 1-dimensional and spanned by $(a^\dagger)^n(e^{-\frac{m\omega}{2\hbar}x^2})$. To better understand this wave function it will be convenient to introduce a change of coordinates:

$$u = \sqrt{\frac{\alpha}{\beta\hbar}}x = \sqrt{\frac{m\omega}{\hbar}}x, \qquad \frac{dx}{du} = \sqrt{\frac{\beta\hbar}{\alpha}}, \qquad \frac{d}{du} = \sqrt{\frac{\beta\hbar}{\alpha}}\frac{d}{dx}.$$

Then we are looking at

$$\begin{aligned}(a^\dagger)^n\big(e^{-\frac{\alpha}{2\beta\hbar}x^2}\big) &= \left(\alpha\sqrt{\frac{\beta\hbar}{\alpha}}u - \beta i(-i\hbar)\sqrt{\frac{\alpha}{\beta\hbar}}\frac{d}{du}\right)^n e^{-\frac{u^2}{2}} \\ &= \big(\sqrt{\alpha\beta\hbar}\big)^n\Big(u - \frac{d}{du}\Big)^n e^{-\frac{u^2}{2}} = \Big(\frac{1}{2}\Big)^{\frac{n}{2}}\Big(u - \frac{d}{du}\Big)^n e^{-\frac{u^2}{2}}.\end{aligned}$$

Write $(u - \frac{d}{du})^n e^{-\frac{u^2}{2}} = H_n(u)e^{-\frac{u^2}{2}}$ where $H_n(u)$ denotes a polynomial in u. This polynomial is called the ***n*th Hermite polynomial**. Clearly $H_0(u) = 1$. Also

$$H_{n+1}(u)e^{-\frac{u^2}{2}} = \Big(u - \frac{d}{du}\Big)\Big(H_n(u)e^{-\frac{u^2}{2}}\Big) = \big[2uH_n(u) - H_n'(u)\big]e^{-\frac{u^2}{2}}.$$

So we get the recursive formula $H_{n+1}(u) = 2uH_n(u) - H_n'(u)$, and a simple induction gives another recursive formula

$$H_{n+1}(u) = 2uH_n(u) - 2nH_{n-1}(u).$$

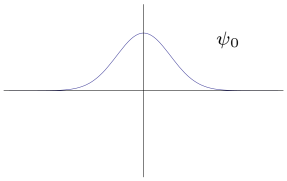
One then quickly computes

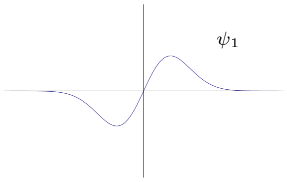
$$\begin{gathered}H_0(u) = 1, \qquad H_1(u) = 2u, \qquad H_2(u) = 4u^2 - 2, \\ H_3(u) = 8u^3 - 12u, \qquad H_4(u) = 16u^4 - 48u^2 + 12.\end{gathered}$$

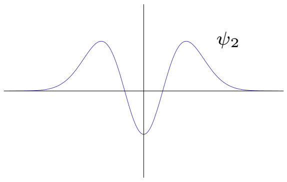
Let ψ_n be the wave function

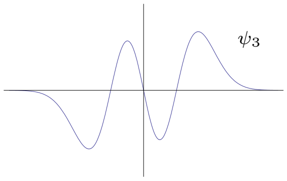
$$\psi_n(x) = C_n(a^\dagger)^n(e^{-\frac{m\omega}{2\hbar}x^2}) = C_n \cdot 2^{-\frac{n}{2}}H_n(u)e^{-\frac{u^2}{2}} = 2^{-\frac{n}{2}}C_nH_n\Big(\sqrt{\tfrac{m\omega}{\hbar}}x\Big)e^{-(\frac{m\omega}{2\hbar})x^2}$$

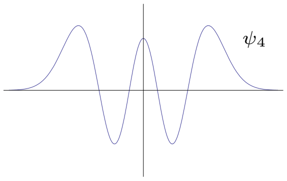
where C_n is a normalization constant that we will determine below. This is our basis for the n-eigenspace of N. Since $H = \hbar\omega(N + \frac{1}{2})$ this is also the eigenspace of H for the eigenvalue $\hbar\omega(n + \frac{1}{2})$. Here are graphs of ψ_n for $0 \leq n \leq 4$:











It is easy to compute the expectation values $\langle X\rangle$ and $\langle P\rangle$ in each state ψ_n. Indeed,

$$\langle X\rangle_n = \int \overline{\psi_n(x)}x\psi_n(x)dx = \int x\psi_n(x)^2dx = 0.$$

The last equality holds because the Hermite polynomials are either even or odd, therefore their squares are even, therefore $x\psi_n(x)^2$ is an odd function. Likewise, using integration by parts we get

$$\langle P\rangle_n = \int \psi_n(x)(-i\hbar)\frac{\partial}{\partial x}\psi_n(x)dx = -i\hbar\frac{(\psi_n(x))^2}{2}\Big|_{-\infty}^{\infty} = 0$$

since $\psi_n(x)$ tends to zero as $x \to \pm\infty$.

Below we will find $(\Delta X)_n$ and $(\Delta P)_n$, but before doing so it is useful to make some observations. We start with a simple calculation:

$$\begin{aligned}\langle a^\dagger\psi_n(x), a^\dagger\psi_n(x)\rangle &= \langle \psi_n(x), aa^\dagger\psi_n(x)\rangle\\ &= \langle \psi_n(x), a^\dagger a\psi_n(x) + [a, a^\dagger]\psi_n(x)\rangle\\ &= \langle \psi_n(x), (N+1)\psi_n(x)\rangle\\ &= \langle \psi_n(x), (n+1)\psi_n(x)\rangle\\ &= (n+1)\langle \psi_n(x), \psi_n(x)\rangle.\end{aligned}$$

So if $\psi_n(x)$ is normalized then we normalize $a^\dagger\psi_n$ by dividing by $\sqrt{n+1}$. That is to say, $\psi_{n+1} = \frac{1}{\sqrt{n+1}}a^\dagger\psi_n$. Applying a to both sides and using $aa^\dagger = N+1$, we get $a\psi_{n+1} = \sqrt{n+1}\psi_n$. Summarizing these (and reindexing), we have

$$\boxed{\begin{aligned} a\psi_n &= \sqrt{n}\,\psi_{n-1}\\ a^\dagger\psi_n &= \sqrt{n+1}\,\psi_{n+1}\end{aligned}}$$

We can use these formulas to determine the normalization constant C_n. The relation $\psi_{n+1} = \frac{1}{\sqrt{n+1}}a^\dagger\psi_n$ shows that $C_{n+1} = \frac{1}{\sqrt{n+1}}C_n$. The constant C_0 normalizes the Gaussian wave function $e^{-\frac{m\omega}{2\hbar}x^2}$, and is therefore equal to $\left(\frac{m\omega}{\pi\hbar}\right)^{\frac{1}{4}}$. Thus, we arrive at

$$\psi_n(x) = \frac{1}{\sqrt{n!}}2^{-n/2}\left(\frac{m\omega}{\hbar\pi}\right)^{1/4}\cdot H_n\left(\sqrt{\frac{m\omega}{\hbar}}\,x\right)e^{-\frac{m\omega}{2\hbar}x^2}.$$

(This identification of C_n will not be needed below, however).

Now we are ready to compute the uncertainties:

Proposition 3.3.11. *$(\Delta X)_n = \sqrt{\left(n+\frac{1}{2}\right)\frac{\hbar}{m\omega}}$, $(\Delta P)_n = \sqrt{\left(n+\frac{1}{2}\right)m\omega\hbar}$.*

Note, as a consequence, that $(\Delta X)_n \cdot (\Delta P)_n = (n+\frac{1}{2})\hbar \geq \frac{\hbar}{2}$, once again confirming the Uncertainty Principle.

Proof. The hard way to go about this is by a direct computation of the integral $\int \overline{\psi_n(x)}x^2\psi_n(x)dx$. An easier method comes from playing around with the operators a and $a^\dagger$.

Observe that $X = (a + a^\dagger)\sqrt{\frac{\hbar}{2m\omega}}$ and $P = -i(a - a^\dagger)\sqrt{\frac{m\omega\hbar}{2}}$. We have

$$\begin{aligned}
\langle X^2\rangle_n = \langle\psi_n|X^2|\psi_n\rangle &= \frac{\hbar}{2m\omega}\langle\psi_n|(a+a^\dagger)^2|\psi_n\rangle \\
&= \frac{\hbar}{2m\omega}\big\langle(a+a^\dagger)\psi_n\big|(a+a^\dagger)\psi_n\big\rangle \\
&= \frac{\hbar}{2m\omega}\Big\langle\sqrt{n}\psi_{n-1}+\sqrt{n+1}\psi_{n+1}\Big|\sqrt{n}\psi_{n-1}+\sqrt{n+1}\psi_{n+1}\Big\rangle \\
&= \frac{\hbar}{2m\omega}\big(n+(n+1)\big) \\
&= \Big(n+\frac{1}{2}\Big)\frac{\hbar}{m\omega}.
\end{aligned}$$

We have used that $a + a^\dagger$ is self-adjoint, and that the ψ_n's are orthonormal.

The computation for $(\Delta P)_n$ is entirely similar. □

Exercise 3.3.12. Compute the probability amplitudes $\langle p|\psi_n\rangle$, as a function of p. Again, the hard way to do this is by a direct computation of $\int \frac{e^{-ixp/\hbar}}{\sqrt{2\pi\hbar}}\psi_n(x)dx$. An easier way is to try to use the operators a and $a^\dagger$. Begin by verifying that $a|p\rangle = i\sqrt{\frac{2}{m\omega\hbar}}|p\rangle + a^\dagger|p\rangle$. Set $F_n(p) = \langle p|\psi_n\rangle$, and prove the recursion relation

$$F_{n+1}(p) = -ip\sqrt{\frac{2}{m\omega\hbar(n+1)}}F_n(p) + \sqrt{\frac{n}{n+1}}F_{n-1}(p).$$

Do a Gaussian integral to determine $F_0(p) = De^{-Gp^2}$ for certain constants D and G, then use this to obtain the first few $F_n(p)$'s.

3.4. Introduction to path integrals

Suppose we are considering a quantum mechanical system with Hamiltonian H. In a typical problem we will have an initial state ψ and we want to understand the time evolution $\psi_t = e^{-\frac{iHt}{\hbar}}\psi$. In our treatment of the free particle we saw that we could write

$$\psi_t(x') = \int_x U(x',t;x,0)\psi(x) \tag{3.4.1}$$

where $U(x',t;x,0)$ (the "propagator") is the probability amplitude for a particle at $(x,0)$ to be detected at (x',t). One could also write $U(x',t;x,0) = \langle x'|e^{-\frac{iHt}{\hbar}}|x\rangle$ which helps make (3.4.1) clearer:

$$\begin{aligned}
\psi_t(x') = \langle x'|\psi_t\rangle = \langle x'|e^{-\frac{iHt}{\hbar}}\psi\rangle = \langle x'|e^{-\frac{iHt}{\hbar}}|\psi\rangle &= \int_x \langle x'|e^{-\frac{iHt}{\hbar}}|x\rangle\langle x|\psi\rangle \\
&= \int_x \langle x'|e^{-\frac{iHt}{\hbar}}|x\rangle\psi(x).
\end{aligned}$$

Our goal will be to better understand the propagator $\langle x'|e^{-iHt/\hbar}|x\rangle$. We do this by breaking it up into n "smaller" probability computations, which we can estimate to first order in t. Write $\Delta t = \frac{t}{n}$. Then $e^{-\frac{iHt}{\hbar}} = \Big(e^{-\frac{iH\Delta t}{\hbar}}\Big)^n$. So we have

$$\begin{aligned}\langle x'|e^{-iHt/\hbar}|x\rangle &= \langle x'|e^{-iH\Delta t/\hbar}\cdots e^{-iH\Delta t/\hbar}|x\rangle \\ &= \int_{\mathbf{x}} \langle x'|e^{-iH\Delta t/\hbar}|x_{n-1}\rangle\langle x_{n-1}|e^{-iH\Delta t/\hbar}|x_{n-2}\rangle\cdots\langle x_1|e^{-iH\Delta t/\hbar}|x\rangle\end{aligned}$$

where the integral is over $\mathbf{x} = (x_1, \ldots, x_{n-1}) \in \mathbb{R}^{n-1}$. Here we have inserted a complete set of states $|x_i\rangle\langle x_i|$ in $n-1$ different places. The above integral adds up the probability amplitudes over all ways to take n steps from x to x'. It will be convenient to write $x = x_0$ and $x' = x_n$.

Now let us assume that our Hamiltonian has the form $H = \frac{P^2}{2m} + V(x)$. Since P and $V(x)$ will not commute we do not have $e^{-\frac{iH\Delta t}{\hbar}} = e^{-\frac{iP^2\Delta t}{2m\hbar}} \cdot e^{-\frac{iV(x)\Delta t}{\hbar}}$, but these two expressions do coincide to first order in Δt (just write out the power series and check it). We will therefore write

$$\begin{aligned}\langle x_{j+1}|e^{-iH\Delta t/\hbar}|x_j\rangle &= \langle x_{j+1}|e^{\frac{-iP^2}{2\hbar m}\Delta t}e^{\frac{-i}{\hbar}V(x)\Delta t}|x_j\rangle + \mathcal{O}(\Delta t^2) \\ &= \langle x_{j+1}|e^{\frac{-iP^2}{2\hbar m}\Delta t}|x_j\rangle e^{\frac{-i}{\hbar}V(x_j)\Delta t} + \mathcal{O}(\Delta t^2)\end{aligned}$$

since $|x_j\rangle$ is an eigenvector for $V(x)$. Now inserting $\int_p |p\rangle\langle p| = 1$ we get that

$$\begin{aligned}\langle x_{j+1}|e^{\frac{-iP^2}{2\hbar m}\Delta t}|x_j\rangle &= \int_p \langle x_{j+1}|e^{\frac{-iP^2}{2\hbar m}\Delta t}|p\rangle\langle p|x_j\rangle && (3.4.2) \\ &= \int_p \langle x_{j+1}|p\rangle\langle p|x_j\rangle e^{\frac{-ip^2}{2\hbar m}\Delta t} \\ &= \int_p \frac{e^{\frac{ip}{\hbar}(x_{j+1}-x_j)}}{2\pi\hbar} e^{\frac{-ip^2}{2\hbar m}\Delta t} \\ &= \frac{1}{2\pi\hbar}\sqrt{\frac{2\pi m\hbar}{i\Delta t}}\, e^{-\left[\frac{(x_{j+1}-x_j)^2}{4\hbar^2}\frac{2\hbar m}{i\Delta t}\right]} && (3.4.3)\end{aligned}$$

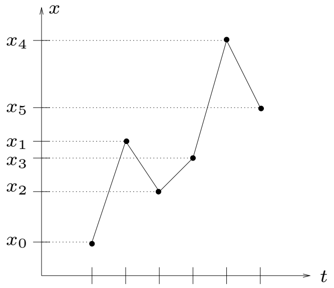
where the third equality uses our Schrödinger Axiom and for the last equality we just computed the Gaussian integral. Putting everything together, we have

$$\langle x_{j+1}|e^{-iH\Delta t/\hbar}|x_j\rangle \approx \sqrt{\frac{m}{2\pi\hbar i\Delta t}} \cdot e^{\frac{i}{\hbar}\left[\frac{m}{2}\left(\frac{x_{j+1}-x_j}{\Delta t}\right)^2\Delta t - V(x_j)\Delta t\right]}.$$

We observe that the term in brackets approximates the action for constant-velocity path of time interval Δt from x_j to x_{j+1}, provided that Δt is small. So we can write

$$\begin{aligned}&\langle x'|e^{-iH\Delta t/\hbar}|x_{n-1}\rangle\langle x_{n-1}|e^{-iH\Delta t/\hbar}|x_{n-2}\rangle\cdots\langle x_1|e^{-iH\Delta t/\hbar}|x\rangle \\ &\qquad \approx \left(\frac{m}{2\pi hi\Delta t}\right)^{\frac{n}{2}} e^{\frac{i}{\hbar}\left[\frac{m}{2}\sum\left(\frac{x_{j+1}-x_j}{\Delta t}\right)^2\Delta t - \sum V(x_j)\Delta t\right]} \\ &\qquad \approx \left(\frac{m}{2\pi hi\Delta t}\right)^{\frac{n}{2}} e^{\frac{i}{\hbar}S(\gamma_{\mathbf{x}})}\end{aligned}$$

where $\gamma_{\mathbf{x}}$ is the piecewise-linear path depicted below.



(each marked time interval has length Δt).

Finally, taking all possibilities for $\mathbf{x} = (x_1, \dots, x_{n-1})$ into account, we get

$$\langle x'|e^{-\frac{iH}{\hbar}t}|x\rangle \approx \Big(\frac{m}{2\pi hi\Delta t}\Big)^{\frac{n}{2}} \int_{\mathbf{x}\in\mathbb{R}^{n-1}} e^{\frac{i}{\hbar}S(\gamma_{\mathbf{x}})}. \tag{3.4.4}$$

Up to this point everything we have said is basically sensible from a mathematical viewpoint, but now we are going to say something that doesn't make much sense. Feynman tells us to consider the limit as $n \to \infty$ and $\Delta t \to 0$. This doesn't quite make sense, in part because the constant in front of the integral blows up. Disregarding this, Feynman's intuition is that as $n \to \infty$ the piecewise-linear paths approximate *all* paths; therefore we should regard the left-hand-side of (3.4.4) as an integral over the space of all paths (where "path" probably means "smooth path" here). Feynman would have us write things like

$$\langle x'|e^{-\frac{iH}{\hbar}t}|x\rangle = \mathcal{N}\int_{\gamma\colon x\rightsquigarrow x'} e^{\frac{i}{\hbar}S(\gamma)}D\gamma \tag{3.4.5}$$

where we imagine $D\gamma$ to be some appropriate measure on the space of all paths, and $\mathcal{N}$ is some appropriate normalization constant. This is called a "path integral." Despite the lack of rigorous mathematical foundation here, physicists have learned to use these path integrals to provide quite a bit of important intuition. They also have accumulated a large bag of magic tricks for calculating such integrals.

Remark 3.4.6. There are two important remarks that should be made right away:

(1) In working up to (3.4.4) we had to calculate a Gaussian integral with respect to p (this was back in (3.4.2)). The convergence of that integral is based on the quadratic part in p, namely $e^{-\frac{i}{2m\hbar}\Delta t p^2}$, and this integral converges only when $Re\left(\frac{i}{2m\hbar}\Delta t\right) > 0$. If we formally write $\Delta t = a + bi$ then we need $b < 0$. So the integral is only well-defined when Δt is assumed to have a small, negative imaginary component. Or said differently, for real Δt we are taking the Gaussian integral as being defined by analytic continuation from the lower half of the complex plane.

(2) The oscillatory integrand in the path integral gives an appealing way for understanding the relationship between classical and quantum mechanics. The integrand is a unit complex number, and as it winds once around the circle one gets complete cancellation and a total contribution of zero. For small $\hbar$ the integrand will wind around the circle extremely quickly, leading to lots of cancellation. So if γ is an arbitrary path, moving γ slightly leads to lots of cancellations in the integral and therefore no overall

contribution. This happens **except** when γ is a critical point of S. So as $\hbar \to 0$ we have that the path integral is concentrated at the critical paths, and the other paths do not contribute to the physics. This is precisely what classical mechanics says happens.

3.4.7. Path Integrals for a Free Particle. In the last section we did some manipulations with probability amplitudes to "derive" the path integral expression. This was useful from a pedagogical perspective, but what is more typical in physics is to start with the path integral heuristic, decide on an ad hoc basis how to assign it some mathematical meaning, and then use this to make computations of probability amplitudes. We will give an example of this in the case of the free particle.

We will need some simple generalizations of Gaussian integrals:

Proposition 3.4.8.

(a) $\displaystyle\int e^{-a[x_2-x_1]^2}e^{-b[x_1-x_0]^2}\,dx_1 = e^{\frac{-ab}{a+b}[x_2-x_0]^2}\sqrt{\frac{\pi}{a+b}}$

(b) $\displaystyle\int e^{-a_n[x_n-x_{n-1}]^2}e^{-a_{n-1}[x_{n-1}-x_{n-2}]^2}\cdots e^{-a_1[x_1-x_0]^2}\,dx_1dx_2\cdots dx_{n-1}$

$= e^{\frac{-\sigma_n(a_n,\ldots,a_1)}{\sigma_{n-1}(a_n,\ldots,a_1)}[x_n-x_0]^2}\sqrt{\frac{\pi^{n-1}}{\sigma_{n-1}(a_n,\ldots,a_1)}}$ *where* σ_i *denotes the ith elementary symmetric function.*

(c) $\displaystyle\iint e^{-a[(x_n-x_{n-1})^2+\cdots+(x_1-x_0)^2]}\,dx_1dx_2\cdots dx_{n-1} = e^{\frac{-a}{n}(x_n-x_0)^2}\sqrt{\frac{\pi^{n-1}}{na^{n-1}}}.$

Proof. For (a), just combine the exponentials and use Proposition 3.1.12. Part (b) follows from (a) by an induction, and (c) is a particular case of (b). □

Consider a free particle moving along $\mathbb{R}$, say starting at x at time t and ending at x' at time t'. The Lagrangian is given by $L = \frac{1}{2}m\dot{x}^2$. Then the path minimizing the action is $\gamma_{min}(s) = \left(\frac{x'-x}{t'-t}\right)(s-t)+x$. The action evaluated on this path is given by

$$\begin{aligned}S(\gamma_{min}) = \int_t^{t'} L(\gamma_{min}(s),\dot{\gamma}_{min}(s))ds &= \int_t^{t'}\frac{1}{2}m\dot{\gamma}_{min}(s)^2ds\\ &= \frac{1}{2}m\left(\frac{x'-x}{t'-t}\right)^2(t'-t)\\ &= \frac{1}{2}m\left[\frac{(x'-x)^2}{t'-t}\right].\end{aligned}$$

In the quantum nechanical world we would like to compute the probability amplitude that a particle at x will be detected at x' after $t'-t$ seconds. In other words we want to compute

$$\langle x',t'|x,t\rangle = \langle x'|e^{-\frac{iH}{\hbar}(t'-t)}|x\rangle = \mathcal{N}\int_{\text{all }\gamma} e^{\frac{i}{\hbar}S(\gamma)}D(\gamma).$$

We make sense of the path integral by breaking the paths into small enough pieces so that they look approximately linear—and above we saw how to understand the action on straight-line paths. If we break our paths γ into n pieces (where n is very large), we get the following approximation to $\langle x', t' | x, t \rangle$:

$$\begin{aligned}\mathcal{N}\int_{\text{all }\gamma} e^{\frac{i}{\hbar}S(\gamma)} D(\gamma) &\approx \mathcal{N}\int e^{\frac{im}{2\hbar}\left[\frac{(x'-x_{n-1})^2}{\Delta t}+\cdots+\frac{(x_1-x)^2}{\Delta t}\right]} dx_1 \cdots dx_{n-1} \\ &= e^{\frac{im}{2\hbar n\Delta t}(x'-x)^2}\mathcal{N}\sqrt{\frac{1}{n}\cdot\Big(\frac{2\pi\hbar\Delta t}{im}\Big)^{n-1}} \quad \text{(using part (c) of (3.4.8)).}\end{aligned}$$

Again we run into problems with this expression not converging for n large, but let's ignore this by pulling all the constants together; then also using that $t' - t = n\Delta t$ we write

$$\langle x', t' | x, t \rangle = Ce^{\frac{im(x'-x)^2}{2\hbar(t'-t)}}$$

where C is a mysterious constant which we will assume does not depend on x and x'. Amazingly, this is a correct result—as we know because we already derived it, using different methods, back in (3.3.6). Recall that there we found

$$\langle x', t' | x, t \rangle = \sqrt{\frac{m}{2\pi i\hbar(t'-t)}}e^{\frac{im(x'-x)^2}{2\hbar(t'-t)}}.$$

It is somewhat typical of path integral methods to have this kind of undetermined constant left in the end. Sometimes one can determine he constant by other techniques, and sometimes one can get by without knowing it. We will talk about these issues much more when we get to quantum field theory.

CHAPTER 4

Maxwell's equations

The mathematical theory of electromagnetic phenomena is one of the great achievements of 19th century physics. It makes for a wonderful story, too: the experiment-based laws obtained by Faraday and Ampère are unified by Maxwell into four fundamental equations, and from these equations one predicts the existence of electromagnetic waves. Given the importance that such waves assumed in so much of twentieth century life, this was really a watershed event.

In the next section we describe Maxwell's equations. It will turn out, though, that these equations are a little unsatisfactory: they do not take on the same form in all coordinate systems. To remedy this one uses differential forms, which we review below. This leads to a "coordinate free" version of Maxwell's equations which is quite general: it can be studied on any manifold with a non-degenerate metric.

This is still not the end of the story, however. Quantum-mechanical considerations lead to another way that Maxwell's equations are unsatisfactory. The magnetic potential, which is a convenient but non-essential tool from the point of view of Maxwell's equations, assumes a much greater significance in the quantum theory. Gauge theory incorporates the magnetic potential at a more fundamental level, building it into the geometry via bundles and connections.

In the next several sections we develop this story. There is quite a bit of mathematical machinery one has to digest here: differential forms, affine connections, principal bundles, curvature, and so on. We attempt to explain some motivation for each new tool before we introduce it, and then after introducing it we try to explain how it helps us out.

4.1. Maxwell's equations

The basic objects we wish to study are electric and magnetic fields. The former concept is easier to explain, because it can be directly measured. If we are in a room where there are various electric charges, currents, and magnets lying around, a charge q placed at point (x, y, z) will have a certain force $\mathbf{F}$ exerted on it. Experiment shows that if one places a charge of $2q$ at the same point, the exerted force is twice as much, and so on. The electric field $\mathbf{E}$ at this point is defined to be $\frac{\mathbf{F}}{q}$. At any moment in time the electric field is therefore a function $\mathbb{R}^3 \to \mathbb{R}^3$. To take into account changes of the field with time, we regard $\mathbf{E}$ as a function $\mathbf{E}(x, y, z, t)$, i.e., as a function $\mathbb{R}^4 \to \mathbb{R}^3$.

The magnetic field is more difficult to describe, because we do not have magnetic charges that we can use to measure it. We know such fields must exist, because we can move a compass around the room and find that it is deflected through different angles at different places and different moments. We also could observe that if an electric charge q is at point (x, y, z) and is *moving* with velocity $\mathbf{v}$, then the force exerted on it will not be the same as if the particle were sitting still. It took quite a bit of experimentation to begin to guess a mathematical form for these forces. For our present purposes we just jump to the punchline: the magnetic field $\mathbf{B}$ is a function $\mathbb{R}^4 \to \mathbb{R}^3$, and it is connected to measurable phenomena by the *Lorentz force law*:

$$\mathbf{F} = q\mathbf{E} + q(\mathbf{v} \times \mathbf{B})$$

for a particle with charge q at a point (x, y, z) moving with velocity $\mathbf{v}$.

We now give Maxwell's equations for the electric and magnetic fields. For these we need a function $\rho(x, y, z, t)$ giving the charge density at each point in space (and each moment in time), and a function $\mathbf{J}(x, y, z, t)$ giving the current density vector. The meaning of charge density is clear enough, but what is the current density vector? It is characterized by the property that if $\mathbf{u}$ is unit vector and one imagines a small screen of area dA placed at (x, y, z, t) and perpendicular to $\mathbf{u}$, then $\mathbf{J} \cdot \mathbf{u}\, dA$ is the total amount of charge moving across the screen per unit time. One can also write $\mathbf{J}(x, y, z, t) = \rho(x, y, z, t)\mathbf{v}_{\text{avg}}(x, y, z, t)$, where $\mathbf{v}_{\text{avg}}$ is the average charge velocity.

Given a region V in $\mathbb{R}^3$, the integral $\iint\limits_{\partial V} (\mathbf{J} \cdot \hat{\mathbf{n}})dS$ is therefore the total charge leaving the region per unit time. But this quantity can also be computed as $-\frac{d}{dt}\left(\iiint\limits_V \rho\, dV\right)$. Using the Divergence Theorem, we therefore have

$$\iiint\limits_V (\nabla \cdot \mathbf{J})dV = \iint\limits_{\partial V} (\mathbf{J} \cdot \hat{\mathbf{n}})dS = -\frac{d}{dt}\left(\iiint\limits_V \rho\, dV\right) = \iiint\limits_V \left(\frac{\partial \rho}{\partial t}\right) dV.$$

Since this holds for any region V, it must be that $\nabla \cdot \mathbf{J} = -\frac{\partial \rho}{\partial t}$.

Here are Maxwell's equations in their so-called *differential form* (or *local form*):

(1) $\nabla \cdot \mathbf{E} = \frac{1}{\epsilon_0}\rho$

(2) $\nabla \cdot \mathbf{B} = 0$ (no magnetic monopoles)

(3) $\nabla \times \mathbf{E} = -\dfrac{\partial \mathbf{B}}{\partial t}$ (Faraday's law)

(4) $\nabla \times \mathbf{B} = \mu_0\epsilon_0 \dfrac{\partial \mathbf{E}}{\partial t} + \mu_0 \mathbf{J}$ (Ampère-Maxwell law).

The ϵ_0 and μ_0 appearing here are certain constants of nature, called the **electric constant** and **magnetic constant**, respectively. The former is also called the *permittivity of free space*, and the latter is the *permeability of free space*. More generally, any material has a permittivity and permeability—these measure how easy or hard it is for electric/magnetic fields to form within that material.

We wish to discuss the physical meaning of each of Maxwell's equations, but this is easier to do if we give the equations in a more global form. Essentially one just uses Stokes's Theorem and the Divergence Theorem to recast (1)–(4) as equations about integrals. For this, let V be any solid, compact region in $\mathbb{R}^3$ (the closure of a bounded open set), and let S be any compact surface in $\mathbb{R}^3$. The *integral form* (or *global form*) of Maxwell's equations becomes:

$$(1)\qquad \oint\!\!\!\oint_{\partial V} \mathbf{E}\cdot\mathbf{n}\,dS = \tfrac{1}{\epsilon_0}\iiint_V \rho\,dV = \tfrac{1}{\epsilon_0}(\text{total charge inside of } V).$$

$$(2)\qquad \oint\!\!\!\oint_{\partial V} \mathbf{B}\cdot\mathbf{n}\,dV = 0.$$

$$(3)\qquad \oint_{\partial S} \mathbf{E}\cdot\mathbf{ds} = \iint_S -\Big(\frac{\partial\mathbf{B}}{\partial t}\Big)\cdot\mathbf{n}\,dS.$$

$$(4)\qquad \oint_{\partial S} \mathbf{B}\cdot\mathbf{ds} = \iint_S \Big(\mu_0\epsilon_0\frac{\partial\mathbf{E}}{\partial t} + \mu_0\mathbf{J}\Big)\cdot\mathbf{n}\,dS.$$

For each equation let us explain why it is equivalent to the corresponding differential form, and also explain its physical significance.

Equation (1): The Divergence Theorem gives $\oint\!\!\oint_{\partial V} \mathbf{E}\cdot\mathbf{n}\,dS = \iiint_V (\nabla\cdot\mathbf{E})\,dV$, and this gives the equivalence between the differential and integral forms of (1).

In words, equation (1) says that the total electric flux through any closed surface is proportional to the total amount of charge inside the surface. This is called **Gauss's Law**.

Equation (2): The two forms are again equivalent by the Divergence Theorem. The equation says that there are no magnetic charges.

Equation (3): Stokes's Theorem says that $\oint_{\partial S}\mathbf{E}\cdot\mathbf{ds} = \oint\!\!\oint_S (\nabla\times\mathbf{E})\cdot\mathbf{n}\,dS$. This immediately gives the equivalence of the two forms.

Equation (3) is Faraday's Law, which says that a changing magnetic field induces an electric current. The reader might have seen a demonstration of this where one moves a bar magnet back and forth through a hoop of wire connected to an ammeter (a device that measures current). While the magnet is moving perpendicularly to the hoop, there is a current. If the magnet stops, the current dissipates. If the magnet moves parallel to the plane of the hoop, there may still be a current but it will be much smaller (because the magnetic flux across the hoop is not changing very much).

Equation (4): The two forms are again equivalent by Stokes's Theorem. This equation is called the Ampère-Maxwell law, and it has an interesting story. Ampère's Law states that an electric current through a wire creates a magnetic field circulating around the wire. On its own this would suggest the (equivalent) equations

$$\nabla\times\mathbf{B} = \mu_0\mathbf{J} \qquad \text{or} \qquad \oint_{\partial S}\mathbf{B}\cdot\mathbf{ds} = \mu_0\iint_S \mathbf{J}\cdot\mathbf{n}\,dS.$$

However, the divergence of a curl is always zero; so the above equations would lead to

$$0 = \nabla\cdot(\nabla\times\mathbf{B}) = \nabla\cdot(\mu_0\mathbf{J}) = -\mu_0\frac{\partial\rho}{\partial t}.$$

That is, Ampère's Law can only be valid when the total charge at each point is constant. Maxwell realized that by adding a new term into Ampère's Law he could remove this obstruction. Indeed, if we have $\nabla\times\mathbf{B} = \mu_0\mathbf{J} + \mathbf{A}$ then we will get

$$0 = \nabla\cdot(\nabla\times\mathbf{B}) = -\mu_0\frac{\partial\rho}{\partial t} + (\nabla\cdot\mathbf{A}).$$

If we take $\mathbf{A} = \mu_0\epsilon_0\frac{\partial \mathbf{E}}{\partial t}$ then the above identity holds by the first of Maxwell's equations.

It is this new term that Maxwell added to Ampère's Law which gives the existence of electromagnetic waves. For simplicity assume we are in free space, so that that $\rho = 0$ and $\mathbf{J} = 0$. Then equation (4) becomes $\nabla \times \mathbf{B} = \mu_0\epsilon_0\frac{\partial \mathbf{E}}{\partial t}$. We then get

$$\begin{aligned}\mu_0\epsilon_0\frac{\partial^2\mathbf{E}}{\partial t^2} = \frac{\partial}{\partial t}\big(\nabla \times \mathbf{B}\big) = \nabla \times \Big(\frac{\partial \mathbf{B}}{\partial t}\Big) &= -\nabla \times (\nabla \times \mathbf{E}) \\ &= -(\nabla(\nabla \cdot \mathbf{E}) - \triangle\mathbf{E}) \\ &= \triangle\mathbf{E}.\end{aligned}$$

Here $\triangle$ is the Laplacian operator $\triangle = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$, and the fourth equality is a generic identity that is readily checked. The fifth equality comes from $\nabla \cdot \mathbf{E} = 0$, using the first Maxwell equation and $\rho = 0$.

The point is that $\triangle\mathbf{E} = (\mu_0\epsilon_0)\frac{\partial^2\mathbf{E}}{\partial t^2}$ is the wave equation—for a wave travelling with velocity equal to $\frac{1}{\sqrt{\mu_0\epsilon_0}}$. A similar analysis gives the same equation but with $\mathbf{E}$ replaced by $\mathbf{B}$. The constants μ_0 and ϵ_0 can be measured experimentally by studying charges, currents, and magnets, so the speed of these waves can be computed—and it turns out to agree with the speed of light! Maxwell was thereby led to the hypothesis that light is an electromagnetic wave.

Remark 4.1.1 (Electromagnetic duality). One of the things that is most apparent from Maxwell's equations is that they are *almost* symmetric in $\mathbf{E}$ and $\mathbf{B}$. The difference is that while there are electric charges in the universe, as far as we know there are no magnetic charges. If there *were* magnetic charges then one would expect a magnetic charge density appearing in equation (2) and a magnetic current density term in equation (4)—this would make things completely symmetric in $\mathbf{E}$ and $\mathbf{B}$. We will return from time to time to this idea of electromagnetic duality.

4.1.2. Invariance under coordinate changes. It is basically self-evident that Maxwell's equations are invariant under translations and rotations of 3-space—that is, they take on exactly the same form if one performs such a coordinate change. Essentially this is just because the divergence and curl have this property. Said differently, the circulation and flux integrals appearing in the integral form of Maxwell's equations are purely geometric invariants of the vector fields $\mathbf{E}$ and $\mathbf{B}$; they are not going to be changed by translating or rotating the vector fields.

Let Aff denote the group of rigid motions of $\mathbb{R}^3$. The group of translations (isomorphic to $\mathbb{R}^3$) is a normal subgroup, and one has the exact sequence

$$1 \to \mathbb{R}^3 \to \text{Aff} \to SO(3) \to 1.$$

Said differently, Aff is a semi-direct product $\text{Aff} = \mathbb{R}^3 \rtimes SO(3)$. Note that Aff is a 6-dimensional, non-compact Lie group. What we have said so far is that Maxwell's equations are invariant under coordinate changes in Aff.

Maxwell's equations are, however, *not* invariant under the so-called "Gallilean coordinate changes." Imagine we have two observers—let us call them Calvin and Hobbes—and that Hobbes is moving with a constant velocity $\mathbf{v}$ with respect to Calvin. Without loss of generality let us assume that the observers have chosen coordinate systems which line up in the y and z directions, and where the velocity $\mathbf{v}$ is occurring only in the x-direction. Let us also assume that the observers cross paths when $t = 0$ according to both of their clocks. Let (x, y, z, t) denote Calvin's coordinate system, and (x', y', z', t') denote Hobbes's. The Gallilean coordinate change is

$$x = x' + vt', \quad y = y', \quad z = z', \quad t = t'.$$

The easiest way to see that Maxwell's equations are not invariant under such coordinate changes is to think about electromagnetic waves. According to Maxwell, electromagnetic waves in free space propogate with a speed c that is determined by the constants of nature μ_0 and ϵ_0. Assuming these constants will be measured to be the same in all experiments done by either Calvin or Hobbes, it follows that both observers will experience electromagnetic waves as travelling with the same velocity c. But if Calvin sends an electromagnetic wave travelling in the positive x-direction, a Gallilean coordinate change clearly shows that Hobbes will observe the wave approaching him with speed $c - v$. This is an apparent contradiction.

At this point we jump ahead many years in our story, past Maxwell to Einstein, Poincaré, Lorentz, and Minkowski. Their solution to this puzzle is to say that Gallilean transformations just do not preserve the laws of physics—so we can't use them. A different kind of transformation must be used in changing from Calvin's coordinate system to Hobbes's, the so-called Lorentz transformations. We will give a very brief tour through this, from a modern perspective.

We start by defining a new inner product on $\mathbb{R}^4$, called the **Minkowski inner product**. This is given by

$$\langle (x, y, z, t), (x', y', z', t') \rangle = xx' + yy' + zz' - c^2 tt'.$$

Here c is the speed of light. Mathematically, it will be convenient for us to choose units so that $c = 1$; for instance, choose the unit of distance to be the light-year and the unit of time to be the year. While this makes the mathematics less cumbersome in places, it also has the negative effect of removing some physical intuition from certain formulas. From time to time we will "put the c's back in" to help illustrate a point.

So now our Minkowski inner product has

$$\langle (x, y, z, t), (x', y', x', t') \rangle = xx' + yy' + zz' - tt'.$$

Define $O(3, 1)$ to be the symmetry group of this form. That is,

$$O(3, 1) = \{P \in GL_4(\mathbb{R}) \mid \langle P\underline{v}, P\underline{w} \rangle = \langle \underline{v}, \underline{w} \rangle \ \ \forall \underline{v}, \underline{w} \in \mathbb{R}^4\}.$$

This is called the **Lorentz group**. Note that we could also write that $O(3, 1) = \{P \in GL_4(\mathbb{R}) | P^T \cdot D \cdot P = D\}$, where D is the diagonal matrix with diagonal entries $1, 1, 1, -1$. Thus, the elements of $O(3, 1)$ are exactly those elements of $GL_4(\mathbb{R})$ of the form $\left[\begin{array}{c|c|c|c} v_1 & v_2 & v_3 & v_4 \end{array}\right]$ with $v_1, v_2, v_3, v_4 \in \mathbb{R}^4$ and

$$\langle v_i, v_j \rangle = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \text{ and } 1 \leq i \leq 3 \\ -1 & \text{if } i = j = 4. \end{cases}$$

Here are some important kinds of elements of $O(3, 1)$:

(1) Any matrix $\left[\begin{array}{c|c} A & 0 \\ \hline 0 & 1 \end{array}\right]$ where $A \in O(3)$. In fact, via this identification $O(3)$ is a subgroup of $O(3, 1)$.

(2) The element $T = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & \text{-1} \end{bmatrix}$, called the "time reversal" operator.

(3) Any matrix of the form $\begin{bmatrix} \alpha & 0 & 0 & \beta \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \beta & 0 & 0 & \alpha \end{bmatrix}$, where $\alpha^2 - \beta^2 = 1$. Such an element is called an "x-boost." Similarly, there are y-boosts and z-boosts, represented by the evident matrices.

Fact: All elements of $O(3,1)$ are products (not necessarily of length two!) of elements of $O(3)$, T, and boosts.

This fact yields, in particular, that $O(3,1)$ is 6-dimensional: three dimensions for the $O(3)$ part, and three dimensions worth of boosts.

We also would like to understand the number of components of $O(3,1)$. The determinant gives a map $\det\colon O(3,1) \longrightarrow \{-1,1\}$, since if $P \in O(3,1)$ then $P^T \cdot D \cdot P = D$ and so $\det P^2 = 1$. This map is clearly surjective, and so $O(3,1)$ has at least two components.

If $P \in O(3,1)$ then

$$-1 = \langle (P_{14}, P_{24}, P_{34}, P_{44}), (P_{14}, P_{24}, P_{34}, P_{44}) \rangle = P_{14}^2 + P_{24}^2 + P_{34}^2 - P_{44}^2.$$

Therefore $P_{44}^2 \geq 1$. So the map $P \mapsto P_{44}$ is a map $O(3,1) \to \mathbb{R} - \{(-1,1)\}$. One readily sees that this map is also surjective (all we really care about here is surjectivity on π_0, though).

Exercise 4.1.3. Show that the map $O(3,1) \longrightarrow \{-1,1\} \times \big(\mathbb{R} - (-1,1)\big)$ given by $P \mapsto (\det P, P_{44})$ is surjective on π_0. Consequently, $O(3,1)$ has at least four components.

We won't prove this, but it turns out that $O(3,1)$ has *exactly* four components. Define $SO^+(3,1)$ to be the connected component of the identity in $O(3,1)$. This is called the **restricted Lorentz group**. Elements of $SO^+(3,1)$ are products of boosts and elements of $SO(3)$.

The Lorentz group consists of all linear maps of $\mathbb{R}^4$ that preserve the Minkowski form. Just as we did for $\mathbb{R}^3$, it is also convenient to consider the affine maps that preserve the form—that is, compositions of Lorentz transformations with translations. This is called the **Poincaré group** P. There is an exact sequence

$$1 \to \mathbb{R}^4 \to P \to O(3,1) \to 1,$$

or we could write $P = \mathbb{R}^4 \rtimes O(3,1)$. Note that P is ten-dimensional. The preimage of $SO^+(3,1)$ under $P \to O(3,1)$ is called the **restricted Poincaré group**, and we will denote this as P^+.

After getting through all of these definitions we can finally get to the point. Special relativity says that the laws of physics should be invariant under the action of the group P^+. One sometimes sees the full Poincaré group here instead, but let us stick with the more conservative statement for now. Special relativity also tells us the specific transformation to use when comparing coordinate systems moving with constant velocity with respect to each other. Returning to Calvin and Hobbes, where Hobbes moves with constant velocity $\mathbf{v}$ in the x-direction with respect to Calvin, the coordinate change is

$$\begin{aligned} x' &= \gamma(x - vt) \\ y' &= y \\ z' &= z \\ t' &= \gamma(t - vx) \end{aligned}$$

where $\gamma = (1-v^2)^{-\frac{1}{2}}$ (recall that the primed coordinate system belongs to Hobbes). That is,

$$\begin{bmatrix} x' \\ y' \\ z' \\ t' \end{bmatrix} = \begin{bmatrix} \gamma & 0 & 0 & -\gamma v \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\gamma v & 0 & 0 & \gamma \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \\ t \end{bmatrix}.$$

Note that the matrix appearing in this formula is an x-boost, as we have the formula $\gamma^2 - (\gamma v)^2 = 1$ by the definition of γ.

Remark 4.1.4. There is an easy way to remember the matrix in the above formula. First, one remembers that motion in the x-direction doesn't affect the y- and z-coordinates. Next one thinks about the corresponding Gallilean transformation, which would have $x' = x - vt$ or a $\begin{bmatrix} 1 & 0 & 0 & -v \end{bmatrix}$ across the top row of the matrix. Completing this to a Lorentz transformation suggests having $\begin{bmatrix} -v & 0 & 0 & 1 \end{bmatrix}$ along the bottom row, but this isn't quite right because then the columns do not have the appropriate Minkowski norms. So this is all fixed by multiplying the top and bottom rows by a constant γ, and then the norm condition forces what γ has to be.

Returning now to Maxwell's equations, the question we should have asked at the beginning was whether these equations are invariant under the action of P^+. The invariance under translations is clear, so the question is really about $SO^+(3,1)$. The $SO(3)$ part is clear, so we further reduce to asking about invariance under boosts.

Now, it turns out that Maxwell's equations **are** invariant under boosts. But something a little surprising happens. In Maxwell's equations, the roles of space and time are very different—or *separate*. But the Lorentz boosts mix up space and time. At first this seems like a bad sign: how could Maxwell's equations be invariant if this is happening? The answer is that there is only one way: the Lorentz transformations have to *also* mix up $\mathbf{E}$ and $\mathbf{B}$. This is a strange idea. It means that the thing Calvin measures as the electric field $\mathbf{E}$ is *not* going to be what Hobbes measures as the electric field $\mathbf{E}$. In other words, the electric field (and likewise the magnetic field) are not really physically invariant concepts!

This is perhaps a sign that we are missing the appropriate language for talking about these concepts. Rather than $\mathbf{E}$ and $\mathbf{B}$ being separate entities, we need to find a language where they appear together as parts of a common object. The language of differential forms solves this problem. This is getting a little ahead of ourselves, but in that language the "common object" is something called the *electromagnetic 2-tensor* $\mathcal{F}$. Maxwell's equations become simply

$$d\mathcal{F} = 0, \qquad *d*\mathcal{F} = \mathcal{J}.$$

We will explain what all this means in the next couple of sections. See Section 4.3.1 for a more detailed discussion of how **E** and **B** get intermixed under Lorentz transformations.

4.2. Differential forms

For some reason it is hard to teach differential forms without them seeming hopelessly complicated and intimidating. In this section we give the basic definitions, together with what motivation we can provide, but this material doesn't really congeal into you start to use it for something. That will take place beginning in Section 4.3.

4.2.1. Motivation. In basic multivariable calculus you are taught about divergence, gradient, and curl. It's usually not presented this way, but this is the first time most people are exposed to a chain complex:

$$0 \to C^\infty(\mathbb{R}^3) \xrightarrow{grad} (\text{Vec. fields on } \mathbb{R}^3) \xrightarrow{curl} (\text{Vec. fields on } \mathbb{R}^3) \xrightarrow{div} C^\infty(\mathbb{R}^3) \to 0.$$

In the first and second spots we mean *smooth* vector fields (by convention we start our numbering at the left, with $C^\infty(\mathbb{R}^3)$ in the 0th spot—so really we are thinking of this as a *cochain* complex).

Not only is this a chain complex, but one learns in calculus that it is exact everywhere except at the 0th spot. For example, if the curl of a field on $\mathbb{R}^3$ is zero then it is a gradient field—this shows that H^1 is zero. (In fact, the only way I can remember the basic relations between div, grad, and curl is by remembering the above cochain complex!) At the 0th spot the homology is not zero: the kernel of grad consists of just the constant functions, and so the 0th homology is isomorphic to $\mathbb{R}$.

Now, we can look at the same complex with an open subset $U \subseteq \mathbb{R}^3$ substituted for $\mathbb{R}^3$ everywhere:

$$0 \to C^\infty(U) \xrightarrow{grad} (\text{Vec. fields on } U) \xrightarrow{curl} (\text{Vec. fields on } U) \xrightarrow{div} C^\infty(U) \to 0.$$

Call this complex $C^\bullet(U)$. Examples show that it might not be exact at places where the original was. For instance, if $U = \mathbb{R}^3 - 0$ then consider the radial vector field $\mathbf{F}(x,y,z) = \frac{1}{(x^2+y^2+z^2)^{3/2}}(x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}})$. The divergence of this field is everywhere zero. But $\iint_S \mathbf{F}\cdot\hat{\mathbf{n}}\,dS = 4\pi$ where S is the unit sphere in $\mathbb{R}^3$, and so Stokes's Theorem shows that $\mathbf{F}$ cannot be the curl of anything. Therefore $C^\bullet(U)$ has some homology in the second spot. With a little legwork one sees that this is the *only* homology—with the exception of the constant functions in the 0th spot, of course—and so $H^0(C^\bullet) = H^2(C^\bullet) = \mathbb{R}$ and $H^1(C^\bullet) = H^3(C^\bullet) = 0$.

It doesn't take long to notice that in the two examples we have done, $U = \mathbb{R}^3$ and $U = \mathbb{R}^3 - 0$, we recovered the singular cohomology of U. It turns out that this is always the case: for any open subset $U \subseteq \mathbb{R}^3$,

$$H^i(C_\bullet(U)) \cong H^i_{\text{sing}}(U;\mathbb{R}) \qquad \text{(de Rham Theorem).}$$

Your experience with homology has perhaps left you jaded about this kind of thing, but it is good to take a moment and marvel at how amazing it must seem to someone first coming to this subject. Start from whatever definition of singular cohomology you like best—cellular, simplicial, or singular—and realize that you are now being told you can calculate it a completely different way, not using any kind of simplices or cells but rather by using *calculus*!

Of course there is an immediate question that springs to mind: we did this for $U \subseteq \mathbb{R}^3$, but what about $U \subseteq \mathbb{R}^n$? Unfortunately it is not so clear how to write down the cochain complex in this case: the first two terms are straightforward enough, but then what happens? I claim that if your goal is to make this generalization, and you sit down and try to do it, you will ultimately be led to differential forms.

4.2.2. The first example of a differential form. We will need a few things from the language of vector bundles. A **vector bundle** is (basically) a fiber bundle where the fibers are vector spaces. For example, given a smooth manifold M of dimension n, the tangent bundle TM is a vector bundle over M. All of the natural constructions you know for vector spaces carry over to vector bundles: given two vector bundles E and F over M, one can form vector bundles $E \oplus F, E \otimes F$, the dual vector bundle E^*, etc.

Given a vector bundle $p : E \longrightarrow M$, define the **space of (smooth) sections** of E to be

$$\Gamma(E) = \{s\colon M \longrightarrow E \,|\, p \circ s = \text{id and } s \text{ is smooth}\}.$$

Note that this is a real vector space: using the addition and scalar multiplication in the fibers of E, we can add and scale sections. (In fact not only is $\Gamma(E)$ a real vector space, it is a module over the ring $C^\infty(M)$).

We define the cotangent bundle of M to be the dual of the tangent bundle, and we use the notation $T^*M = (TM)^*$.

Definition 4.2.3. *The vector space of* ***differential*** 1-***forms*** *on* M *is* $\Omega^1(M) = \Gamma(T^*M)$.

This definition is short and sweet, easy to remember, but I think this is a place where things get fuzzy for someone learning this for the first time. Because how does one “picture” a section of the cotangent bundle, and why was this the “right” definition to choose in the first place?

We answer these questions by a basic observation: Given a smooth manifold M, if we have $v \in T_xM$ and $f \in C^\infty(M)$, then we should be able to come up with a directional derivative $\partial_v f \in \mathbb{R}$. As soon as one has the idea of doing this, it is clear how to define it: pick any smooth curve $\gamma\colon (-1,1) \longrightarrow M$ such that $\gamma(0) = x$ and $\gamma'(0) = v$, then define

$$(\partial_v f)_x = \frac{d}{dt}\Big(f(\gamma(t))\Big)\bigg|_{t=0}.$$

One needs to prove that this is well-defined, but this is an easy check. One also checks that this construction is linear in v, as one would expect: so we actually have a linear map $\partial_{(-)}f\colon T_xM \longrightarrow \mathbb{R}$. Since we've done this for every $x \in M$, we have produced exactly a section of the cotangent bundle T^*M!

The point I want to make is that it's useful to think of 1-forms not just as what they *are*, but as what they *do for you*: 1-forms are processes that take any point $x \in M$ and any tangent vector $v \in T_xM$ and produce a real number. The canonical example is the “directional derivative” process described above. Any smooth function $f\colon M \to \mathbb{R}$ has an associated 1-form called df, defined by

$$df_x(v) = (\partial_v f)_x.$$

As to the question of why Definition 4.2.3 was the "right" definition, the answer is that it is simply the natural home of the directional derivative process.

REMARK 4.2.4. Let $f\colon \mathbb{R}^3 \longrightarrow \mathbb{R}$ and $x, v \in \mathbb{R}^3$. In multivariable calculus one learns that $(\partial_v f)(x) = (\vec{\nabla} f)(x) \cdot v$, and afterwards one thinks more about the gradient than directional derivatives. Why didn't we follow this procedure for more general manifolds above? The reason is that the gradient vectors exist in $\mathbb{R}^3$ only because of the dot product, which induces an isomorphism $\mathbb{R}^3 \longrightarrow (\mathbb{R}^3)^*$ given by $v \mapsto (x \mapsto x \cdot v)$. Our directional derivative $(\partial_{\underline{v}} f)_x$ lives in $(\mathbb{R}^3)^*$, but using this isomorphism there is a corresponding vector in $\mathbb{R}^3$—and this is the gradient.

On a general smooth manifold M one lacks the dot product, and so one can only look at directional derivatives themselves. On a *Riemannian* manifold, however, the story is different; because there one does have a metric, and this can be used to define gradient fields. We will have more to say about this a little later.

To summarize what has happened in this section, there is really only one important point: we defined the space of 1-forms $\Omega^1(M)$ and we produced a map

$$d\colon C^\infty(M) \to \Omega^1(M), \qquad f \mapsto df.$$

Here df is to be thought of as something like "all of the directional derivatives of f, bundled together".

4.2.5. Exterior products and Hodge duals. The way I've decided to proceed is to do all of the algebra together, in one swoop, before progressing from 1-forms to n-forms. That is our goal in the present section. I suppose this is because for me algebra always seems "easy", whereas geometry is "hard". So temporarily wipe anything about the geometry of manifolds from your brain, and let V be a real vector space of dimension n.

Let $V^{\otimes k}$ denote the k-th tensor power of V, which has dimension n^k. There is a Σ_k action on $V^{\otimes k}$ defined by

$$\sigma(v_1 \otimes v_2 \otimes \cdots \otimes v_k) = (-1)^\sigma v_{\sigma(1)} \otimes v_{\sigma(2)} \otimes \cdots \otimes v_{\sigma(k)},$$

where $(-1)^\sigma$ denotes the sign of σ. Define the **space of invariants** by

$$\left[V^{\otimes k}\right]^{\Sigma_k} = \{\omega \in V^{\otimes k} \mid \sigma(\omega) = \omega,\ \forall \sigma \in \Sigma_k\}.$$

Let $j\colon \left[V^{\otimes k}\right]^{\Sigma_k} \hookrightarrow V^{\otimes k}$ denote the inclusion. There is a natural retraction $\rho\colon V^{\otimes k} \to \left[V^{\otimes k}\right]^{\Sigma_k}$ given by

$$\rho(\omega) = \frac{1}{k!}\left(\sum_{\sigma \in \Sigma_k} \sigma(\omega)\right).$$

It is easy to see that $\rho \circ j = \mathrm{id}$.

We set $V^{\otimes k}/\Sigma_k = V^{\otimes k}/\langle \sigma\omega - \omega \mid \omega \in V^{\otimes k}, \sigma \in \Sigma_k\}$. This is called the **orbit space** or **coinvariants** of Σ_k acting on $V^{\otimes k}$. We have the following diagram:

$$\begin{tikzcd}
\left[V^{\otimes k}\right]^{\Sigma_k} \arrow[r, tail, "j"'] \arrow[rr, bend left, "\tilde{j}"] & V^{\otimes k} \arrow[r, two heads] \arrow[d, "\rho"'] & V^{\otimes k}/\Sigma_k \arrow[dl, dotted, "\tilde{\rho}"] \\
& \left[V^{\otimes k}\right]^{\Sigma_k} &
\end{tikzcd}$$

Here $\tilde{j}$ is the composite across the top row, and $\tilde{\rho}$ is induced from ρ using that $\rho(\sigma\omega) = \rho(\omega)$. One can easily check that $\tilde{\rho} \circ \tilde{j} = \mathrm{id}$ and $\tilde{j} \circ \tilde{\rho} = \mathrm{id}$. So the spaces of invariants and coinvariants are isomorphic.

Analysts tend to define $\bigwedge^k(V)$ as $[V^{\otimes k}]^{\Sigma_k}$, whereas algebraists tend to define it as $[V^{\otimes k}]_{\Sigma_k}$. Over characteristic zero fields like $\mathbb{R}$ it makes no difference, because the spaces are isomorphic (we needed characteristic zero when we divided by $k!$ in our definition of ρ). [As an aside, over characteristic p fields the invariants $[V^{\otimes k}]^{\Sigma_k}$ are denoted $\Gamma^k(V)$ and called the **divided powers** of V. When one attempts to do calculus in charcteristic p, as in algebraic de Rham or crystalline cohomology, these divided powers play a big role.]

For our present purposes we will stick with the analysts' convention of defining $\bigwedge^k(V) = [V^{\otimes k}]^{\Sigma_k}$. Given $v_1, v_2 \ldots, v_k \in V$, we define

$$v_1 \wedge v_2 \wedge \cdots \wedge v_k = \rho(v_1 \otimes v_2 \otimes \cdots \otimes v_k).$$

Exercise 4.2.6. Show that if $e_1, e_2, \ldots, e_n$ is a basis for V, then

$$\{e_{i_1} \wedge e_{i_2} \wedge \ldots \wedge e_{i_k} \,|\, 1 \leq i_1 < i_2 < \cdots < i_k \leq n\}$$

is a basis for $\bigwedge^k(V)$, and thus dim $\bigwedge^k(V) = \binom{n}{k}$.

Remark 4.2.7. Note that $\bigwedge^n(V) \cong \mathbb{R}$, but not canonically. That is, there is no natural choice of basis for $\bigwedge^n(V)$.

Induced bilinear forms. Now suppose V has a symmetric bilinear form $\langle -, - \rangle$. There is an induced form on $V^{\otimes k}$ defined by taking

$$\langle v_1 \otimes v_2 \otimes \cdots \otimes v_k, w_1 \otimes w_2 \otimes \cdots \otimes w_k \rangle = \prod_{i=1}^{k} \langle v_i, w_i \rangle$$

and extending linearly. Since $\bigwedge^k(V) \hookrightarrow V^{\otimes k}$, the exterior product inherits a form by restriction. From the definition of ρ we have

$$\begin{aligned}
&\langle v_1 \wedge \cdots \wedge v_k, w_1 \wedge \cdots \wedge w_k \rangle = \\
&\qquad \sum_{\sigma, \theta \in \Sigma_k} \frac{(-1)^\sigma (-1)^\theta}{(k!)^2} \langle v_{\sigma(1)} \otimes v_{\sigma(2)} \otimes \cdots \otimes v_{\sigma(k)}, w_{\theta(1)} \otimes w_{\theta(2)} \otimes \cdots \otimes w_{\theta(k)} \rangle \\
&\qquad = \sum_{\theta \in \Sigma_k} \frac{(-1)^\theta}{k!} \langle v_1 \otimes v_2 \otimes \cdots \otimes v_k, w_{\theta(1)} \otimes w_{\theta(2)} \otimes \cdots \otimes w_{\theta(k)} \rangle \\
&\qquad = \frac{1}{k!} \det(\langle v_i, w_j \rangle)_{i,j}.
\end{aligned}$$

The second equality comes from the fact that all the terms where $\sigma = \theta$ are the same, or more generally if $g \in \Sigma_k$ then the terms of the sum where $\theta = g\sigma$ are all the same.

Analysts find it convenient to throw away the $\frac{1}{k!}$ appearing in the inner product of k-forms. This bothers me a little, as one then has to remember that the inner products on $\bigwedge^k(V)$ and $V^{\otimes k}$ are inconsistent. But we will again follow the analysts here: so we redefine the bilinear form on $\bigwedge^k V$ to be

$$\langle v_1 \wedge v_2 \wedge \cdots \wedge v_k, w_1 \wedge w_2 \wedge \cdots \wedge w_k \rangle = \det(\langle v_i, w_j \rangle)_{i,j}$$

Example 4.2.8. Let e_1, e_2, e_3 be an orthonormal basis for V. Then

$$\langle e_1 \wedge e_2, e_2 \wedge e_3 \rangle = \det \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = 0, \qquad \langle e_1 \wedge e_2, e_1 \wedge e_2 \rangle = \det \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = 1.$$

In general, if $e_1, e_2, \ldots, e_n$ is an orthonormal basis for V then the associated basis $\{e_{i_1} \wedge e_{i_2} \wedge \ldots \wedge e_{i_k}\}$ for $\bigwedge^k(V)$ is also orthonormal.

If the original form on V is nondegenerate, then we get an induced isomorphism

$$V \xrightarrow{\cong} V^*, \qquad v \mapsto \langle v, - \rangle.$$

Let α denote the inverse isomorphism. Using this, we can transplant the form on V to a form on V^*: for $f, g \in V^*$ define

$$\langle f, g \rangle = \langle \alpha(f), \alpha(g) \rangle.$$

This form on V^* then yields induced forms on $\bigwedge^k(V^*)$ as discussed above. In short, if V has a nondegenerate symmetric bilinear form then there is an induced form on all of the usual "natural" constructions one can think to apply to V.

Hodge $*$-operator. Note that $\bigwedge^k(V)$ and $\bigwedge^{n-k}(V)$ have the same dimension, since $\binom{n}{k} = \binom{n}{n-k}$. So $\bigwedge^k(V) \cong \bigwedge^{n-k}(V)$, but not canonically. Fix a nondegenerate symmetric bilinear form on V and a nonzero element $\mathrm{vol} \in \bigwedge^n(V)$. It turns out that there is a unique isomorphism $\bigwedge^k(V) \to \bigwedge^{n-k}(V)$, denoted $\omega \mapsto *\omega$, which satisfies the property

$$\mu \wedge (*\omega) = \langle \mu, \omega \rangle \, \mathrm{vol}$$

for all $\mu, \omega \in \bigwedge^k(V)$. The proof of this can be safely left as an exercise, but it will also become clear after the examples we do below. This isomorphism is called the **Hodge $*$-operator**.

Example 4.2.9. Consider $V = \mathbb{R}^3$ with the dot product, and $\mathrm{vol} = e_x \wedge e_y \wedge e_z$. If we want to compute $*e_x$ we note that $*e_x \in \bigwedge^2(V)$ and

$$\begin{aligned} e_x \wedge (*e_x) &= 1 \cdot \mathrm{vol} \\ e_y \wedge (*e_x) &= 0 \\ e_z \wedge (*e_x) &= 0. \end{aligned}$$

So $*e_x = e_y \wedge e_z$.

Exercise 4.2.10. Consider the setting from the above example. If $u, v \in \mathbb{R}^3$ check that $*(u \wedge v) = u \times v$ where $\times$ denotes the cross product.

In general, suppose $e_1, e_2, \ldots, e_n$ is an orthogonal basis for V (not necessarily othonormal!) and take $\mathrm{vol} = e_1 \wedge \cdots \wedge e_n$. Write

$$\{e_1, e_2, \ldots, e_n\} = \{e_{i_1}, e_{i_2}, \ldots, e_{i_k}\} \amalg \{e_{j_1}, e_{j_2}, \ldots, e_{j_{n-k}}\}.$$

It is clear from our examples that $*(e_{i_1}\wedge\cdots\wedge e_{i_k})$ is a scalar multiple of $e_{j_1}\wedge\cdots\wedge e_{j_k}$, and it is just a matter of determining the scalar. In fact it is not hard to prove that

(4.2.11)

$$*(e_{i_1}\wedge\cdots\wedge e_{i_k}) = \left(\frac{\text{vol}}{e_{i_1}\wedge\cdots\wedge e_{i_k}\wedge e_{j_1}\wedge\cdots\wedge e_{j_{n-k}}}\right)\left(\prod_{r=1}^{k}\langle e_{i_r}, e_{i_r}\rangle\right)e_{j_1}\wedge\cdots\wedge e_{j_{n-k}}.$$

The first fraction to the right of the equals sign requires some explanation. Here both the numerator and the denominator are nonzero elements of the one-dimensional vector space $\bigwedge^n(V)$, therefore they are scalar multiples of each other: our fraction notation is just meant to indicate the appropriate scalar. Incidentally, the above formula verifies the uniqueness of the $*$-operator, and as a definition it also serves to prove the existence.

Example 4.2.12. Consider $\mathbb{R}^4$ with the Minkowski product and $\text{vol} = e_x\wedge e_y\wedge e_z\wedge e_t$. Using the above formula one finds that

$$\begin{aligned} *e_t &= (-1)(-1)(e_x\wedge e_y\wedge e_z) = e_x\wedge e_y\wedge e_z\\ *(e_y\wedge e_z) &= 1\cdot 1\, e_x\wedge e_t = e_x\wedge e_t\\ *(e_x\wedge e_t) &= 1\cdot -1\, e_y\wedge e_z = -e_y\wedge e_z. \end{aligned}$$

Corollary 4.2.13. *If $\omega\in\bigwedge^k(V)$, then $*(*\omega) = (-1)^{k(n-k)}(-1)^x\omega$ where x is the number of -1's in signature of the original bilinear form on V.*

Proof. There exists an orthogonal basis for V where each $\langle e_j, e_j\rangle$ is 1 or -1. It suffices to check the identity for $\omega = e_{i_1}\wedge\cdots\wedge e_{i_k}$, and this is done by applying (4.2.11) twice. □

We have now completed our detour through pure algebra, and we can go back to thinking about geometry.

4.2.14. Differential k-forms. Recall that any natural construction we can make for vector spaces extends to something we can also do for vector *bundles*. In particular, if $E\to X$ is a vector bundle then we can form its exterior powers $\bigwedge^k(E)\to X$.

Definition 4.2.15. *Let M be a smooth manifold. The space of **differential k-forms** on M is $\Omega^k(M) = \Gamma(\bigwedge^k(T^*M))$.*

Once again, based merely on this definition the precise nature of what a k-form really *is* can seem a bit obtuse. There are a few things to keep in mind to help with this:

(1) A k-form is a process ω which to every $x\in M$ and vectors $v_1,\ldots,v_k\in T_xM$ assigns a real number $\omega_x(v_1,\ldots,v_k)$, and this real number must be alternating and linear in the v_i's.
(2) Suppose $x^1,\ldots,x^n$ are local coordinates on M. Then $dx^1,\ldots,dx^n$ give a local trivialization of T^*M, and hence the forms $dx^{i_1}\wedge\cdots\wedge dx^{i_k}$ for $i_1<\cdots<i_k$ give a local trivialization of $\bigwedge^k(T^*M)$. This means that any k-form on M can be written locally as

(4.2.16)
$$\omega = \sum_{i_1,\ldots,i_k} f_{i_1,\ldots,i_k}dx^{i_1}\wedge dx^{i_2}\wedge\cdots\wedge dx^{i_k}$$

where each $f_{i_1,\ldots,i_k}$ is a smooth real-valued function defined in the coordinate neighborhood.

(3) Note that if V is a vector space then $\bigwedge^0(V)$ is canonically isomorphic to $\mathbb{R}$. So if $E \to M$ is a vector bundle then $\bigwedge^0(E) \to M$ is isomorphic to the trivial bundle $M \times \mathbb{R} \to M$. Therefore $\Omega^0(M)$ consists of the sections of the rank one trivial bundle, or in other words $\Omega^0(M) = C^\infty(M)$.

We next define the de Rham complex. We have already defined $d\colon \Omega^0(M) \to \Omega^1(M)$, and there is a unique way to extend this to maps $d\colon \Omega^k(M) \to \Omega^{k+1}(M)$ which is natural and satisfies the Leibniz rule $d(a \wedge b) = (da \wedge b) + (-1)^{|a|}(a \wedge db)$. If $\omega \in \Omega^k(M)$ is as in (4.2.16) then one is forced to define

$$d\omega = \sum_{i_1,\ldots,i_k} df_{i_1,\ldots,i_k} \wedge dx^{i_1} \wedge dx^{i_2} \wedge \cdots \wedge dx^{i_k}.$$

One readily checks that $d^2 = 0$ (this boils down to the fact that partial derivatives commute), and thus we have a chain complex. This is called the de Rham complex, and its cohomology will be denoted $H^i_{dR}(M) = H^i(\Omega^\bullet(M))$.

Note that these constructions are contravariantly functorial. If $f\colon M \to N$ is a smooth map then there are induced maps $f^*\colon \Omega^k(N) \to \Omega^k(M)$: if $\omega \in \Omega^k(N)$ then $f^*(\omega)$ is the k-form on M defined by $f^*(\omega)_x(v) = \omega_{f(x)}\big((Df_x)(v)\big)$ for $x \in M$ and $v \in T_xM$. Symbolically this is a mouthful, but it is just the evident thing. It is not hard to check that f^* induces a map of cochain complexes $\Omega^\bullet(N) \to \Omega^\bullet(M)$ and therefore yields induced maps on cohomology groups.

The de Rham theorem states that there are natural isomorphisms $H^i_{dR}(M) \cong H^i_{Sing}(M;\mathbb{R})$. It is easy enough to see how one might prove this: basically one just needs to check that $H^*_{dR}(-)$ satisfies the Mayer-Vietoris and homotopy invariance axioms. Mayer-Vietoris is easy, for if $\{U, V\}$ is an open cover of M then there is an evident short exact sequence of complexes $0 \to \Omega^\bullet(U \cap V) \to \Omega^\bullet(U) \oplus \Omega^\bullet(V) \to \Omega^\bullet(M) \to 0$. This just comes from the fact that differential forms are sections of a bundle, and hence forms defined on open sets can be patched together if they agree on the intersection. The homotopy invariance property $H^*_{dr}(M) \cong H^*_{dR}(M \times \mathbb{R})$ is not as obvious, but it is not difficult—it's essentially an exercise in calculus. See any basic text on differential forms, for instance [**BT**].

Finally, we wrap up this section by returning to the beginning of this whole discussion. Let $U \subseteq \mathbb{R}^3$ be open and consider the de Rham complex

$$0 \longrightarrow \Omega^0(U) \longrightarrow \Omega^1(U) \longrightarrow \Omega^2(U) \longrightarrow \Omega^3(U) \longrightarrow 0.$$

We claim that up to isomorphism this is precisely the grad-curl-div complex with which we began the section. To see this, note for instance that

$$\Omega^2(U) = \Gamma(\textstyle\bigwedge^2 T^*U) \cong \Gamma(\bigwedge^2 TU) \cong \Gamma(\bigwedge^1 TU) = \Gamma(TU).$$

Here the first isomorphism uses the metric on TU (i.e., the dot product in $\mathbb{R}^3$) to give an isomorphism $TU \cong T^*U$. The second isomorphism is via the Hodge $*$-operator. Finally, note that $\Gamma(TU)$ is simply the space of smooth vector fields on U. Applying these ideas to each spot of the de Rham complex, we get the following picture:

$$\begin{tikzcd}
0 \arrow[r] & \Omega^0(U) \arrow[r] \arrow[d, equal] & \Omega^1(U) \arrow[r] \arrow[d, equal] & \Omega^2(U) \arrow[r] \arrow[d, equal] & \Omega^3(U) \arrow[r] \arrow[d, equal] & 0 \\
 & C^\infty(U) \arrow[dr, dotted, "e_0"'] & \Gamma(T^*U) \arrow[d, equal, "\text{metric}"] & \Gamma(\wedge^2(T^*U)) \arrow[d, equal, "\text{metric}"] & \Gamma(\wedge^3(T^*U)) \arrow[d, equal, "\text{metric}"] & \\
 & & \Gamma(TU) \arrow[dr, dotted, "e_1"'] & \Gamma(\wedge^2(TU))] \arrow[d, equal, "\text{Hodge}"] & \Gamma(\wedge^3(TU)) \arrow[d, equal, "\text{Hodge}"] & \\
 & & & \Gamma(TU) \arrow[dr, dotted, "e_2"'] & \Gamma(\wedge^0(TU)) \arrow[d, equal] & \\
 & & & & C^\infty(U) &
\end{tikzcd}$$

The maps e_i are obtained just by chasing around the diagram (that is, they are just the obvious composites).

EXERCISE 4.2.17. Check that the maps e_0, e_1, and e_2 are in fact grad, curl, and div.

4.3. A second look at Maxwell's equations

Let $M = \mathbb{R}^4$ with the Minkowski inner product, and let $\mathbf{E}$ and $\mathbf{B}$ denote the electric and magnetic fields as usual. Recall from the end of Section 4.1 that we wanted to find an object that unified these fields somehow. Since $\mathbf{E}$ has three components, and $\mathbf{B}$ has three components, this suggests that the object we're looking for should have six degrees of freedom—which makes a form in $\bigwedge^2(\mathbb{R}^4)$ a natural choice.

With this in mind, write $\mathbf{E} = E_x\hat{\mathbf{i}} + E_y\hat{\mathbf{j}} + E_z\hat{\mathbf{k}}$ and similarly for $\mathbf{B}$. Define a 2-form $\mathfrak{f} \in \Omega^2(M)$ by

$$\mathfrak{f} = B_x(dy\wedge dz) - B_y(dx\wedge dz) + B_z(dx\wedge dy) + E_x(dx\wedge dt) + E_y(dy\wedge dt) + E_z(dz\wedge dt).$$

This is called the **electromagnetic 2-form**. For the moment put aside the objection that this definition came out of thin air, and that there were some mysterious choices here—for instance, the sign change on the B_y term. Instead of trying to explain these things let's simply do a revealing computation:

$$\begin{aligned} d\mathfrak{f} = \frac{\partial B_x}{\partial x} dx \wedge dy \wedge dz - \frac{\partial B_y}{\partial y} dy \wedge dx \wedge dz + \frac{\partial B_z}{\partial z} dz \wedge dx \wedge dy \\ + \frac{\partial B_x}{\partial t} dt \wedge dy \wedge dz - \frac{\partial B_y}{\partial t} dt \wedge dx \wedge dz + \frac{\partial B_z}{\partial t} dt \wedge dx \wedge dy \\ + \left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} \right) dx \wedge dy \wedge dt + \left(\frac{\partial E_z}{\partial x} - \frac{\partial E_x}{\partial z} \right) dx \wedge dz \wedge dt \\ + \left(\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} \right) dy \wedge dz \wedge dt \end{aligned}$$

$$\begin{aligned} = (\nabla \cdot \mathbf{B})\, dx\, dy\, dz + \left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} + \frac{\partial B_z}{\partial t} \right) dx\, dy\, dt \\ + \left(\frac{\partial E_z}{\partial x} - \frac{\partial E_x}{\partial z} - \frac{\partial B_y}{\partial t} \right) dx\, dz\, dt + \left(\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} + \frac{\partial B_x}{\partial t} \right) dy\, dz\, dt. \end{aligned}$$

We can now observe that $d\mathfrak{f} = 0$ if and only if $\nabla \cdot \mathbf{B} = 0$ and $\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$. So $d\mathfrak{f} = 0$ is a simple, compact, and coordinate-free way of encapsulating two of Maxwell's equations! We will see in just a moment that one can obtain the other two Maxwell equations by considering $*\mathfrak{f}$ and $d(*\mathfrak{f})$.

Note quite generally that if M is a n-manifold with a smooth metric, then the metric gives us an isomorphism $TM \xrightarrow{\cong} T^*M$ and hence induced isomorphisms $\bigwedge^k(TM) \xrightarrow{\cong} \bigwedge^k(T^*M)$. We will call this "metric duality". Furthermore, if M is oriented then we have the Hodge $*$-operator giving us an isomorphism of bundles

$$\textstyle\bigwedge^k(TM) \underset{*}{\xrightarrow{\cong}} \bigwedge^{n-k}(TM) \qquad \text{and} \qquad \bigwedge^k(T^*M) \underset{*}{\xrightarrow{\cong}} \bigwedge^{n-k}(T^*M).$$

We'll call these isomorphisms "Hodge duality".

Returning to $M = \mathbb{R}^4$ with the Minkowski metric, the metric duals of $\mathbf{E}$ and $\mathbf{B}$ are the 1-forms

$$\mathcal{B} = B_x\, dx + B_y\, dy + B_z\, dz \qquad \text{and} \qquad \mathcal{E} = E_x\, dx + E_y\, dy + E_z\, dz.$$

Using (4.2.11), compute that

$$\begin{aligned} *(dx \wedge dt) &= \quad\ \ 1 \cdot (-1) dy \wedge dz = -dy \wedge dz \\ *(dx \wedge dt) &= (-1) \cdot (-1) dx \wedge dz = \quad dx \wedge dz \\ *(dx \wedge dt) &= \quad\ \ 1 \cdot (-1) dx \wedge dy = -dx \wedge dy \end{aligned}$$

It follows at once that our electromagnetic 2-form $\mathfrak{f}$ is

$$\mathfrak{f} = \mathcal{E} \wedge dt - *(\mathcal{B} \wedge dt).$$

Applying $*$ to both sides of the above equation, we get that

$$*\mathfrak{f} = \mathcal{B} \wedge dt + *(\mathcal{E} \wedge dt).$$

Note that the difference between $\mathfrak{f}$ and $*\mathfrak{f}$ involves the change $(\mathcal{E}, \mathcal{B}) \mapsto (\mathcal{B}, -\mathcal{E})$. This is electromagnetic duality.

Writing things out in complete detail, we have

$$*\mathfrak{f} = B_x\, dx \wedge dt + B_y\, dy \wedge dt + B_z\, dz \wedge dt - E_x\, dy \wedge dz + E_y\, dx \wedge dz - E_z\, dx \wedge dy.$$

Either by doing the computation by hand, or else applying electromagnetic duality to our computation of $d\mathfrak{f}$, one finds that

$$d(*\mathfrak{f}) = -(\nabla\cdot\mathbf{E})\,dx\wedge dy\wedge dz + \left(\frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y} - \frac{\partial E_z}{\partial t}\right) dx\wedge dy\wedge dt \\ + \left(\frac{\partial B_z}{\partial x} - \frac{\partial B_x}{\partial z} + \frac{\partial E_y}{\partial t}\right) dx\wedge dz\wedge dt + \left(\frac{\partial B_z}{\partial y} - \frac{\partial B_y}{\partial z} - \frac{\partial E_x}{\partial t}\right) dy\wedge dz\wedge dt.$$

Therefore

$$*(d(*\mathfrak{f})) = -(\nabla\cdot\mathbf{E})\,dt + \left(\frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y} - \frac{\partial E_z}{\partial t}\right) dz \\ - \left(\frac{\partial B_z}{\partial x} - \frac{\partial B_x}{\partial z} + \frac{\partial E_y}{\partial t}\right) dy + \left(\frac{\partial B_z}{\partial y} - \frac{\partial B_y}{\partial z} - \frac{\partial E_x}{\partial t}\right) dx.$$

Define the current density 1-form by $\mathcal{J} = -\rho\,dt + J_x\,dx + J_y\,dy + J_z\,dz$. Note that $J_x\,dx + J_y\,dy + J_z\,dz$ is the metric dual of the current density vector field $\mathbf{J}$. Then the second two Maxwell's equations are equivalent to $*(d(*\mathfrak{f})) = \mathcal{J}$.

Putting everything together, we have found that Maxwell's equations are equivalent to the two equations

$$d\mathfrak{f} = 0 \qquad \text{and} \qquad *\,d*\,\mathfrak{f} = \mathcal{J}.$$

One thing that is very nice is that this version of the equations is totally coordinate-free. Note also that we don't have to restrict ourselves to $M = \mathbb{R}^4$; we could define an electromagnetic field on any oriented manifold M, equipped with a non-degenerate metric, to be a 2-form $\mathfrak{f} \in \Omega^2(M)$ satisyfing the above equations with respect to some fixed current density $\mathcal{J} \in \Omega^1(M)$.

4.3.1. Invariance of Maxwell's equations under coordinate change. Our motiviation for introducing the language of differential forms was that it was supposed to make the invariance of Maxwell's equations under Lorentz tranformations easier and more natural. Let us now see if this really happened.

Let $P\colon \mathbb{R}^4 \to \mathbb{R}^4$ be any smooth map, and assume $\mathfrak{f} \in \Omega^2(\mathbb{R}^4)$ satisfies Maxwell's equations. We see immediately that $d(P^*\mathfrak{f}) = P^*(d\mathfrak{f}) = 0$. So $P^*\mathfrak{f}$ automatically satisfies the first two Maxwell's equations. For the second two equations we need for P to be compatible with the $*$-operator. This will happen if P preserves the metric and the volume form. So assume $P \in O(3,1)$ and $\det P = 1$. Then $*d*\,(P^*\mathfrak{f}) = P^*(*d*\,\mathfrak{f}) = P^*\mathcal{J}$.

We have verified that if $\mathfrak{f}$ satisfies the Maxwell's equations with current density 1-form $\mathcal{J}$ then $P^*\mathfrak{f}$ satisfies the same equations with current density 1-form $P^*\mathcal{J}$. This is all there is to invariance. Hopefully this seems very easy! Having a coordinate free version of Maxwell's equations really pays off.

In some sense this is the complete story as far as coordinate-invariance goes, but it is worthwhile to consider a specific example. Let's return to our two observers Calvin and Hobbes, where Hobbes is moving with respect to Calvin with constant velocity v in the x-direction. To transform from Calvin's coordinate system to Hobbes's we are supposed to use an x-boost, namely the matrix

$$P = \begin{pmatrix} \gamma & 0 & 0 & -\gamma v \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\gamma v & 0 & 0 & \gamma \end{pmatrix}$$

where $\gamma^2 - \gamma^2 v^2 = 1$, or $\gamma^2 = \dfrac{1}{1-v^2}$. So our coordinate change is given for 1-forms by

$$\begin{pmatrix} dx' \\ dy' \\ dz' \\ dt' \end{pmatrix} = \begin{pmatrix} \gamma & 0 & 0 & -\gamma v \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\gamma v & 0 & 0 & \gamma \end{pmatrix} \cdot \begin{pmatrix} dx \\ dy \\ dz \\ dt \end{pmatrix}$$

(recall that the primed coordinate system belongs to Hobbes). The inverse of this matrix is a similar matrix in which v has been replaced with $-v$, and so

$$\begin{pmatrix} dx \\ dy \\ dz \\ dt \end{pmatrix} = \begin{pmatrix} \gamma & 0 & 0 & \gamma v \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \gamma v & 0 & 0 & \gamma \end{pmatrix} \cdot \begin{pmatrix} dx' \\ dy' \\ dz' \\ dt' \end{pmatrix}.$$

We calculate $P^*\mathfrak{f}$ by taking the definition of $\mathfrak{f}$ and substituting $dx = \gamma dx' + (\gamma v)dt'$, and similarly for dx, dy, and dz. We get

$$\begin{aligned} P^*\mathfrak{f} &= B_x\,dy'\,dz' - B_y\,(\gamma\,dx' + \gamma v\,dt')\,dz' + B_z\,(\gamma\,dx' + \gamma v\,dt')\,dy' \\ &\qquad + E_x\,(\gamma\,dx' + \gamma v\,dt')\,(\gamma v\,dx' + \gamma dt') + E_y\,dy'\,(\gamma v\,dx' + \gamma dt') \\ &\qquad + E_z\,dz'\,(\gamma v\,dx' + \gamma dt') \\ &= B_x\,dy'\,dz' - (B_y\gamma + \gamma v E_z)dx'\,dz' + (\gamma B_z - \gamma v E_y)\,dx'\,dy' \\ &\qquad + E_x\,dx'\,dt' + (\gamma E_y - \gamma v B_z)\,dy'\,dt' + (\gamma E_z + \gamma v B_y)\,dz'\,dt'. \end{aligned}$$

Examining this final formula, we have that

$$\begin{aligned} E'_x &= E_x & B'_x &= B_x \\ E'_y &= \gamma E_y - \gamma v B_z & B'_y &= \gamma B_y + \gamma v E_z \\ E'_z &= \gamma E_z + \gamma v B_y & B'_z &= \gamma B_z - \gamma v E_y \end{aligned}$$

So one sees very clearly how the electric and magnetic fields get intermixed under a Lorentz transformation.

4.3.2. The electromagentic potential. Let us first return to the original form of Maxwell's equations, with vector fields $\mathbf{E}$ and $\mathbf{B}$. Two of the equations say that

$$\nabla \cdot \mathbf{B} = 0 \qquad \text{and} \qquad \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}.$$

Since $\mathbf{B}$ has zero divergence, one has $\mathbf{B} = \nabla \times \mathbf{A}$ for some vector field $\mathbf{A}$. Such an $\mathbf{A}$ is called a **magnetic potential**, or a **vector potential**.

But now we can write

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} = -\frac{\partial}{\partial t}(\nabla \times \mathbf{A}) = -\nabla \times \frac{\partial \mathbf{A}}{\partial t},$$

or $\nabla \times \left(\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t}\right) = 0$. Therefore $\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t} = -\nabla\phi$ for some $\phi \colon \mathbb{R}^3 \to \mathbb{R}$. This ϕ is called an **electric potential**, or a **scalar potential**.

Note that neither $\mathbf{A}$ nor ϕ are unique. The magnetic potential $\mathbf{A}$ is well-defined only up to the addition of a gradient field, and the electric potential ϕ is well-defined only up to addition of constants.

These potentials can also be discussed from the differential forms perspective. Since $d\mathfrak{f} = 0$ and we are working on $\mathbb{R}^4$, we have $\mathfrak{f} = dA$ for some 1-form A (using that $H^1(\mathbb{R}^4)=0$). This A is called the **electromagnetic potential 1-form**, and is well-defined up to addition of terms dg where $g \in C^\infty(\mathbb{R}^4)$.

The 1-form A may be written in coordinates as $A = A_t\,dt + A_x\,dx + A_y\,dy + A_z\,dz$ for $A_t, A_x, A_y, A_z \in C^\infty(\mathbb{R}^4)$. Compared to the discussion in the vector field terminology, A_t is essentially the ϕ and $A_x\,dx + A_y\,dy + A_z\,dz$ is essentially the $\mathbf{A}$ (or really, its metric dual).

The point of the electromagnetic potential is that half of Maxwell's equations are easy to solve, and choosing a potential does this. Because the potential is not unique, there is a tendency to regard the potential itself as not physically significant—it is just a convenient tool, not important at the same level as the electromagnetic fields themselves. However, over the course of the next few sections we will see that the potential has a much deeper significance than this would suggest. Once quantum mechanics enters the picture the role of the potential is somehow much more fundamental, and we will find ways of building it deeper into the theory. A small sign of this heightened significance shows up in the next subject, where we construct a Lagrangian for the theory.

4.3.3. A Lagrangian approach to electromagnetism. Given a charged particle with charge q and mass m moving in an electromagnetic field, recall that the Lorentz force law says that $\mathbf{F}_{EM} = q\mathbf{E} + q(\mathbf{v}\times\mathbf{B})$. Thinking back to classical mechanics, one might have the idea of finding a Lagrangian which gives rise to the equations of motion corresponding to this force. In the case where the force could be written as the negative gradient of a potential, the Lagrangian was just $L = T - V$ where T was the kinetic energy and V was the potential. The Lorentz force law, however, is not of this form: note, for instance, that the force depends on the velocity of the particle and not just its position.

So motion in an electromagnetic field doesn't immediately fit with the simple things we know about Lagrangians and classical mechanics. Still, with a little legwork and ingenuity one *can* write down a Lagrangian which gives rise to the appropriate equations of motion. Here it is:

$$L(x_1,x_2,x_3,\dot{x}_1,\dot{x}_2,\dot{x}_3) = \left(\sum_{i=1}^{3}\frac{1}{2}m\dot{x}_i^2\right) - q\cdot\phi(x_1,x_2,x_3) + q\begin{pmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{pmatrix}\cdot\mathbf{A}(x_1,x_2,x_3).$$

This formula is suggestive of the "kinetic minus potential" paradigm for Lagrangians, but the thing being subtracted is not a true potential (note the dependence on velocity, for instance). It's still useful to think of this term as some kind of "generalized potential," however.

Let us check that the Euler-Lagrange equations for L really do recover the Lorentz force law. The Euler-Lagrange equations are

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}_i}\right) = \frac{\partial L}{\partial x_i},$$

which gives

$$\frac{d}{dt}\left(m\dot{x}_i + qA_i\right) = -q\frac{\partial \phi}{\partial x_i} + q\sum_{j=1}^{3}\dot{x}_j\frac{\partial A_j}{\partial x_i},$$

or equivalently

$$m\ddot{x}_i + q\frac{\partial A_i}{\partial t} + q\sum_{j=1}^{3}\frac{\partial A_i}{\partial x_j}\dot{x}_j = -q\frac{\partial \phi}{\partial x_i} + q\sum_{j=1}^{3}\dot{x}_j\frac{\partial A_j}{\partial x_i}.$$

So we get

$$\begin{aligned} m\ddot{x}_i &= -q\left(\frac{\partial \phi}{\partial x_i} + \frac{\partial A_i}{\partial t}\right) + q\sum_{j=1}^{3}\left(\dot{x}_j\frac{\partial A_j}{\partial x_i} - \dot{x}_j\frac{\partial A_i}{\partial x_j}\right) \\ &= -q\left(\frac{\partial \phi}{\partial x_i} - \frac{\partial A_i}{\partial t}\right) + q\left[\begin{pmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{pmatrix} \times (\nabla \times \mathbf{A})\right]_i \\ &= q\mathbf{E}_i + q\left[\begin{pmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{pmatrix} \times \mathbf{B}\right]_i = i\text{th component of } \mathbf{F}_{EM}, \end{aligned}$$

where we have written $[\mathbf{v}]_i$ for the ith component of a vector $\mathbf{v}$.

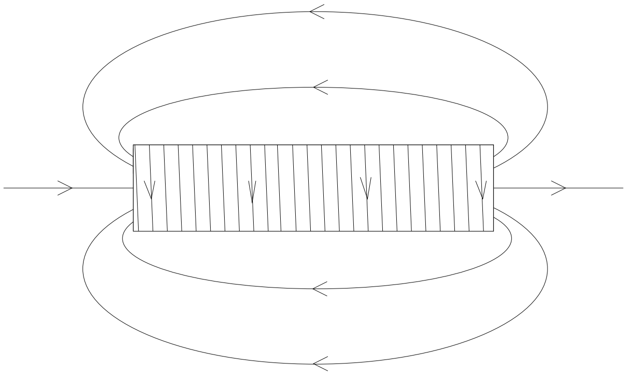
In the next section we will use the Lagrangian for electromagnetism to explain an intriguing quantum-mechanical effect.

4.4. The Aharanov-Bohm effect

In our approach to electromagnetism we started with the idea that the primary objects of study are the electric and magnetic fields $\mathbf{E}$ and $\mathbf{B}$. We later saw that in $\mathbb{R}^3$ Maxwell's equations imply the existence of an electromagnetic potential: in particular, of a magnetic potential $\mathbf{A}$ such that $\operatorname{curl}\mathbf{A} = \mathbf{B}$. This potential function is only defined up to addition of a gradient field, and because of this it seems of less primary importance in physical law. The Aharanov-Bohm effect, however, demonstrates the flaw in this point of view. It gives an example where the presence of a magnetic field affects objects *even in regions of space where the magnetic field is zero!* Quantum mechanics explains these effects in terms of the magnetic potential $\mathbf{A}$, which can be nonzero even when $\mathbf{B}$ vanishes. Our goal in this section is to explain this example.

4.4.1. The basic effect.

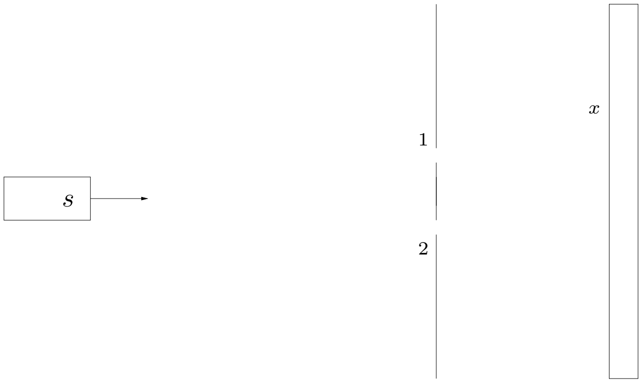
A solenoid is a hollow tube with a wire that is tightly coiled around it. As current flows through the wire it generates a magnetic field, somewhat reminiscent of that of a bar magnet (but one in the shape of a cylinder). Here is a picture:



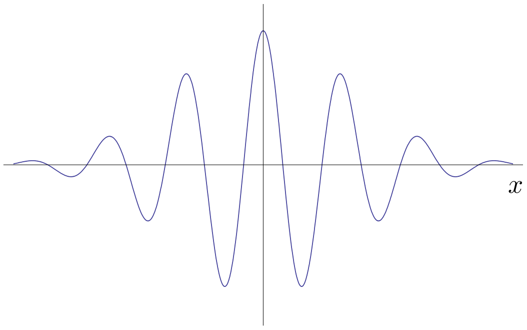
If the solenoid is lengthened, the magnetic field lines outside the cylinder become more spread apart. In the limiting case of an infinite solenoid it turns out that there is no magnetic field outside of the tube, whereas inside the tube the magnetic field is constant (pointing in a direction parallel to the tube). This can all be proven mathematically, and we will do some of the analysis below.

Even though the magnetic field is zero outside of the tube, the magnetic *potential* is nevertheless nonzero there—we will see this as part of our analysis. It turns out that one can choose a potential whose field lines circulate around the axis of the solenoid; the length of the potential vectors steadily grows inside of the tube, and then once it reaches the tube itself it stops growing and is just constant outside of the tube. Again, we will do the analysis below—for the moment just accept that this is the result.

Now consider a standard quantum-mechanical double slit experiment, as shown in the following diagram. Electrons are fired at a screen with two slits, and the electrons that make it through the slits land on a second screen:

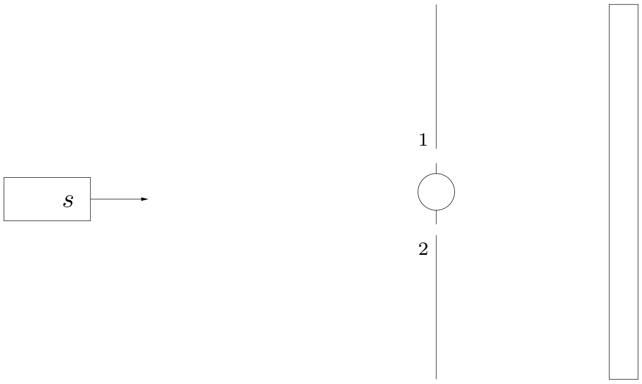


One finds that the number of electons arriving at points x on the second screen forms an interference pattern as follows:



Let us be clear what we mean by this. After firing millions of electrons at the screen and marking where they land, one would get a sequence of bands which are dense at the center and sparse near the edges. The density of the bands, as a function of x, is given by the norm square of the above graph.

Next imagine placing a solenoid in the experiment, as shown below (the tube of the solenoid is coming out of the paper):



It is important that electrons leaving a slit cannot move *around* the solenoid on their way to the screen, and likewise that electrons from the source cannot move around the solenoid on their way to the slits (we will see the reason for this below). That is why we have placed it as indicated. The presence of the solenoid creates a magnetic field inside of the circular region, but the magnetic field remains zero in the regions where the electrons are travelling.

If one now performs the experiment again, with current flowing through the solenoid, one finds that the interference pattern on the screen is *shifted*. What could explain this? At first it seems that electrons that are not passing through the solenoid must be unaffected, since there is no magnetic field in this region. If this is so then the shifting must be caused by electrons passing *through* the solenoid. But there are experimental ways of ruling this out. For instance, if the radius of the solenoid is made very small (but keeping the same magnetic flux) presumably fewer electrons are able to pass through it; and yet the shifting in the interference pattern remains the same. One could take this further and imagine a thought experiment in which the solenoid is made so small that basically there is zero probability of electrons passing through it—and yet if the interference pattern remains the same no matter how small the solenoid is, that seems to rule out any explanation having to do with happenings inside the solenoid. Alternatively, perhaps one could construct the solenoid out of a material where the electrons can definitely not pass through it; so if having the current on or off still results in interference patterns that are shifted from each other, one again knows that this is not caused by anything happening inside the solenoid.

So experiment tells us that we must look outside the solenoid for an explanation of the shifting interference pattern. And the only thing that is different *outside* the solenoid is the magnetic potential, not the magnetic field. This shows that in the world of quantum mechanics the magnetic potential has more of a physical reality than one might first believe.

Let us now make this analysis a bit more quantitative, and show exactly how the magnetic potential explains the observations of this experiment.

Recall that $\langle 1|s\rangle$ denotes the probability amplitude of an electron leaving s to reach position 1. We will also be concerned with the amplitudes $\langle 2|s\rangle$, $\langle x|1\rangle$, and $\langle x|2\rangle$ where x denotes a point on the screen. The Feynman approach to quantum mechanics allows us to think of $\langle 1|s\rangle$ as a path integral over all possible paths from s to 1:

$$\langle 1|s\rangle = \int_\gamma e^{\frac{i}{\hbar}S(\gamma)}\,D\gamma.$$

Similar descriptions are valid for the other amplitudes. Here S is the action. For the experiment without the solenoid it is the action for a free particle.

Note that the probability amplitude of an electron passing through slit 1 and reaching the screen at x is $\langle x|1\rangle\langle 1|s\rangle$. Therefore the probability of an electron leaving source s and arriving at x is

$$P(x) = \big|\langle x|1\rangle\langle 1|s\rangle + \langle x|2\rangle\langle 2|s\rangle\big|^2.$$

Now add the solenoid to the situation. As we saw in Section 4.3.3, the presence of electromagnetism leads to a new action given by

$$S_{new}(\gamma) = S(\gamma) + q\cdot\int_\gamma \mathbf{A}\cdot\mathbf{ds}.$$

Let $S_{em}(\gamma)$ denote the term $q\int_\gamma \mathbf{A}\cdot\mathbf{ds}$.

Every path from s to 1 has the same value under S_{em}, because the curl of $\mathbf{A}$ is zero in this region. This is why it was important that the electrons not be able to move around the solenoid in their journey from the source to a slit! We can therefore write

$$\langle 1|s\rangle_{new} = \int_\gamma e^{\frac{i}{\hbar}(S(\gamma)+S_{em}(\gamma))}\,D\gamma = e^{\frac{i}{\hbar}S_{em}(\gamma_{1s})}\int_\gamma e^{\frac{i}{\hbar}S(\gamma)}\,D\gamma = e^{\frac{i}{\hbar}S_{em}(\gamma_{1s})}\langle 1|s\rangle$$

where γ_{1s} is any fixed path from s to 1. A similar analysis applies to paths from 1 to x, giving us

$$\langle x|1\rangle_{new}\langle 1|s\rangle_{new} = e^{\frac{i}{\hbar}S_{em}(\gamma_1)}\cdot\langle x|1\rangle\langle 1|s\rangle$$

where now γ_1 denotes any path from s to x passing through 1. Of course we likewise have

$$\langle x|2\rangle_{new}\langle 2|s\rangle_{new} = e^{\frac{i}{\hbar}S_{em}(\gamma_2)}\cdot\langle x|1\rangle\langle 1|s\rangle$$

where γ_2 is any path from s to x passing through 2. Putting all of this together, we get that

$$\begin{aligned}P_{new}(x) &= \big|e^{\frac{i}{\hbar}S_{em}(\gamma_1)}\cdot\langle x|1\rangle\langle 1|s\rangle + e^{\frac{i}{\hbar}S_{em}(\gamma_2)}\cdot\langle x|2\rangle\langle 2|s\rangle\big|^2\\ &= \big|\langle x|1\rangle\langle 1|s\rangle + e^{\frac{i}{\hbar}S_{em}(\gamma_2-\gamma_1)}\cdot\langle x|2\rangle\langle 2|s\rangle\big|^2\end{aligned}$$

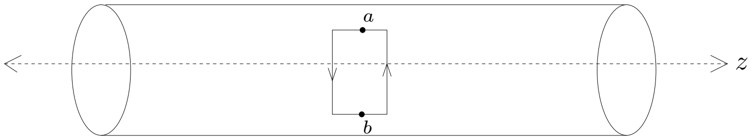
where $\gamma_2 - \gamma_1$ denotes the loop that follows γ_2 and then follows γ_1 backwards. But by Stokes's Theorem $\int_{\gamma_2-\gamma_1} \mathbf{A}\cdot\mathbf{ds} = \iint_{interior} \mathbf{B}\cdot\hat{n}\,dA = \Phi$, where Φ is the total magnetic flux through the solenoid. So we have

$$P_{old}(x) = \left|\langle x|1\rangle\langle 1|s\rangle + \langle x|2\rangle\langle 2|s\rangle\right|^2,$$
$$P_{new}(x) = \left|\langle x|1\rangle\langle 1|s\rangle + e^{\frac{i}{\hbar}q\Phi}\cdot\langle x|2\rangle\langle 2|s\rangle\right|^2.$$

One can see from these formulas that the interference pattern will be different in the presence of the solenoid.

4.4.2. Mathematical analysis of the solenoid. Consider an infinitely-long solenoid centered around the z-axis. The symmetry of the situation makes it clear that the magnetic field $\mathbf{B}$ at each point will either be directed radially towards/from the center of the solenoid, or else will point entirely in the z-direction. Consideration of the picture at the beginning of this section shows that it is definitely the latter: at all points $\mathbf{B}$ is parallel to the z-axis. Symmetry then also tells us the magnitude of $\mathbf{B}$ will only depend on the distance to the z-axis.

Based on the above, one can deduce from Ampère's Law that $\mathbf{B}$ will vanish outside the solenoid and be constant inside the solenoid. Indeed, imagine a very narrow rectangular loop of wire of width dx placed inside the solenoid:



If we write $\mathbf{B}(x) = B(x)\hat{\mathbf{k}}$ then $\int_{\text{loop}} \mathbf{B}\cdot\mathbf{ds}$ is approximately $[B(b)-B(a)]dx$, whereas Ampère's Law says this integral must be zero (because no current flows across the cross-sectional area of the loop, and the electric field is static). One deduces that $B(a) = B(b)$, and this will hold for all points a and b having the same z-value inside the solenoid. We can also make this argument for two points outside the solenoid: but here it is evident that as one gets infinitely far away the magnetic field must be zero, therefore the magnetic field must be zero everywhere.

Let R denote the radius of the solenoid. We have argued that

$$\mathbf{B}(x,y,z) = \begin{cases} 0 & \text{if } x^2+y^2 > R^2 \\ C\hat{\mathbf{k}} & \text{if } x^2+y^2 \le R^2 \end{cases}$$

where C is some constant. We next look for a magnetic potential $\mathbf{A}$, so we want to solve $\operatorname{curl}\mathbf{A} = \mathbf{B}$. There are many solutions and we just need to find *one*, and we would expect it to be radially symmetric. With a little more experience in basic electromagnetism we could motivate the following guess more, but for the present purposes let us just treat it as a guess:we will look for a potential of the form

$$\mathbf{A}(x,y,z) = g(r)\cdot\left[\frac{-y\hat{\mathbf{i}} + x\hat{\mathbf{j}}}{x^2+y^2}\right]$$

where $r = \sqrt{x^2+y^2}$. Note that this is a vector field circulating around the z-axis, whose magnitude varies with r.

A routine computation shows that

$$\operatorname{curl}\mathbf{A} = \frac{g'(r)}{r}\hat{\mathbf{k}}$$

and we want this to equal $\mathbf{B}$. It follows that $g'(r) = rC$ for $r < R$, and so we can take $g(r) = \frac{r^2C}{2}$. It also follows that $g'(r) = 0$ for $r > R$, and so for $r > R$ we will have $g(r) = g(R) = \frac{R^2C}{2}$. Note that $\pi R^2 C$ is just the magnetic flux through a cross-section of the solenoid. If we denote this flux as Φ, we can write

$$\mathbf{A}(x,y,z) = \begin{cases} \frac{r^2\Phi}{2\pi R^2}\left[\dfrac{-y\hat{\mathbf{i}} + x\hat{\mathbf{j}}}{x^2+y^2}\right] & \text{if } r \leq R, \\ \frac{\Phi}{2\pi}\left[\dfrac{-y\hat{\mathbf{i}} + x\hat{\mathbf{j}}}{x^2+y^2}\right] & \text{if } r \geq R. \end{cases}$$

This completes our derivation of a magnetic potential for the solenoid. Note that we didn't actually need the specific form of this potential for our analysis of the Aharanov-Bohm effect; however, it is still nice to be able to look at a concrete formula.

4.5. Magnetic monopoles and topology

The Maxwell equation $\nabla \cdot \mathbf{B} = 0$ comes from the fact that, as far as we know, there are no magnetic charges (also called "monopoles") in the universe. But the observation that electric and magnetic fields are nearly dual to each other might make one wonder if maybe there *are* magnetic charges, and possibly we just haven't found them yet. In the process of thinking about this, Dirac found a quantum-mechanical argument showing that if magnetic charges exist then electric charges (and magnetic charges too) must be quantized. Of course we know from experiment that electric charges *are* quantized, and we have no good explanation for this, so Dirac's argument is very intriguing. Even more intriguing is that his argument is essentially topological! In the end it boils down to the fact that $H^2(S^2) \cong \mathbb{Z}$.

Imagine that we had a magnetic monopole of strength μ, sitting at the origin of $\mathbb{R}^3$. The magnetic field generated by such a monopole would point radially away from the origin (or towards it, depending on the sign of μ), be spherically symmetric, and it would have to be divergence-free away from the origin. As always, this forces

$$\mathbf{B}(x,y,z) = \frac{\mu}{\rho^2} \cdot \left[\frac{x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}}{\rho}\right]$$

where $\rho = \sqrt{x^2+y^2+z^2}$. In particular, for a sphere of radius R about the origin we would have

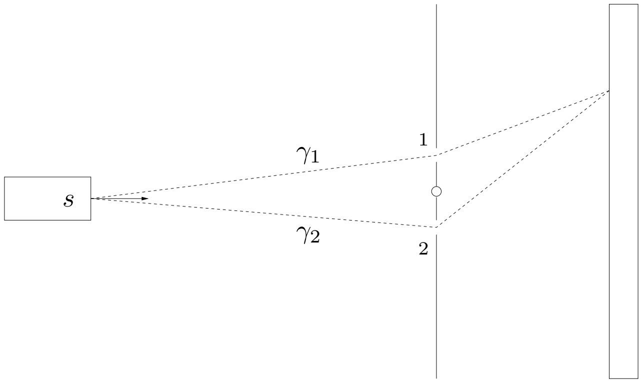
$$\int_{S_R} (\mathbf{B}\cdot\hat{\mathbf{n}})\, dA = \frac{\mu}{R^2} \cdot 4\pi R^2 = 4\pi\mu.$$

The field $\mathbf{B}$ is defined and divergence-free on $\mathbb{R}^3 - 0$, but unlike situations we have considered before there is no magnetic potential for $\mathbf{B}$. This is because the deRham complex for $\mathbb{R}^3 - 0$ is not exact in degree 2, due to $H^2(\mathbb{R}^3 - 0) \neq 0$.

We *can* obtain a magnetic potential by picking any ray W from the origin to infinity, and looking only on the space $\mathbb{R}^3 - W$. For this space is contractible, and therefore has vanishing H^2. The subspace W doesn't even have to be a ray, it can be any reasonable curve going out to infinity—in the present context W is called a "Dirac string."

Let z_+ be the non-negative z-axis, and let z_- denote the non-positive z-axis. Let $U_+ = \mathbb{R}^3 - z_-$ and $U_- = \mathbb{R}^3 - z_+$. These cover $\mathbb{R}^3 - 0$, and on each piece of the cover there exists a magnetic potential for $\mathbf{B}$. Let $\mathbf{A}_+$ denote a potential defined on U_+, and let $\mathbf{A}_-$ denote a potential defined on U_-. Of course $\mathbf{A}_+$ and $\mathbf{A}_-$ will not agree on $U_+ \cap U_-$, or else they would patch together to give a potential defined on all of $\mathbb{R}^3 - 0$.

Now let us run a double slit experiment, with and without a magnetic monopole in the middle. Note that this experiment is taking place in $\mathbb{R}^3$, although we can imagine a scenario where our particles are constrained to move in a plane. Let us assume the motion is taking place in the xy-plane, and that our picture shows a "top view" of the experiment:



As we saw in our discussion of the Aharanov-Bohm effect (see the previous section), in the double slit experiment without the magnetic monopole the probability of an electron from the source reaching spot x on the screen is

$$P(x) = \big|\langle x|1\rangle\langle 1|s\rangle + \langle x|2\rangle\langle 2|s\rangle\big|^2.$$

We also saw that in the presence of a magnetic field the probability is

$$P_{mag}(x) = \big|\langle x|1\rangle\langle 1|s\rangle + e^{\frac{i}{\hbar}\cdot q\Phi_+}\langle x|2\rangle\langle 2|s\rangle\big|^2$$

where $\Phi_+ = \int_{\gamma_2-\gamma_1} \mathbf{A}_+ \cdot \mathbf{ds}$. Here we are thinking of our plane (minus the origin) as lying inside of U_+ and using the magnetic potential $\mathbf{A}_+$. Also, recall that our notation has all probability amplitudes like $\langle x|1\rangle$ referring to the double-slit experiment *without* the magnetic monopole.

But now notice that we could just as well have regarded our plane (minus origin) as lying inside of U_-, and used the magnetic potential $\mathbf{A}_-$. This gives the formula

$$P_{mag}(x) = \big|\langle x|1\rangle\langle 1|s\rangle + e^{\frac{i}{\hbar}\cdot q\Phi_-}\langle x|2\rangle\langle 2|s\rangle\big|^2$$

where $\Phi_- = \int_{\gamma_2-\gamma_1} \mathbf{A}_- \cdot \mathbf{ds}$. Since both formulas for $P_{mag}(x)$ are valid, we must have $e^{\frac{i}{\hbar}q\Phi_+} = e^{\frac{i}{\hbar}q\Phi_-}$. This holds if and only if

$$\frac{q\Phi_+}{\hbar} - \frac{q\Phi_-}{\hbar} \in 2\pi\mathbb{Z}. \qquad (4.5.1)$$

Since we know $\mathbf{B}$ precisely, we can actually compute Φ_+ and Φ_- in terms of the monopole strength μ. By Stokes's Theorem we can write

$$\Phi_+ = \int_{\gamma_2-\gamma_1} \mathbf{A}_+ \cdot \mathbf{ds} = \iint_{D_+} \mathbf{B}\cdot\hat{\mathbf{n}}\, dA = \mu \cdot \frac{4\pi}{2} = 2\pi\mu$$

where D_+ is a disk in U_+ bounding $\gamma_2 - \gamma_1$, and where the third equality comes from the fact that we may as well assume D_+ is the top half of the unit sphere (by assuming $\gamma_2 - \gamma_1$ is the equator). Likewise

$$\Phi_- = \int_{\gamma_2-\gamma_1} \mathbf{A}_- \cdot \mathbf{ds} = \iint_{-D_-} \mathbf{B}\cdot\hat{\mathbf{n}}\, dA = -\mu \cdot \frac{4\pi}{2} = -2\pi\mu.$$

The unfortunate notation $-D_-$ denotes the bottom half of the unit sphere but with the opposite of the usual orientation, which is required by Stokes's Theorem for compatibility with the orientation of $\gamma_2 - \gamma_1$.

Equation (4.5.1) gives $\Phi_+ - \Phi_- \in \frac{2\pi\hbar}{q}\mathbb{Z}$. Using $\Phi_+ = 2\pi\mu = -\Phi_-$ we immediately get the Dirac quantization condition

$$\mu q \in \frac{\hbar}{2}\mathbb{Z}.$$

So given the existence of a magnetic monopole of strength μ, the only possible electric charges are $\frac{n\hbar}{2\mu}$ for $n \in \mathbb{Z}$. The argument does not show that electric charges occur corresponding to all possible values of n, but in any case we can agree that the set of possible electric charges will be a $\mathbb{Z}$-submodule of $\frac{\hbar}{2\mu}\mathbb{Z}$ (since two electric charges of strengths q_1 and q_2 could presumably be bound together in some way to create an effective charge of.strength $q_1 + q_2$). Any such $\mathbb{Z}$-submodule looks like $(\frac{n\hbar}{2\mu})\mathbb{Z}$ for some n, showing that the possible electric charges are precisely the integral multiples of some basic charge $\frac{n\hbar}{2\mu}$.

The same argument of course applies to magnetic charge strengths. If q is an electric charge that occurs in nature, then the set of possible magnetic charges will be $(\frac{k\hbar}{2q})\mathbb{Z}$ for some integral k. So both electric and magnetic charges will be quantized.

We close this section by remarking that the above arguments are essentially topological: we are using that $\mathbb{R}^3 - 0$ can be covered by two contractible pieces whose intersection is $\mathbb{R}^2 - 0 \simeq S^1$. The $\mathbb{Z}$ appearing in the Dirac quantization condition is essentially $H^1(S^1)$, or alternatively $H^2(S^2)$. We will return to this topic after we have introduced principal bundles and gauge theory.

Part 2

Delving deeper into quantum physics

CHAPTER 5

Spin and Dirac's theory of the electron

A rotation of $\mathbb{R}^3$ is determined by an axis, an angle, and a direction (clockwise or counterclockwise, relative to some orientation of the axis). The set of all rotations forms a group under composition, denoted $SO(3)$. It should be no surprise that this group plays a large role in physical theories—it is important to be able to transform results from one coordinate system to a rotated one, and also many basic physical problems have spherical symmetry.

Now, *of course* it is true that rotating by 360 degrees is the same as the identity; this hardly needs to be said out loud. It turns out, however, that there exist mathematical constructs where you rotate them by 360 degrees and it's *not* the identity. This is somewhat of a shock. (If you rotate them by 720 degrees they are back the way they started, though—and unlike the 360 degree version this rule is absolute). Even more shockingly, these strange constructs turn out to be crucial to giving quantum-mechanical descriptions of electrons and other particles. This is the theory of spin.

It will take several pages to give a complete explanation of this theory, but I want to try to give the general picture in just a few paragraphs. At first the whole situation seems preposterous: rotation by 360 degrees is the identity in $SO(3)$, that is unquestionable; and if two things are the same, what could it possibly mean to have a world where they are *not* the same? The key is to understand that there is a conceptual difference between "large rotations" and "small rotations". One makes a large rotation about some axis by taking a "small rotation" and doing enough of it. Keep this in mind while I describe the situation more rigorously.

The first thing that is important is that $\pi_1(SO(3)) = \mathbb{Z}/2$. You don't need to remember why this is true, just accept it; but I will explain it carefully in Section 5.1 below. The $\mathbb{Z}/2$ tells us that if $X \to SO(3)$ is the universal cover then every point of $SO(3)$ has two points in the fiber. One can show that the space X has a group structure that is inherited from $SO(3)$ (again, don't worry about picturing this for the moment, just accept it). In a small neighborhood of the identity these groups look pretty much the same! One can think of an infinitesimal neighborhood of the identity in $SO(3)$ as consisting of "small rotations", and so the infinitesimal elements of X have the same description—to be rigorous here we would say that X and $SO(3)$ have the same Lie algebra. This is important: we now see that there are these "two worlds"—X and $SO(3)$—both of whose elements can be obtained by taking "small rotations" and then doing more and more of them. We will think of an element of X as a kind of "generalized rotation". By the way, the fact that $\pi_1(SO(3)) = \mathbb{Z}/2$ shows that X is the *only* connected covering space of $SO(3)$—so these are really the only two "worlds" made from small rotations in this way.

Since $X \to SO(3)$ is a 2-fold covering, there are exactly two elements of X that map to the identity. One is the identity of X, the other is something else—call it E, just to have a name for it. Choose an axis in $\mathbb{R}^3$ and consider the gradual rotatation that starts at the identity and slowly rotates about the axis until the rotation degree is 360. This is a path in $SO(3)$ (a loop, in fact). Lift it to a path in X that also starts at the identity, and you will find that the ending point is the element E. So working in X, gradually rotating more and more—until we get to 360 degrees—doesn't give us the identity anymore! If you take a loop in $SO(3)$ that gradually rotates through 720 degrees about an axis, then lifting that into X yields a path that *does* still end at the identity. Generalized rotations through 720 degrees are still the identity, even in X.

How does all of this appear in quantum mechanics? Recall that a state of a quantum mechanical system is described by a vector ψ in some complex vector space $\mathcal{H}$. Rotating the physical system should put it in a different state, so we expect $SO(3)$ to act on the vector space $\mathcal{H}$. If we rotate the system through 360 degrees about an axis, we can certainly agree that physically nothing should be different—all measurements should give the same answers. But here is the problem! Because it is built into the framework of quantum mechanics that all multiples of ψ describe the same physics—they all give rise to the same probabilities. So knowing that our state gives all the same measurements as the initial state ψ doesn't necessarily imply that it is *equal* to ψ, only that they are multiples of each other. This opens the door: as we rotate through greater and greater angles approaching 360 degrees, the state ψ only needs to gradually move to a *multiple* of itself. Mathematical arguments show that this multiple can only be 1 or -1 (essentially hinging on the fact that $\pi_1 SO(3) = \mathbb{Z}/2$); so doing the 360 rotation *twice* will always bring us back to the original state ψ. If the multiple is 1 we have an honest-to-god representation of $SO(3)$, whereas if the multiple is -1 what we have is a representation of the mysterious group X (what mathematicians now call a *spin representation*).

So far this is all well and fine, but why would the universe prefer representations of X to representations of $SO(3)$? Is there a reason to expect, naively, that X should appear in physics? The best answer I can give is the following. The simplest nontrivial quantum systems are ones where there are two basic states—where the associated complex vector space has dimension 2. It turns out that it is mathematically impossible to give an honest representation of $SO(3)$ on a 2-dimensional vector space, except for the trivial representation. However it *is* possible to give interesting representations of X in this setting. I like to think of things this way: once the basic machinery of quantum mechanics was in place, it was inevitable that people would run into quantum systems where there were only two basic states; and as soon as that happened, the appearance of X was basically guaranteed.

I will spend the next several sections trying to say everything in the above paragraphs in a more rigorous manner. Incidentally, the group X is officially known as Spin(3). There is a version of the above story for $SO(n)$ as well, leading to a group called Spin(n). These spin groups are of tremendous importance in modern mathematics. We will begin by learning more about them.

5.1. The Spin groups

Let G be any topological group, and let $p\colon X \to G$ be a covering space. Choose a point x in X in the fiber over the identity element e. I claim that X has a structure of topological group where x is the identity and where p is a group homomorphism; moreover, the structure having these two properties is unique. Indeed, consider the lifting problem

$$\begin{tikzcd}
 & & X \arrow[d, "p"] \\
X \times X \arrow[r, "p\times p"] \arrow[urr, dotted, "\lambda"] & G \times G \arrow[r, "\mu"] & G.
\end{tikzcd}$$

The image of $\pi_1(X \times X, (x, x))$ under the horizontal map is precisely $p_*(\pi_1(X, x))$, and so elementary covering-space theory says that there is a unique lifting λ sending (x, x) to x. This gives a multiplication X for which p is a homomorphism. It is not transparent that the multiplication is associative, but one observes that $\lambda(x, \lambda(y, z))$ and $\lambda(\lambda(x, y), z)$ both give liftings for the map $X \times X \times X \to G$ sending $(x, y, z) \mapsto p(x)p(y)p(z)$. Uniqueness of liftings in covering spaces yields the required associativity in X, and similar arguments work for the unital condition and the existence of inverses.

In particular, observe that the universal covering space of G has a group structure (once one chooses a basepoint in the fiber over the identity).

We next recall that $\pi_1(SO(n)) = \mathbb{Z}/2$ for $n \geq 3$. The explanation for this comes in two parts. First one argues that $SO(3)$ is homeomorphic to $\mathbb{R}P^3$. Construct a map $D^3 \to SO(3)$ by sending a vector $u \in D^3$ to the rotation about the axis $\langle u\rangle$ through an angle of $180|u|$ degrees, counterclockwise from the viewpoint of a person standing on the tip of u with his body directed radially outward from the origin. Notice that this description doesn't quite make sense when u is the origin, since $\langle u\rangle$ is not a line—but for vectors close to the origin the rotation is through very small angles, and so it makes sense to send 0 to the identity rotation. Finally, observe that if $|u| = 1$ then both u and $-u$ get sent to the same rotation. So one obtains a map $\mathbb{R}P^3 \to SO(3)$, and one readily checks that this is a bijection. The spaces involved are compact and Hausdorff, so any continuous bijection is a homeomorphism.

At this point we know that $\pi_1(SO(3)) = \mathbb{Z}/2$. To extend to higher n, consider the standard action of $SO(n)$ on $S^{n-1} \subseteq \mathbb{R}^n$. The action is transitive, so S^{n-1} is a homogeneous space for $SO(n)$. The stabilizer of the basis vector e_1 is just $SO(n-1)$, and hence there is a fiber bundle

$$SO(n-1) \to SO(n) \twoheadrightarrow S^{n-1}.$$

The long exact sequence for homotopy groups immediately yields isomorphisms $\pi_1(SO(n-1)) \cong \pi_1(SO(n))$ for $n \geq 4$. This completes the argument that $\pi_1(SO(n)) = \mathbb{Z}/2$ for $n \geq 3$.

The universal cover of $SO(n)$, with its induced group structure, is denoted $\text{Spin}(n)$ and called the nth **spin group**. The kernel of $\text{Spin}(n) \to SO(n)$ is a 2-element subgroup, so let's denote it $\{I, E\}$ where I is the identity and E is the nontrivial element. So $\text{Spin}(n)/\{I, E\} \cong SO(n)$.

Note that in a neighborhood of the identity the spaces $\text{Spin}(n)$ and $SO(n)$ are homeomorphic (because $\text{Spin}(n) \to SO(n)$ is a covering space). So the infinitesimal structure of these two groups is the same. More rigorously, we say that they have the same Lie algebra.

Let L be any line in $\mathbb{R}^3$, and let $\sigma\colon [0, 1] \to SO(3)$ be the path such that $\sigma(t)$ is clockwise rotation about the z-axis through $360t$ degrees (clockwise relative to some fixed frame of reference). This is a loop in $SO(3)$, and using the homeomorphism $SO(3) \cong \mathbb{R}P^3$ from above one readily sees that it represents the nontrivial generator of $\pi_1(SO(3))$. By covering space theory there is a unique lift of σ to a path $\tilde{\sigma}\colon [0,1] \to \mathrm{Spin}(n)$ such that $\tilde{\sigma}(0) = I$. This *cannot* be a loop in $\mathrm{Spin}(n)$, because $\mathrm{Spin}(n)$ was the universal cover of $SO(n)$ [if it were a loop it would be null-homotopic, which would imply that σ is also null-homotopic—and this is not true]. So it must be that $\tilde{\sigma}(1) = E$.

We can describe the mathematics of the above paragraph more informally as follows. The idea of "small rotation about an axis L" has essentially equal meanings in the groups $SO(n)$ and $\mathrm{Spin}(n)$—because they have the same infinitesimal structure near the identity. In $SO(n)$ if one keeps rotating more and more, then after 360 degrees one gets back to the identity. If one tries to match these rotations in $\mathrm{Spin}(n)$, then after 360 degrees one is *not* at the identity, but rather at the exotic element E.

It is often useful to have an explicit model for the groups $\mathrm{Spin}(n)$, with elements that one can manipulate by hand. This is relatively easy to provide when $n = 3$. Since $SO(3)$ is homeomorphic to $\mathbb{R}P^3$, it must be that $\mathrm{Spin}(3)$ is homeomorphic to the sphere S^3. One has a convenient group structure on S^3 coming from multiplication of unit quaternions, and we claim that this *is* the group $\mathrm{Spin}(3)$. To see this it will suffice to exhibit $SO(3)$ as a quotient of S^3 by a group of order 2.

Recall that the division algebra $\mathbb{H}$ of quaternions has a norm given by $|x| = x\overline{x}$, and that $|xy| = |x| \cdot |y|$ for any two quaternions. So if q is a unit quaternion then both left- and right-multiplication by q preserves the norm. Moreover, the norm coincides with the standard Euclidean norm on the underlying real vector space, with $1, i, j, k$ being an orthonormal basis.

For $q \in \mathbb{H}$ let $\Upsilon_q\colon \mathbb{H} \to \mathbb{H}$ be given by $x \mapsto qx\overline{q}$. The map Υ_q preserves the norm and sends 1 to 1, hence it also preserves the orthogonal complement to 1. This complement is the subspace $\langle i, j, k\rangle$ of purely imaginary elements. Identifying this subspace with $\mathbb{R}^3$ in the evident way, the map Υ_q restricts to an orthogonal transformation of $\mathbb{R}^3$. In this way we obtain a group homomorphism $S^3 \to O(3)$, and the fact that S^3 is connected implies that the image must land in $SO(3)$. It takes a little work, but one can indeed show that the image *equals* $SO(3)$ and that the kernel is $\{1, -1\}$ (see Example 5.1.1 below for a complete analysis). This proves that $\mathrm{Spin}(3) \cong S^3$.

To get an explicit model for $\mathrm{Spin}(n)$ for $n > 3$ one has to work a little harder. The standard approach is to use Clifford algebras, which essentially play the role that $\mathbb{H}$ did when $n = 3$. The reader can safely ignore the following paragraph for the moment, as we will not need it anytime soon, but it seems reasonable to quickly give this description here.

Let Cl_n be the quotient of the tensor algebra $\mathbb{R}\langle e_1, \ldots, e_n\rangle$ by the relations $e_i \otimes e_j + e_j \otimes e_i = -2\delta_{ij}$. This relation identifies a 2-tensor with a 0-tensor, and so it preserves the parity of tensors. So it makes sense to talk about even and odd tensors inside of Cl_n. If $a_1^2 + \cdots + a_n^2 = 1$, check that $u = a_1e_1 + \cdots a_ne_n$ satisfies $u^2 = -1$; in particular, u is a unit. Let Cl_n^* denote the multiplicative group of units, which we have just seen contains the sphere S^{n-1}. Let $\mathrm{Pin}(n) \subseteq \mathrm{Cl}_n^*$ be the subgroup *generated by* the elements of S^{n-1}. Finally, let $\mathrm{Spin}(n)$ be the set of even tensors inside of $\mathrm{Pin}(n)$. It takes a little work to produce a map $\mathrm{Spin}(n) \to SO(n)$ and prove it has the desired properties, but it is not terribly difficult. See [**LM**, Chapter 1] for more details.

The following material will be useful when we discuss the Pauli spin matrices:

Example 5.1.1 (Quaternions and rotations). Let $q \in \langle i, j, k\rangle$ be a unit quaternion. The real subspace $\langle 1, q\rangle \subseteq \mathbb{H}$ is a subalgebra of $\mathbb{H}$ that is isomorphic to the complex numbers, with q corresponding to i. The unit circle in this copy of the complex numbers is the set of elements

$$q_\theta = \cos(\theta) + \sin(\theta)q.$$

In essence, the lesson here is that any purely imaginary unit quaternion is "as good as i, j, or k". Every unit quaternion lies on exactly one of these circles, with q corresponding to the (normalized) imaginary part of the given quaternion. Of course we are seeing the Hopf bundle $S^1 \to S^3 \to S^2$ here.

Another useful fact that goes with the above is that for any $x \in \langle i, j, k\rangle$ that is orthogonal to q, the vectors 1, q, x, and qx constitute an oriented, orthonormal frame for $\mathbb{H}$.

For any $v \in \mathbb{R}^3$ and any angle θ, let $R_{v,\theta}$ denote the rotation of $\mathbb{R}^3$ about the axis $\langle v\rangle$, through an angle θ, that is oriented counterclockwise with respect to a peron who stands at the origin and has his or her head pointing in the direction of v. It is not immediately clear how to write down the matrix in $SO(3)$ corresponding to $R_{v,\theta}$, or how to express a composition $R_{w,\alpha}R_{v,\beta}$ in the same form. It turns out that quaternions give a good way to do these things, based on the group homomorphism $S^3 \to SO(3)$ described above.

We claim that for any unit quaternion q in $\langle i, j, k\rangle$ and any angle θ, the conjugation map Υ_{q_θ} (given by $x \mapsto q_\theta x \overline{q}_\theta$) coincides with the rotation $R_{q,2\theta}$ on $\langle i, j, k\rangle$. To see this, first note that Υ_{q_θ} clearly fixes q. If $x \in \langle i, j, k\rangle$ is orthogonal to q then $xq = -qx$ and so

$$\begin{aligned}\Upsilon_{q_\theta}(x) = q_\theta x \overline{q_\theta} &= \big(\cos\theta + (\sin\theta)q\big)x\big(\cos\theta - (\sin\theta)q\big)\\ &= (\cos^2\theta - \sin^2\theta)x + (2\sin\theta\cos\theta)qx\\ &= (\cos 2\theta)x + (\sin 2\theta)qx.\end{aligned}$$

Given that q, x, qx is an orthonormal frame for $\langle i, j, k\rangle$, the above formulas describe the desired rotation. Incidentally, this analysis proves that $S^3 \to SO(3)$ is surjective: the rotations $R_{v,\theta}$ are in the image, and they generate $SO(3)$. It also shows that a given rotation will have exactly two preimages, of the form q and $-q$.

So given a unit vector $v = (v_1, v_2, v_3)$ in $\mathbb{R}^3$ and an angle θ, the rotation $R_{v,\theta}$ is the image under $S^3 \to SO(3)$ of the quaternion

$$v_{\theta/2} = \cos(\tfrac{\theta}{2}) + \sin(\tfrac{\theta}{2})(v_1 i + v_2 j + v_3 k).$$

The composition $R_{v,\alpha}R_{w,\beta}$ is the image of the quaternion $v_{\alpha/2} \cdot w_{\beta/2}$; so if we compute this product and separate it into its real and imaginary parts, the normalization of the imaginary part gives us the axis of rotation. The angle of rotation is obtained by dividing the norms of the imaginary and real parts and applying arctangent. These formulas are not 'easy' to execute by hand, but they are conceptually very straightforward (and easy for computers). Likewise, if we want the matrix in $SO(3)$ corresponding to $R_{v,\theta}$ then the columns are obtained by conjugating each of i, j, and k by $v_{\theta/2}$ and taking the i, j, k-coordinates of the result. One gets formulas that are quadratic in the v_i's. By comparison, if one tries to write down the matrix for $R_{v,\theta}$ using Euler angles or some similar mechanism, it is quite cumbersome to get simple formulas involving the v_i's.

5.2. Projective representations and other fundamentals

Let G be any group and let V be a complex vector space. A representation of G on V is simply a group homomorphism $G \to \mathrm{GL}(V)$. That is, for any $g \in G$ one has a linear automorphism $\rho_g \colon V \to V$, and $\rho_{gh} = \rho_g \rho_h$. If G is a *topological* group one usually requires that the map $G \to \mathrm{GL}(V)$ be continuous.

A **projective representation** of G on V is a map of sets $\rho \colon G \to \mathrm{GL}(V)$ together with a function $c \colon G \times G \to \mathbb{C}^*$ such that

$$\rho_{gh} = c(g,h)\rho_g \rho_h$$

for every $g, h \in G$. In other words, we almost have a representation except that the group relation only holds up to multiplication by diagonal matrices. Again, this doesn't quite capture everything we want when G is a topological group, as an extra continuity condition seems appropriate. The condition that $G \to \mathrm{GL}(V)$ be continuous is too strong, though. We will say more about this in a moment.

Let $D \subseteq \mathrm{GL}(V)$ be the set of scalar transformations (constant diagonal matrices). Then D is in the center of $\mathrm{GL}(V)$ and is therefore normal, so we may form the group $\mathrm{PGL}(V) = \mathrm{GL}(V)/D$. A projective representation of G therefore yields a group homomorphism $G \to \mathrm{PGL}(V)$. When G is a topological group we will demand that this map be continuous.

Remark 5.2.1. The notion of projective representation might seem artificial and unappealing at first. Why assume that ρ only preserves products up to scalar factors? Without a motivating example this seems unnatural. For our purposes the motivation is from quantum mechanics, where a vector and its multiples all represent the same "underlying physics". If one doesn't allow projective representations then the theory ends up being too limited, and doesn't model what actually happens in the real world.

In case it is not clear, the name 'projective representation' comes from the fact that there is an induced action on projective space. The standard action of $\mathrm{GL}(V)$ on the projective space $\mathbb{P}(V)$ induces an action of $\mathrm{PGL}(V)$ on $\mathbb{P}(V)$, and therefore an action of G by restriction.

If the map $G \to \mathrm{PGL}(V)$ lifts to $G \to \mathrm{GL}(V)$ then we say that the projective representation lifts to an honest representation. Note that we may always form the pullback

$$\begin{tikzcd}
\tilde{G} \arrow[r] \arrow[d, two heads] & \mathrm{GL}(V) \arrow[d, two heads] \\
G \arrow[r] & \mathrm{PGL}(V),
\end{tikzcd}$$

and that the projective representation of G therefore gives rise to an honest representation of $\tilde{G}$.

Write $\mathrm{PGL}_n(\mathbb{C})$ for $\mathrm{PGL}(\mathbb{C}^n)$. We can also form groups $\mathrm{PSL}_n(\mathbb{C})$ and PU_n by quotienting the groups $\mathrm{SL}_n(\mathbb{C})$ and SU_n by their subgroups of scalar matrices. There are inclusions $\mathrm{PSU}_n \hookrightarrow \mathrm{PSL}_n(\mathbb{C}) \hookrightarrow \mathrm{PGL}_n(\mathbb{C})$, and one readily checks that the second of these is an isomorphism. We could have also formed the group PU_n, but it is isomorphic to PSU_n for the same reason.

Note that the subgroups of scalar matrices in both U_n and $\mathrm{SL}_n(\mathbb{C})$ are isomorphic to the group μ_n of nth roots of unity, which is isomorphic to $\mathbb{Z}/n$. We have two fiber bundles (in fact, covering spaces):

$$\begin{tikzcd}
\mu_n \arrow[r, "="] \arrow[d] & \mu_n \arrow[d] \\
SU_n \arrow[r, tail] \arrow[d, two heads] & \mathrm{SL}_n(\mathbb{C}) \arrow[d, two heads] \\
\mathrm{PSU}_n \arrow[r, tail] & \mathrm{PSL}_n(\mathbb{C}).
\end{tikzcd}$$

Say that an element $\alpha \in \mathrm{PGL}_n(\mathbb{C})$ is **unitary** if its action on $\mathbb{C}P^{n-1}$ preserves orthogonality: if L_1 and L_2 are orthogonal lines in $\mathbb{C}^n$ then $\alpha(L_1)$ and $\alpha(L_2)$ are still orthogonal. Here orthogonality is determined relative to the standard Hermitian product. It is an easy exercise to check that an element of $\mathrm{GL}_n(\mathbb{C})$ preserves orthogonality if and only it if is the product of a scalar matrix and a unitary matrix. This shows that the subgroup of $\mathrm{PGL}_n(\mathbb{C})$ consisting of the unitary elements is precisely PSU_n, as one might expect. A unitary projective representation of a group G therefore gives a map $G \to \mathrm{PSU}_n$.

Let G be a topological group with projective unitary representation $G \to \mathrm{PSU}_n$. If G is simply-connected then covering space theory shows that this map lifts to a map of spaces $\rho\colon G \to SU_n$ that preserves the identity. Covering space theory also guarantees that this will be a group homomorphism (because $(g,h) \mapsto \rho(gh)$ and $(f,h) \mapsto \rho(g)\rho(h)$ lift the same map $G \times G \to \mathrm{PSU}_n$). So for a simply-connected topological group every projective representation comes from an honest representation.

After all of these generalities, let us consider some specific examples.

5.2.2. Projective representations of S^1. A 1-dimensional projective representation is a map into $PSL_1(\mathbb{C}) = \{I\}$, and so this is not very interesting. Let us instead consider 2-dimensional projective representations, and for simplicity let us also assume they are unitary. Consider the following diagram:

$$\begin{tikzcd}
 & & SU_2 \arrow[d] \\
S^1 \arrow[r, "2"'] \arrow[urr, dotted] & S^1 \arrow[r] & PSU_2
\end{tikzcd}$$

where $S^1 \to PSU_2$ is our given representation and $S^1 \to S^1$ is the degree 2 map. The fiber sequence $\mu_2 \to SU_2 \to PSU_2$ shows that $\pi_1(PSU_2) = \mathbb{Z}/2$, and therefore the composite across the horizontal row induces the trivial map on π_1. Covering space theory then shows that there is a lifting $S^1 \to SU_2$ that preserves the identity, as indicated. One readily checks that this is a group map, and so we have a unitary representation of S^1 on $\mathbb{C}^2$. Now, we know that up to isomorphism all such representations decompose into sums of irreducibles; and the irreducible representations of S^1 are 1-dimensional, with one for every integer $m \in \mathbb{Z}$, given by $z.v = z^m v$ for $z \in S^1$. So up to a change of basis in $\mathbb{C}^2$, our map $S^1 \to SU_2$ has the form

$$z \mapsto \begin{bmatrix} z^m & 0 \\ 0 & z^{-m} \end{bmatrix}$$

for some $m \in \mathbb{Z}$. Equivalently, we have that our original map $S^1 \to PSU_2$ has (after a change of basis in $\mathbb{C}^2$) the form

$$e^{i\theta} \mapsto \begin{bmatrix} e^{im\theta/2} & 0 \\ 0 & e^{-im\theta/2} \end{bmatrix} = \cos(\tfrac{m\theta}{2})I + i\sin(\tfrac{m\theta}{2})\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

for some $m \in \mathbb{Z}$. It might look like the values for $\theta = 0$ and $\theta = 2\pi$ are different, but remember that the matrix denotes an element of PSU_2—which is really a *coset* in SU_2 for the subgroup $\{I, -I\}$. As a coset, the values for $\theta = 0$ and $\theta = 2\pi$ are the same.

Changing basis amounts to congugating by an invertible matrix P, which we might as well assume lies in SU_2. So the unitary representation we started with was $S^1 \to PSU_2$ given by

$$e^{i\theta} \mapsto P\Big(\cos(\tfrac{m\theta}{2})I + i\sin(\tfrac{m\theta}{2})\left[\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right]\Big)P^{-1} = \cos(\tfrac{m\theta}{2})I + i\sin(\tfrac{m\theta}{2})\Big[P\left[\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right]P^{-1}\Big].$$

We have therefore seen that all projective unitary representations of S^1 on $\mathbb{C}^2$ are of this form.

At this point we need to ask ourselves what is special about the matrices $P\left[\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right]P^{-1}$, as P ranges through the elements of SU_2. Is there a way of describing such matrices intrinsically? One thing that is easy to see is that they are all Hermitian matrices of trace zero. To say more than this, recall that $M_{2\times 2}(\mathbb{C})$ has a Hermitian inner product given by $\langle X, Y \rangle = \sum_{i,j} x_{ij}\overline{y_{ij}}$. Notice that the condition that a matrix have trace zero is equivalent to it being orthogonal to the identity matrix. If $P \in U_2$ then an easy calculation shows that

$$\langle PX, PY \rangle = \langle X, Y \rangle = \langle XP, YP \rangle$$

for all matrices X and Y. In particular, since $\left[\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right]$ has norm $\sqrt{2}$ so do all the conjugates $P\left[\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right]P^{-1}$.

Consider the space of all Hermitian, trace zero matrices in $M_{2\times 2}(\mathbb{C})$. This has dimension 3 over $\mathbb{R}$. The set of all matrices $P\left[\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right]P^{-1}$ lies inside this space, and a little work shows that it is precisely the sphere of radius $\sqrt{2}$ (the last paragraph showed the subset in one direction).

Our discussion has proven the following result:

PROPOSITION 5.2.3. *Every group homomorphism $S^1 \to PSU_2$ has the form*

$$e^{i\theta} \mapsto \cos(\tfrac{m\theta}{2})I + i\sin(\tfrac{m\theta}{2})J$$

where $m \in \mathbb{Z}$ and J is Hermitian, trace zero, and has norm $\sqrt{2}$. Every such matrix J squares to the identity.

The space consisting of the above matrices J is a sphere inside of a 3-dimensional Euclidean vector space (the Hermitian, trace zero matrices). It makes sense to choose an orthogonal basis $\sigma_1, \sigma_2, \sigma_3$ for this vector space, all of norm $\sqrt{2}$, and then to represent elements J as $a_1\sigma_1 + a_2\sigma_2 + a_3\sigma_3$ where $\sum a_i^2 = 1$. The standard choice is

$$\sigma_1 = \sigma_x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad \sigma_2 = \sigma_y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \qquad \sigma_3 = \sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

(The three elements chosen here are very natural, at least up to sign, but their exact ordering is not: the choice of ordering was largely determined by historical accident. See ?????) For later use it is convenient for us to record the following relations:

$$\sigma_x^2 = \sigma_y^2 = \sigma_z^2 = I, \qquad \sigma_x\sigma_y = -\sigma_y\sigma_x = i\sigma_z, \qquad [\sigma_x, \sigma_y] = 2i\sigma_z.$$

All relations obtained from these by cyclically permuting the symbols x, y, and z are also satisfied.

REMARK 5.2.4. As J runs over all Hermitian, trace zero matrices of norm $\sqrt{2}$, the matrix iJ runs over all anti-Hermitian, trace zero matrices of norm $\sqrt{2}$. Whereas $J^2 = I$ one has $(iJ)^2 = -I$. This is a trivial observation, but one should keep it in mind when reading different treatments of this basic theory.

5.2.5. Projective representations of $SO(3)$. Our goal is to analyze projective, unitary representations of $SO(3)$ on $\mathbb{C}^2$. That is, we will analyze group homomorphisms $SO(3) \to PSU_2$.

First recall the fiber sequence $\mu_2 \to SU_2 \to PSU_2$. We claim that this is isomorphic to the fiber sequence $\mathbb{Z}/2 \to S^3 \to SO(3)$. A unit quaternion acts on $\mathbb{H}$ by left multiplication, and this is a unitary map of complex vector spaces if $\mathbb{H}$ is given the complex structure of *right* multiplication. If we write $q = z_1 + jz_2$ with $z_1, z_2 \in \mathbb{C}$, then $qj = z_1 j + jz_2 j = j\overline{z_1} - \overline{z_2}$. So with respect to the complex basis $1, j$ for $\mathbb{H}$, the matrix for left-multiplication-by-q is

$$\phi(q) = \begin{bmatrix} z_1 & -\overline{z_2} \\ z_2 & \overline{z_1} \end{bmatrix}.$$

This describes a group homomorphism $\phi\colon S^3 \to SU_2$ that is readily checked to be an isomorphism. Since it maps -1 to $-I$, it descends to an isomorphism on the quotients $\tilde{\phi}\colon SO(3) \to PSU_2$.

REMARK 5.2.6. The map ϕ is not only a group homomorphism, but it is the restriction of an $\mathbb{R}$-linear map $\mathbb{H} \to M_{2\times 2}(\mathbb{C})$. In particular, if $q = q_0 + q_1 i + q_2 j + q_3 k$ then

$$\begin{aligned} \phi(q) &= q_0 I + q_1\phi(i) + q_2\phi(j) + q_3\phi(k) \\ &= q_0 I + q_1 \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix} + q_2 \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} + q_3 \begin{bmatrix} 0 & -i \\ -i & 0 \end{bmatrix}. \end{aligned}$$

The matrices $\phi(i)$, $\phi(j)$, and $\phi(k)$ probably remind you of the Pauli spin matrices; in fact, we have

$$\phi(i) = i\sigma_z, \qquad \phi(j) = -i\sigma_y, \qquad \phi(k) = -i\sigma_x.$$

The awkwardness to the ordering here is due to the unfortunate historical conventions for the Pauli spin matrices; we will have more to say about this below.

REMARK 5.2.7. Another nice property of the map ϕ is that it essentially preserves the metrics. To be precise, recall that $\mathbb{H}$ has an inner product given by $\langle q_1, q_2\rangle = Re(q_1\overline{q_2})$, where $Re(-)$ is the real part of a quaternion. The group SU_2 has the metric induced by the Hermitian inner product on $M_{2\times 2}(\mathbb{C})$, which is actually real-valued when one restricts to SU_2. It is easy to check that

$$\langle \phi(q_1), \phi(q_2)\rangle = 2\langle q_1, q_2\rangle.$$

In particular, ϕ preserves the orthogonality relation.

If G is a topological group and our goal is to understand group homomorphisms $G \to PSU_2$, then since $PSU_2 \cong SO(3)$ it follows that such group homomorphisms correspond (up to conjugacy) to real representations of G on $\mathbb{R}^3$. In particular, we can apply this to $G = SO(3)$ itself. As it happens, one knows *all* the real representations of $SO(3)$, and the only representations on $\mathbb{R}^3$ are the trivial representation and the standard representation (see Appendix B). So up to isomorphism we know all the projective representations of $SO(3)$ on $\mathbb{C}^2$. They are given by the trivial map $SO(3) \to PSU_2$ and the isomorphism $\tilde{\phi}\colon SO(3) \to PSU_2$ given above.

As we have remarked before, any group homomorphism $SO(3) \to PSU_2$ lifts to a group map $S^3 \to SU_2$ by covering space theory (in this case using that S^3 is simply-connected). The lift of $\tilde{\phi}$ is just the map ϕ that has already been described.

The following result summarizes our discussion so far:

Proposition 5.2.8. *Up to isomorphism there is only one nontrivial, projective, unitary representation of $SO(3)$ on $\mathbb{C}^2$. It corresponds to the honest (non-projective) representation of S^3 on $\mathbb{C}^2$ given by the map $\phi\colon S^3 \to SU_2$.*

The nontrivial projective representation of $SO(3)$ on $\mathbb{C}^2$ may also be described in the following useful way. Recall that every element of $SO(3)$ is a rotation $R_{v,\theta}$ where v is a nonzero vector in $\mathbb{R}^3$ and θ is an angle; we are using the notation from Example 5.1.1. Fixing v, the set of all $R_{v,\theta}$ form a subgroup of $SO(3)$ that is isomorphic to S^1. We therefore obtain a projective representation of S^1 on $\mathbb{C}^2$, which by the previous section we know has the form

$$e^{i\theta} \mapsto \cos(\tfrac{m\theta}{2})I + i\sin(\tfrac{m\theta}{2})J_v$$

for some $m \in \mathbb{Z}$ and some Hermitian, trace zero matrix J_v. We claim that $m = 1$ and that things can be arranged so that if $v = (v_x, v_y, v_z)$ is a unit vector then

$$J_v = -(v_x\sigma_x + v_y\sigma_y + v_z\sigma_z)$$

where σ_x, σ_y, and σ_z are the Pauli spin matrices. To see this we work our way around the square

$$\begin{tikzcd}
S^3 \arrow[r, "\phi"] \arrow[d] & SU_2 \arrow[d] \\
SO(3) \arrow[r] & PSU_2.
\end{tikzcd}$$

We saw in Example 5.1.1 that a preimage of $R_{v,\theta}$ is the quaternion $v(\theta) = \cos(\theta/2) + \sin(\theta/2)[iv_1 + jv_2 + kv_3]$. To apply ϕ we use Remark 5.2.6 to get

$$\begin{aligned}
\phi(v(\theta)) &= \cos(\tfrac{\theta}{2})I + \sin(\tfrac{\theta}{2}) \cdot \Big[v_1\phi(i) + v_2\phi(j) + v_3\phi(k)\Big] \\
&= \cos(\tfrac{\theta}{2})I - i\sin(\tfrac{\theta}{2})\Big[-v_1\sigma_z + v_2\sigma_y + v_3\sigma_x\Big].
\end{aligned}$$

This isn't quite what we wanted, although it is very close. It certainly allows us to identify $m = 1$. To get things in the form of our claim, we need to apply the automorphism c of S^3 that sends $i \mapsto -k$, $j \mapsto j$, and $k \mapsto k$ (this is conjugation by the quaternion $\frac{1}{\sqrt{2}}(1 + j)$).

Remark 5.2.9. We can see at this point that the "best" choice for the Pauli matrices...

We will need to know about the automorphisms of S^3. Note first that there is a group map $S^3 \to \mathrm{Aut}(S^3)$ sending q to the map C_q given by $C_q(x) = qx\bar{q}$ (conjugation by q). The kernel clearly contains ± 1, and we claim this is equality. For one has $C_q = C_r$ if and only if $q^{-1}r$ commutes with all of $\mathbb{H}$, which happens only if $q^{-1}r \in \mathbb{R}$. Since q and r have norm 1, this is equivalent to $q = \pm r$.

It follows that there is an induced map $SO(3) \to \mathrm{Aut}(S^3)$, and this is an injection.

Suppose $f\colon S^3 \to S^3$ is an automorphism. The element -1 is the only non-identity element of S^3 that squares to 1, and so $f(-1) = -1$. The purely imaginary elements of S^3 are the set of all elements that square to -1, so f sends purely imaginary elements to purely imaginary elements. It follows that $f(i)$ and $f(j)$ are purely imaginary elements that anticommute.

Let $P = \{(q_1, q_2) \in S^3 \times S^3 \,|\, q_1^2 = -1 = q_2^2,\ q_1q_2 = -q_2q_1\}$. Then there is a map of spaces $\mathrm{Aut}(S^3) \to P$ that sends f to $(f(i), f(j))$. We claim the composite

$$SO(3) \longrightarrow \mathrm{Aut}(S^3) \longrightarrow P.$$

is a homeomorphism. To see this, identify $\langle i, j, k\rangle$ with $\mathbb{R}^3$ in our usual way. Then P is the set of orthonormal 2-frames in $\mathbb{R}^3$, and the above composite sends a matrix R to the pair consisting of its first two columns. This is a homeomorphism because for a matrix in $SO(3)$ the third column is determined by the first two using the cross product.

PROPOSITION 5.2.10. *The maps $SO(3) \to \mathrm{Aut}(S^3)$ and $\mathrm{Aut}(S^3) \to P$ are both homeomorphisms. As a consequence, every automorphism of S^3 extends to an $\mathbb{R}$-algebra automorphism of $\mathbb{H}$.*

PROOF. There are many different ways to prove this, depending on how much technology one is willing to assume. It is somewhat longwinded, but not difficult, to give a completely naive proof that automorphisms of S^3 are determined by where they map i and j; hence $\mathrm{Aut}(S^3) \to P$ is injective. The reader may wish to work this out as an exercise. [Hint: Use that every purely imaginary quaternion q belongs to a unique 1-dimensional subgroup.] Rather than take this approach, we prove instead that $SO(3) \to \mathrm{Aut}(S^3)$ is surjective using a little Lie theory.

Let $\alpha \in \mathrm{Aut}(S^3)$, and let $\pi\colon S^3 \to SO(3)$ denote our standard projection. Then $\pi\alpha$ and π give two orthogonal representations of S^3 on $\mathbb{R}^3$. Lie theory tells us that there is only one such representation, so these two are isomorphic. Hence there is an $A \in SO(3)$ such that $(\pi\alpha)(x) = A\pi(x)A^{-1}$ for all $x \in S^3$. But $A = \pi(q)$ for some $q \in S^3$, so we have $(\pi\alpha)(x) = \pi(qx\bar{q})$ for all x. This says that C_q and α are both liftings in the diagram

$$\begin{tikzcd}
 & S^3 \arrow[d, "\pi"] \\
S^3 \arrow[ur, dotted] \arrow[r, "\pi\alpha"] & SO(3).
\end{tikzcd}$$

Since both liftings map 1 to 1, covering space theory says they are the same. □

Since we have already seen that S^3 and SU_2 are isomorphic Lie groups, the following is an immediate corollary:

Corollary 5.2.11. *There is a bijection between Lie group isomorphisms $S^3 \to SU_2$ and the set*

$$\{(M_1, M_2) \,|\, M_1, M_2 \in SU_2,\ M_1^2 = M_2^2 = -I,\ M_1M_2 = -M_2M_1\}.$$

The bijection sends $f\colon S^3 \to SU_2$ to $(f(i), f(j))$. Every Lie group isomorphism $S^3 \to SU_2$ extends to a map of $\mathbb{R}$-algebras $\mathbb{H} \to M_{2\times 2}(\mathbb{C})$.

For $R \in S^3$ write $\bar{R}$ for the image of R under our map $S^3 \to SO(3)$. The following result looks odd, and I don't quite know what to say about it. But we will need it in the next section.

Corollary 5.2.12. *Let $f\colon S^3 \to SU_2$ be a group homomorphism, and write $e_1 = i$, $e_2 = j$, $e_3 = k$, and $M_i = f(e_i)$. Then for every $s \in \{1, 2, 3\}$ and every $R \in S^3$ one has*

$$f(R)M_sf(R)^{-1} = \sum_t \bar{R}_{ts}M_t.$$

Proof. This follows from the fact that f extends to a map of $\mathbb{R}$-algebras $\mathbb{H} \to M_{2\times 2}(\mathbb{C})$. One simply computes that

$$f(R)M_sf(R)^{-1} = f(R)f(e_s)f(R)^{-1} = f(Re_sR^{-1}).$$

But recall that $\bar{R}$ is precisely the map that sends e_s to Re_sR^{-1}, so that

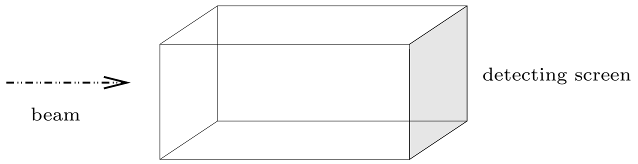
$$Re_sR^{-1} = \sum \bar{R}_{ts}e_t.$$

Applying f to this, and using that f preserves sums, the desired result follows at once. □

Remark 5.2.13. We have stated the above result in the form it will be used later, but it is perhaps best understood as a result about the Lie group S^3 on its own. The adjoint action of S^3 on T_IS^3 is a 3-dimensional real, irreducible representation. The action of S^3 on $\mathbb{R}^3$ via the projection $S^3 \to SO(3)$ is another such representation. As there is only one 3-dimensional, irreducible, real representation of S^3, these must be isomorphic. The formula in the corollary essentially gives the isomorphism. We will let the reader ponder this, without explaining more.

5.3. The story of electron spin

Imagine that you have a box in which there is a certain magnetic field. You pick a fixed type of neutrally charged particle to study (e.g., a fixed type of atom) and you send a beam of these particles into the box. At the far end of the box is a detecting screen, which allows you to see where the particles hit:



Suppose that when you run the experiment you find that there are exactly two locations where the particles hit the screen: let us call them position 1 and position 2. No matter what you do to the particles *before* they enter the box, you find that when they leave the box they always pass through one of those two positions—although you do find that the percentage of particles passing through each position can be made to vary. What conclusions can we make from this? The inevitable conclusion is that our particles come in two types—I suppose we could call them "type 1" and "type 2"— perhaps depending on some kind of internal structure to the particle that we don't understand. But whatever this internal structure may be, we can infer that it interacts with a magnetic field.

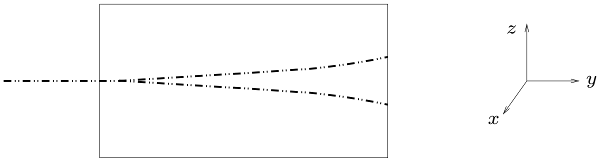
Next you try the experiment with other types of particles. You find that some types leave the box in *three* locations, some in four, some in five, and so on. The set of possibilities is always discrete. You begin to wonder more about what "internal structures" could produce these results.

The experiment can also be carried out for charged particles, although here there is an extra factor one must consider. A charged particle moving through a magentic field will feel a Lorentz force, proportional to its charge and velocity—and this force will add to whatever deflection is being caused by the particles' internal structure. If we are careful, however, we can compensate for this by adding an electric field into the box that will exactly cancel out the Lorentz force. Our experiment is then equivalent to what we did for neutrally charged particles, and we again observe the same type of behavior.

The "internal structure" that is relevant to these experiments is now called **spin**. In analogy with classical physics, physicists think about this structure in terms of the particles having an internal magnetic moment—like a tiny bar magnet. These analogies are not perfect, though: in a beam of tiny bar magnets one would expect the orientation of the magnets to be continuously distributed amongst all possible orientations, and when these internal magnetic moments interact with the magnetic field one would expect a continuous pattern of deflecting particles showing up on the detection screen. This is not what happens. The fact that there are only two orientations possible—and that these orientations are seemingly unaffected by anything we do to the particles before they enter the box—is completely outside the experience of any classical phenomena. It is best not to take these classical analogies too seriously.

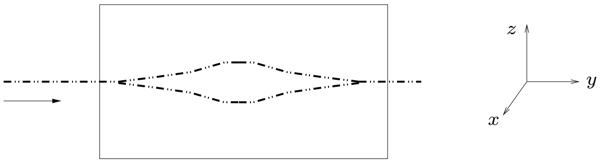
For the moment we would like to completely ignore the physicists' view of spin as an internal magnetic moment, and instead just treat it as a mysterious internal structure that we don't understand. The surprising thing is that even treating it in these terms—as a "black box", so to speak—we can still figure out most of its important properties, almost purely by mathematical reasoning. The following discussion is largely based on Volume III, Chapter 6 of the Feynman lectures [**FLS**].

Let us start by being more specific about the kind of experiment we want to consider. We assume the particles enter the box in a beam directed along the positive y-axis, and that the magnetic field is chosen to be constant in both the x- and y-directions and to have a large gradient in the z-direction. What we find when we do this experiment is that the beam of particles is separated into two beams, deflected in the positive and negative z-directions by the same amount. The particles hit the screen a little bit above and below the y-axis, where "above" and "below" refer to the z-direction.



side view

By coupling one of these boxes with someting like a "mirror-image" box, we could obtain a machine that briefly separates the two beams and then recombines them as follows:



side view

The net effect of this machine is to do nothing: the beam that leaves the box is essentially the same as the beam that enters the box. However, by placing an impenetrable screen in the middle of the machine to block one of the beams, we can obtain a filter: a machine that only lets through one of the two types of particles. Let us call this particular machine an "S-machine". Let us call the particles that get deflected into the positive z-direction the "$+S$" particles, with the opposite type being the "$-S$" particles.

Now we consider a quantum-mechanical model of this situation. We have an observable quantity, namely whether a particle is in a $+S$ state or a $-S$ state. We expect to have a Hermitian operator corresponding to this observable, and we might as well denote this operator as S. There will be exactly two eigenspaces of S, which we *for simplicity* assume are 1-dimensional (this assumption could easily be wrong, but let us stick with it for the moment and see what it leads to). Choose a unit vector from each eigenspace and call them $|{+S}\rangle$ and $|{-S}\rangle$. Our Hilbert space $\mathcal{H}$ of possible states is then the 2-dimensional complex vector space spanned by these two orthonormal basis vectors (we are for the moment ignoring other observables such as position and momentum, instead imagining a toy model in which only the S-state matters).

Note that a typical state of our quantum system has the form $a|+S\rangle + b|-S\rangle$ where $a, b \in \mathbb{C}$ and we may assume $|a|^2 + |b|^2 = 1$ by normalization. This represents a state of the incoming beam in which the percentages of particles in the $+S$ and $-S$ states are $|a|^2$ and $|b|^2$, respectively.

Next consider stringing together two S-machines, as shown below. If we block off the $-S$ beam in the left machine, then the particles coming out of that machine are all guaranteed to be in the $|+S\rangle$ state; they will then *all* be deflected upward by the second machine. This seems clear enough. But now suppose that we rotate the second S-machine through some angle about the y-axis? What happens then? We will find that some percentage of the particles are forced "upward" with respect to the second machine (i.e., in the positive direction of the new z-axis), and the remaining particles are force "downward"—but what exactly are these percentages, and how do we compute them?

To answer these questions let us work in a little more generality. For any oriented, orthonormal frame $\mathcal{F} = (v_1, v_2, v_3)$, let $S_{\mathcal{F}}$ be a machine of type S that has been rotated so that its original positive x-axis is aligned along v_1, its positive y-axis is aligned along v_2, etc. We can just as well denote this machine as S_R, where $R \in SO(3)$ is the rotation matrix whose column vectors are those in $\mathcal{F}$. The eigenspaces for S_R will be two orthogonal complex lines in $\mathcal{H}$; let us denote them L_R^+ and L_R^-. Physicists tend to go ahead and choose unit vectors $|+S_R\rangle$ and $|-S_R\rangle$ in these eigenspaces, but we need to be careful about this because there is no *canonical* choice—in each case there is an S^1's worth of unit vectors. When we were just making one choice, in the case of the original machine S all by itself, there was no big deal. But if we want to make a choice for every $R \in SO(3)$, we need to worry about whether this can be done continuously. Let us avoid the issue for the moment by simply not making these choices.

Let $\mathbb{P}_{1,1}(\mathcal{H})$ denote the space of pairs (K_1, K_2) of orthogonal lines in $\mathcal{H}$. For notational simplicity write $L^+ = L_I^+$ and $L^- = L_I^-$; so L^+ and L^- are the spans of $|+S\rangle$ and $|-S\rangle$. We have a map $F\colon SO(3) \to \mathbb{P}_{1,1}(\mathcal{H})$ that sends $R \in SO(3)$ to the pair (L_R^+, L_R^-). A first goal is to understand what this map looks like. Here are the few things that seem transparent:

- The identity I is sent to the pair (L^+, L^-).
- Every rotation $R_{z,\theta}$ about the z-axis is also sent to the above pair.
- The rotation $R_{y,180}$ is sent to the pair (L^-, L^+).
- For any line l in the xy-plane, the rotation $R_{l,180}$ is also sent to (L^-, L^+).

It turns out that there are many continuous maps $SO(3) \to \mathbb{P}_{1,1}(\mathcal{H})$ having these properties. However, there is an extra condition we might expect our map F to satisfy: since there is no preferred frame of reference in space, we should be able to replace S with a rotated version S_R and get the same results. What F does in a neighborhood of R should look the same as what F does is a neighborhood of the identity. In trying to make this mathematically precise, one is quickly led to the assumption that there is an *action* by $SO(3)$ on $\mathbb{P}_{1,1}(\mathcal{H})$, and that our map F is simply applying the action to the fixed element (L^+, L^-).

Note that if (K_1, K_2) is a pair of orthogonal lines in $\mathcal{H}$, then K_2 is completely determined by K_1; therefore $\mathbb{P}_{1,1}(\mathcal{H})$ is homeomorphic to $\mathbb{P}(\mathcal{H})$. It seems reasonable to assume that the action of $SO(3)$ on $\mathbb{P}_{1,1}(\mathcal{H})$ comes from a unitary action on $\mathbb{P}(\mathcal{H})$, and moreover that this action is linear—in other words, we assume that we have a projective representation of $SO(3)$ on $\mathcal{H}$. From now on we will use our chosen basis $|+S\rangle$, $|-S\rangle$ to identify $\mathcal{H}$ with $\mathbb{C}^2$.

At this point we are looking for group homomorphisms $SO(3) \to PSU_2$ such that

(i) Each rotation $R_{z,\theta}$ is sent to a matrix of the form $\left[\begin{smallmatrix} \alpha & 0 \\ 0 & \bar{\alpha} \end{smallmatrix}\right]$ (really a coset of this matrix in $PSU_2 = SU_2/D$).

(ii) For any line l in the xy-plane, the rotation $R_{l,180}$ is sent to a matrix of the form $\left[\begin{smallmatrix} 0 & \alpha \\ -\bar{\alpha} & 0 \end{smallmatrix}\right]$.

These are equivalent to the conditions $F(R_{z,\theta}) = (L^+, L^-)$ and $F(R_{l,180}) = (L^-, L^+)$. We saw in Section 5.2 that any group homomorphism $SO(3) \to PSU_2$

lifts to a map $\Gamma\colon S^3 \to SU_2$:

$$\begin{tikzcd}
S^3 \arrow[r] \arrow[d, two heads] & SU_2 \arrow[d, two heads] \\
SO(3) \arrow[r] & PSU_2.
\end{tikzcd}$$

We therefore look for homomorphisms $S^3 \to SU_2$ satisfying the properties below:

(i) For any angle θ, the quaternion $(\cos\theta) + (\sin\theta)k$ is sent to a matrix of the form $\begin{bmatrix} \gamma_1 & 0 \\ 0 & \overline{\gamma}_1 \end{bmatrix}$.

(ii) i is sent to a matrix $\begin{bmatrix} 0 & \alpha \\ -\overline{\alpha} & 0 \end{bmatrix}$, and j is sent to a matrix $\begin{bmatrix} 0 & \beta \\ -\overline{\beta} & 0 \end{bmatrix}$.

As discussed in ????, a homomorphism $S^3 \to SU_2$ is specified by giving two matrices $M_1, M_2 \in SU_2$ with the properties that

$$M_1^2 = M_2^2 = -I, \quad M_1M_2 = -M_2M_1.$$

One sets $M_3 = M_1M_2$, and the map $S^3 \to SU_2$ is then given by

$$v_0 + v_1 i + v_2 j + v_3 k \mapsto v_0 I + v_1 M_1 + v_2 M_2 + v_3 M_3.$$

Comparing this to property (ii) above, we see that

$$M_1 = \begin{bmatrix} 0 & \alpha \\ -\overline{\alpha} & 0 \end{bmatrix}, \qquad M_2 = \begin{bmatrix} 0 & \beta \\ -\overline{\beta} & 0 \end{bmatrix},$$

and for $M_1M_2 = -M_2M_1$ we must have that $\frac{\bar{\alpha}}{\alpha} = -\frac{\bar{\beta}}{\beta}$, or equivalently $\alpha^2 = -\beta^2$. The image of k is the matrix

$$M_3 = \begin{bmatrix} -\alpha\bar{\beta} & 0 \\ 0 & -\bar{\alpha}\beta \end{bmatrix},$$

and property (i) follows for free.

Note that there is not a unique homomorphism $S^3 \to SU_2$ of this form: we are free to choose any $\alpha \in S^1$ and then choose β to be either $i\alpha$ or $-i\alpha$. The convention amongst physicists is to take $\alpha = -i$ and $\beta = -1$, really just for historical reasons. The map $S^3 \to SU_2$ is then

$$\begin{aligned} v_0 + v_1 i + v_2 j + v_3 k &\mapsto v_0 I + v_1 \left[\begin{smallmatrix} 0 & -i \\ -i & 0 \end{smallmatrix}\right] + v_2 \left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right] + v_3 \left[\begin{smallmatrix} -i & 0 \\ 0 & i \end{smallmatrix}\right] \\ &= v_0 I - iv_1\sigma_x - iv_2\sigma_y - iv_3\sigma_z. \end{aligned} \tag{5.3.1}$$

Returning to our $SO(3)$ action on $\mathbb{P}(\mathcal{H})$, we have found in particular that the rotations $R_{x,\theta}$, $R_{y,\theta}$, and $R_{z,\theta}$ act as the matrices

(5.3.2)

$$\cos(\tfrac{\theta}{2})I - i\sin(\tfrac{\theta}{2})\sigma_x, \quad \cos(\tfrac{\theta}{2})I - i\sin(\tfrac{\theta}{2})\sigma_y, \quad \text{and} \quad \cos(\tfrac{\theta}{2})I - i\sin(\tfrac{\theta}{2})\sigma_z,$$

respectively.

Remark 5.3.3. Let us now return to the question of whether it is possible to continuously select a basis of eigenvectors $|+S_R\rangle$, $|-S_R\rangle$, for each rotation

$R \in SO(3)$. Recalling that $PSU_2 = PU_2$, this amounts to finding a lifting in the diagram

$$\begin{tikzcd}
& U_2 \arrow[d] \\
SO(3) \arrow[r] \arrow[ur, dotted] & PU_2.
\end{tikzcd}$$

But $SO(3) \to PU_2$ is an isomorphism on π_1 (recall again that $PU_2 = PSU_2$), $\pi_1(SO(3)) = \mathbb{Z}/2$, and $\pi_1 U_2 \cong \mathbb{Z}$: these three facts show that there cannot be such a lifting. It is not possible to continuously select our eigenvectors, for all R at the same time.

If we precompose with $S^3 \to SO(3)$ then of course there *is* such a lifting, in fact we can lift into SU_2:

$$\begin{tikzcd}
& & SU_2 \arrow[d] \\
S^3 \arrow[r] \arrow[urr, dotted] & SO(3) \arrow[r] & PSU_2.
\end{tikzcd}$$

This lifting is just the map Γ constructed earlier. So we *can* continuously select our eigenvectors at the expense of replacing $SO(3)$ by its double cover. For any $q \in S^3$, write $|{+S_q}\rangle$ and $|{-S_q}\rangle$ for this chosen basis (these are just the columns of $\Gamma(q)$).

Remark 5.3.4. You might have been annoyed by the minus signs that appear in equations (5.3.1) and (5.3.2). These arise because physicists prefer to work with coordinate-change matrices rather than the usual matrix of a linear transformation. Suppose $T\colon \mathbb{C}^2 \to \mathbb{C}^2$ is given by $T(x) = Ax$, where A is invertible. Let $f_1 = T(e_1)$ and $f_2 = T(e_2)$, which is a new basis for $\mathbb{C}^2$. For any $v \in \mathbb{C}^2$ write $v = v_1e_1 + v_2e_2$ and also $v = c_1f_1 + c_2f_2$. Then the coordinates (v_1, v_2) and (c_1, c_2) are related by the formula

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = A^{-1} \cdot \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}.$$

We leave this as an easy exercise for the reader. If A is unitary then we can replace A^{-1} with $A^{\dagger}$.

For $v \in \mathcal{H}$, write $C_{+,q}$ and $C_{-,q}$ for the coordinates of v with respect to the basis $|{+S_q}\rangle$, $|{-S_q}\rangle$. When $q = 1$ we omit it from the subscript. Using the observations from the preceding paragraph, we have that

$$\begin{aligned}
\begin{bmatrix} C_{+,q} \\ C_{-,q} \end{bmatrix} = \Gamma(q)^{-1} \cdot \begin{bmatrix} C_+ \\ C_- \end{bmatrix} &= (v_0 I - iv_1\sigma_x - iv_2\sigma_y - v_3\sigma_z)^{\dagger} \begin{bmatrix} C_+ \\ C_- \end{bmatrix} \\
&= (v_0 I + iv_1\sigma_x + iv_2\sigma_y + iv_3\sigma_z) \begin{bmatrix} C_+ \\ C_- \end{bmatrix}.
\end{aligned}$$

A physicist would say that rotating through θ degrees about the x-axis amounts to multiplying the coordinates in $\mathcal{H}$ by the matrix $\cos(\frac{\theta}{2})I + i\sin(\frac{\theta}{2})\sigma_x$, and similarly for the other coordinate axes. Compare this to [**FLS**, Volume III, Table 6.2].

Remark 5.3.5 (Commentary on the Feynman lectures). We recommend that the reader take a look at Feynman's introduction to spin from [**FLS**, Volume III, Chapters 5–6]. It is well done, and it gives a sense of how physicsts think about these things. Here we make a couple of remarks on how parts of [**FLS**, Chapter 6] correspond to the treatment we have given above.

At the end of [**FLS**, Section III.6.2] Feynman gives an argument for why the transformation matrices can all be assumed to have determinant 1. The mathematician reading this will immediately note that while the argument is fine for one matrix at a time, it cannot be made to work for all matrices at once—because a square root cannot be chosen continuously on the complex numbers. In our treatment, this part of the argument corresponds instead to the identification of PU_2 with PSU_2.

Feynman investigates the $SO(3)$-action on $\mathcal{H}$ by first restricting to the three S^1's corresponding to rotations about the coordinate axes. In terms of the underlying physics, rotations about the z-axis do nothing to the physical states, whereas 180 degree rotations about the x- and y-axes will reverse the notion of up and down. He then looks at various composites of rotations and makes arguments about what the action must be. Essentially this amounts to a bare hands proof that there is only one nontrivial, 2-dimensional, projective representation of $SO(3)$. We might phrase things this way: looking at the nontrivial projective action of $SO(3)$ on $\mathbb{C}^2$, we see that the action of each S^1 inside $SO(3)$ is nontrivial. So given a specific action to study, as soon as one knows the nontriviality of the y-rotations this *forces* the nontriviality of the z-rotations. This is the crux of Feynman's argument. Even though the action of z-rotations on the physical states is trivial, the action on the quantum states cannot be.

Physical terminology

So far in this chapter we have given a detailed examination of the eigenspaces of S_R as R ranges through all rotations in $SO(3)$. We have not said anything about the eigen*values*. This is because the exact values here are largely a question of convention: if the Hermitian operator S has eigenvalues $\lambda_\pm$, then for any $a, b \in \mathbb{R}$ the Hermitian operator $aI + bS$ has the same eigenspaces but has eigenvalues equal to $a + b\lambda_\pm$. Thus, by choosing our operator appropriately we can arrange for its eigenvalues to take on any two distinct real values. The convention in physics is to have them be the numbers $\frac{1}{2}$ and $-\frac{1}{2}$. From now on we assume that S has been arranged to have this property. Particles of the type we have studied in this section are called "spin-$\frac{1}{2}$ particles".

Particles in the eigenspace of S for eigenvalue $\frac{1}{2}$ are said to be **spin up** particles, whereas those in the state for eigenvalue $-\frac{1}{2}$ are called **spin down**. More precisely, we should say that these are spin up or spin down *with respect to the frame S*. But recall that if R is a rotation about the z-axis then the eigenspaces of S_R are the same as those of S; so we can actually say spin up or spin down *with respect to the z-axis* without the meaning becoming obscured. The same goes for any oriented line in $\mathbb{R}^3$; we can talk about spin up or spin down particles with respect to this line.

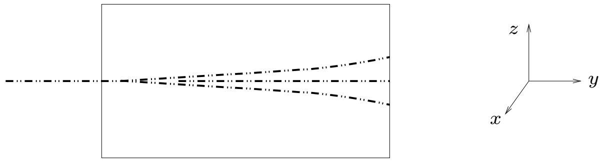
Note that a spin up particle with respect to the z-axis will be neither spin up nor spin down with respect to the x-axis (for example). Rather, the particle corresponds to a state $a|{+x}\rangle + b|{-x}\rangle$ where $|{+x}\rangle$ and $|{-x}\rangle$ are two chosen eigenstates representing spin up and spin down for the x-axis.

The machine S with which we started the discussion is called a **Stern-Gerlach experiment**. The original Stern-Gerlach experiment was done with silver atoms as the particles. The reason for silver is the following: the atom has 47 protons, 61 neutrons, and 47 electrons. It is neutrally charged. The electrons appear in pairs, a spin-up and spin-down electron in each pair, but with one electron left over. The net spin of the atom is dominated by this one unmatched electron (the protons and neutrons also have spin structures, but because of their much heavier masses the effects of these structures on the experiment turn out to be significantly less than the effects of the electron). So essentially the experiment measures the spin of an electron, but in a very clever way. Similar experiments have been done for other atoms with an unmatched electron in the outer shell.

5.4. Spin one and higher

When a beam of electrons is sent into a Stern-Gerlach experiment, it is split into two beams—one is deflected in the up direction and the other in the down direction. We say that electrons come in two types, spin up and spin down with respect to the z-axis of the experiment. There also exist classes of particles that a Stern-Gerlach experiment separates into *three* types—or even more. Our object in this section is to discuss the general theory of such particles.

The following picture shows a Stern-Gerlach experiment separating a beam into three pieces:



side view

Proceeding as in our discussion of spin $\frac{1}{2}$ particles, we postulate that the particles we are dealing with have an internal structure and that the states of such a structure are described by a finite-dimensional Hilbert space $\mathcal{H}$. For the particles from the experiment depicted above this would seem to be a 3-dimensional space, but let us deal with the most general situation we can imagine.

The Stern-Gerlach experiment shows there is some quantity we can measure about the internal state of our particles, and this will correspond to an observable—Hermitian operator S on $\mathcal{H}$. Because the universe allows us to use machines corresonding to different rotations, we must have a projective action of $SO(3)$ on $\mathcal{H}$. The action must be unitary because it must preserve orthogonality between states. Irreducibility?

We have arrived at the mathematical problem of understanding all projective, unitary representations of $SO(3)$ on $\mathbb{C}^n$, for various values of n. In the last section we discussed $n = 2$. Consider what is by now the familiar lifting problem:

$$\begin{tikzcd}
& SU_n \arrow[d, two heads] \\
SO(3) \arrow[ur, dotted] \arrow[r] & PSU_n.
\end{tikzcd}$$

Recall that $\pi_1 SU_n = 0$: to see this, start with $SU_2 \cong S^3$ and then inductively use the fiber sequences $SU_{n-1} \to SU_n \to S^{2n-1}$. So the above lifting exists if and only if the map $\pi_1 SO(3) \to \pi_1 PSU_n$ is trivial. The map $SU_n \to PSU_n$ has fiber μ_n, so it follows that $\pi_1 PSU_n \cong \mathbb{Z}/n$. Our induced map on π_1 therefore has the form $\mathbb{Z}/2 \to \mathbb{Z}/n$, which is automatically trivial when n is odd. So when n is odd, every projective representation comes from an honest representation. When n is even such a lifting may or may not exist, but we are guaranteed a lifting

$$\begin{tikzcd}
S^3 \arrow[r, dotted] \arrow[d] & SU_n \arrow[d] \\
SO(3) \arrow[r] & PSU_n.
\end{tikzcd}$$

At this point we need to recall more from the basic theory of Lie groups. One knows *all* the irreducible complex representations of S^3, and they are described as follows. We use the fact that $S^3 \cong SU_2$. Let H_n be the vector space of complex, homogeneous, degree n polynomials in the formal variables z_1 and z_2. Let SU_2 act on this space by linear combinations of z_1 and z_2 (????). It turns out these are irreducible representations, and they are the only ones. Note that $\dim H_n = n+1$.

Complex representations of $SO(3)$ are precisely complex representations of S^3 on which -1 acts as the identity. The irreducible representations of $SO(3)$ are therefore $H_0, H_2, H_4, \ldots$. Note that H_k has dimension $k+1$.

5.5. Lie algebra methods

We saw in the last section that a particle's internal spin structure is determined by an irreducible representation of S^3. Representations of a simply-connected Lie group are in bijective correspondence with representations of their Lie algebra; the latter should be thought of as the *infinitesimal* representations of the group. Physicists like to use the Lie algebra point-of-view because it gives rise to observables. In this section we briefly review this theory.

Start with a group map $\rho\colon S^3 \to SU_n$, a unitary representation of S^3 on $\mathbb{C}^n$ via matrices of determinant one. Differentiating ρ at the identity gives

$$D\rho\colon T_I S^3 \to T_I SU_n,$$

which is a map of Lie algebras. The tangent space $T_I SU_n$ is the space of trace zero, skew-Hermitian, $n \times n$ matrices, and the Lie algebra structure is the commutator.

We could analyze the Lie algebra $T_I S^3$ directly, but let us instead use the projection $S^3 \to SO(3)$ which is a local diffeomorphism near the identity. So $T_I S^3 \to T_I SO(3)$ is an isomorphism. The Lie algebra for $SO(3)$ is usually denoted $\mathfrak{so}(3)$, and consists of the real, skew-symmetric matrices where the bracket is the commutator. The standard generators are obtained by taking infinitesimal rotations about the three coordinate axes. A counterclockwise rotation through θ degrees about the x-axis has matrix

$$R_{\theta,x} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}.$$

Applying $\frac{d}{d\theta}\big|_{\theta=0}$ gives

$$R_x = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}.$$

One can remember this without the computation by remembering that a small counterclockwise rotation about the x-axis moves the postive y-axis a little bit into the positive z-direction, whereas it moves the positive z-axis a little bit into the negative y-direction. Using similar mnemonics, or else doing the actual calculation, one finds that

$$R_y = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix} \qquad \text{and} \qquad R_z = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

From these matrices one readily checks that $[R_x, R_y] = R_z$, $[R_y, R_z] = R_x$, and $[R_z, R_x] = R_y$. Note that the latter two identities are obtained from the first by cyclic permutation of the symbols x, y, z.

We have at this point determined the Lie algebra $\mathfrak{so}(3)$: as a vector space it is $\mathbb{R}\langle R_x, R_y, R_z\rangle$, and the bracket is given by the above formulas.

Our map $D\rho$ gives a map of Lie algebras $\mathfrak{so}(3) \to \mathfrak{su}(n)$. The target consists of trace zero, skew-Hermitian matrices, but we would prefer to deal with Hermitian matrices because these correspond to physical observables. So set $J_x = -iR_x$, $J_y = -iR_y$, and $J_z = -iR_z$. We then have

$$[J_x, J_y] = iJ_z$$

and the cyclic permutations of this identity. The complex Lie algebra generated by J_x, J_y, and J_z (which is isomorphic to the complexification of $\mathfrak{so}(3)$) is a copy of $\mathfrak{sl}(2,\mathbb{C})$. To see this explicitly, pick one of the J's to be our torus—physicists like to pick J_z. Then set

$$J_+ = J_x + iJ_y, \qquad J_- = J_x - iJ_y.$$

Check that

$$[J_+, J_-] = 2J_z, \qquad [J_z, J_+] = J_+, \qquad [J_z, J_-] = -J_-.$$

The usual description for $\mathfrak{sl}(2,\mathbb{C})$ is in terms of generators e, f, and h subject to the relations $[e, f] = h$, $[h, e] = 2e$, $[h, f] = 2f$. So our isomorphism should have

$$J_+ \longleftrightarrow e, \qquad J_- \longleftrightarrow f, \qquad J_z \longleftrightarrow \tfrac{1}{2}h.$$

At this point we understand that we have an irreducible representation of $\mathfrak{sl}(2,\mathbb{C})$ on $\mathbb{C}^n$. The standard theory of such representations says that the eigenvalues of h will be integers and will take the form $-r, -r+2, -r+4, \dots, r-2, r$ where $r+1 = n$. The e operator raises the eigenvalue by 2, whereas the f operator lowers it by 2. Physicists like to work with J_z instead of h, and the eigenvalues of J_z will be

$$-\tfrac{r}{2}, \quad -\tfrac{r}{2}+1, \quad -\tfrac{r}{2}+2, \quad \dots, \quad \tfrac{r}{2}-1, \quad \tfrac{r}{2}.$$

The operator J_+ raises these eigenvalues by 1 (that is, it sends the λ-eigenspace into the $(\lambda+1)$-eigenspace), whereas J_- lowers them by 1. Physicists write j instead of $\frac{r}{2}$, so we will do this as well.

Pick a unit vector e_1 lying in the eigenspace for $\lambda = j$. Let e_k be the normalization of $(L_-)^{k-1}(e_1)$, which will be a basis for the eigenspace with eigenvalue $j-k+1$. An easy computation shows that

$$J_- e_k = \sqrt{(2j-k+1)k} \cdot e_{k+1} \qquad \text{and} \qquad J_+ e_{k+1} = \sqrt{(2j-k+1)k} \cdot e_k.$$

For example, write $J_- e_1 = \lambda e_2$ and note that we know $\lambda \in \mathbb{R}_{>0}$. Compute

$$\begin{aligned}\lambda^2 = \langle \lambda e_2 | \lambda e_2 \rangle = \langle J_- e_1 | J_- e_1 \rangle = \langle e_1 | J_+ J_- e_1 \rangle &= \langle e_1 | (2J_z + J_- J_+) e_1 \rangle \\ &= \langle e_1 | 2j e_1 \rangle \\ &= 2j.\end{aligned}$$

This determines λ, and at the same time shows that $J_+ e_2 = \frac{1}{\lambda} J_+ J_- e_2 = \frac{2j}{\lambda} e_2$. Now repeat these same steps to analyze $J_- e_3$, and inductively to analyze each $J_- e_k$.

At this point it is routine to determine the action of all of our operators on the basis elements e_k: the only ones we have not done explicitly are J_x and J_y, but here one uses $J_x = \frac{1}{2}(J_+ + J_-)$ and $J_y = \frac{1}{2i}(J_+ - J_-)$. In other words, we can write down the matrices representing each of these operators with respect to the $\{e_i\}$ basis.

Physicists denote the elements of our basis $\{e_k\}$ by $|j,m\rangle$ where $-j \leq m \leq j$ and $j - m \in \mathbb{Z}$. So $|j,m\rangle$ is an eigenvector for J_z with eigenvalue m, and in terms of our notation $|j,m\rangle = e_{j-m+1}$. The basis $\{|j,m\rangle\}_m$ is orthonormal, and we have that

$$\begin{aligned}J_+ |j,m\rangle &= \sqrt{(j+m-1)(j-m)}\ |j,m+1\rangle \\ J_- |j,m\rangle &= \sqrt{(j+m)(j+1-m)}\ |j,m-1\rangle.\end{aligned}$$

Note that $J_+ J_- |j,m\rangle = (j+m)(j+1-m)\,|j,m\rangle$.

Physicists also like to use the operator $\mathbf{J}^2 = J_x^2 + J_y^2 + J_z^2$. Notice that $J_+ J_- = J_x^2 + J_y^2 + J_z$ and so

$$\mathbf{J}^2 = J_+ J_- + J_z^2 - J_z.$$

Applying $\mathbf{J}^2$ to $|j,m\rangle$ therefore gives

$$\mathbf{J}^2 = \big((j+m)(j+1-m) + m^2 - m\big)|j,m\rangle = j(j+1)|j,m\rangle.$$

Since the eigenvalue is independent of m, we see that $\mathbf{J}^2$ acts diagonally on $\mathbb{C}^n$ as scalar multiplication by $j(j+1)$.

The operator $\mathbf{J}^2$ is called the **total angular momentum operator**, whereas J_x, J_y, and J_z are the operators for **angular momentum about the *x*-, *y*-, and *z*-axes**.

. This gives an orthonormal basis for $\mathbb{C}^n$, and with respect to this basis the matrix for J_z is

$$J_z = \begin{bmatrix} j & 0 & \cdots & 0 \\ 0 & j-1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & -j \end{bmatrix}$$

For $j = \frac{1}{2}$ the matrices look as follows:

$$J_z = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{bmatrix}, \quad J_+ = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad J_- = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad J_x = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad J_y = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

For $j = 1$ they look like

$$J_z = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \quad J_+ = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad J_- = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

$$J_x = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad J_y = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Based on these examples the reader will find it a simple matter to write down the J-matrices corresponding to any value of j.

5.6. The Schrödinger-Pauli equation

Up until now we have discussed the quantum spin states of a particle that sits in one physical position and is frozen in time. Such spin states are given by an irreducible, unitary representation of the group S^3. In this section our goal is to develop the theory for a more typical quantum particle—one that spreads its existence throughout space as a probability wave—and then to let time flow. For spin-$\frac{1}{2}$ particles this leads to the differential equation of the section's title.

Let $\mathcal{H}$ denote the space of quantum states for a spin-$\frac{1}{2}$ particle: it is a two-dimensional complex vector space. For the purposes of our discussion fix an orthonormal basis for $\mathcal{H}$ and call the vectors U and D; these might be the states corresponding to spin up and spin down with respect to the z-axis, for example.

A particle without an internal spin state is described by a wave function $\psi\colon \mathbb{R}^3 \to \mathbb{C}$. Each complex number $\psi(\mathbf{x})$ is the probability amplitude for the particle to be measured at location $\mathbf{x}$, and ψ is assumed to belong to $L^2(\mathbb{R}^3, \mathbb{C})$. But for a spin-$\frac{1}{2}$ particle we need more information: at each location $\mathbf{x}$ there should be given probability amplitudes for detecting the particle in either the U or D states. In other words, the quantum state of our spin-$\frac{1}{2}$ particle should be described by a wave function $\psi\colon \mathbb{R}^3 \to \mathcal{H}$. Note that such a wave function is equivalent to having two functions $\psi_1, \psi_2\colon \mathbb{R}^3 \to \mathbb{R}$, where $\psi(\mathbf{x}) = \psi_1(\mathbf{x})U + \psi_2(\mathbf{x})D$. So the wave function for a spin-$\frac{1}{2}$ particle may be regarded as a pair of ordinary wave functions, assembled into a vector $\left[\begin{smallmatrix} \psi_1(\mathbf{x}) \\ \psi_2(\mathbf{x}) \end{smallmatrix}\right]$; this is the way things are often portrayed in introductory physics texts.

The Hilbert space of quantum states for a spin-$\frac{1}{2}$ particle will be taken to be $\mathcal{H}_{spin} = L^2(\mathbb{R}^3, \mathcal{H})$. Note that there is an isomorphism

$$L^2(\mathbb{R}^3, \mathcal{H}) \cong L^2(\mathbb{R}^3, \mathbb{R}) \otimes \mathcal{H},$$

and this allows us to regard elements of $\mathcal{H}_{spin}$ in two different ways. We will go back and forth between these, usually without comment.

Remark 5.6.1. We have elected to use the wave function model of quantum states in our present discussion, but recall from Chapter ???? that this can be avoided. For a particle in $\mathbb{R}^3$ without spin, we must have a Hilbert space with orthonormal basis given by $|\mathbf{x}\rangle$ for $\mathbf{x} \in \mathbb{R}^3$. A state ψ will then yield amplitudes $\langle\psi|\mathbf{x}\rangle$, and the wave function is simply the collection of all these amplitudes. For a spin-$\frac{1}{2}$ particle our orthonormal basis must instead consist of vectors $|\mathbf{x}, U\rangle$ and $|\mathbf{x}, D\rangle$ for $\mathbf{x} \in \mathbb{R}^3$. A state ψ will then give rise to amplitudes $\langle\psi|\mathbf{x}, U\rangle$ and $\langle\psi|\mathbf{x}, D\rangle$. So it is possible to have our entire discussion abstractly, without ever mentioning wave functions at all. Note that whatever Hilbert space $\mathcal{V}$ we are using to model the states of a spinless particle, it is clearly $\mathcal{V} \otimes \mathcal{H}$ that models the states of our particle with spin.

We next discuss observables, which will be Hermitian operators on $\mathcal{H}_{spin}$. It seems reasonable enough that the momentum of a particle should not involve its spin amplitudes at all, and so we postulate that the operator for momentum in the x-direction (say) is simply $i\hbar\frac{\partial}{\partial x} \otimes \mathrm{id}$. Note that in writing this we use the tensor product description of $\mathcal{H}_{spin}$. If we choose a basis for $\mathcal{H}$ and use it to denote elements of $\mathcal{H}$ as column vectors, then we are simply saying that P_x is the operator

$$\begin{bmatrix} i\hbar\frac{\partial}{\partial x} & 0 \\ 0 & i\hbar\frac{\partial}{\partial x} \end{bmatrix}.$$

At this point we have described the quantum states for spin-$\frac{1}{2}$ particles in $\mathbb{R}^3$. The next goal is to let time flow. Recall that this is accomplished by writing down an appropriate Hamiltonian H (a Hermitian operator on $\mathcal{H}_{spin}$) and then wave functions change in time via the formula

$$\psi_t = e^{-\frac{i}{\hbar}Ht}\psi.$$

If we write $\psi(\mathbf{x}, t) = \psi_t(\mathbf{x})$, then this is equivalent to the corresponding Schrödinger equation

$$i\hbar\tfrac{\partial\psi}{\partial t} = H\psi.$$

The challenge is to write down the correct Hamiltonian for modelling the physics.

For a particle of mass m in free space physicists use the Hamiltonian $H = \frac{1}{2m}(P_x^2 + P_y^2 + P_z^2)$, which in our case is the 2×2 matrix

$$H = -\frac{\hbar^2}{2m}\begin{bmatrix} \nabla^2 & 0 \\ 0 & \nabla^2 \end{bmatrix}.$$

This says that as time flows the spin states of the particle evolve independently, just as if there were no spin at all. This agrees with what one observes in experiments.

Things become trickier when there is an electromagnetic field in the picture. Now the particle's spin structure interacts with the field, and that must be built into our Hamiltonian. Classically, physicists learned to do this in the following way. Pick a vector potential $\mathbf{A}$ for the magnetic field (so that $\operatorname{curl}\mathbf{A} = \mathbf{B}$) and a scalar potential ϕ for the electric field (so that $-\nabla\phi = \mathbf{E} + \frac{\partial\mathbf{A}}{\partial t}$). See the discussion in Section 4.3.2 if you don't recall this. The classical Hamiltonian for a particle of charge q moving in our electromagnetic field is

$$H = \frac{1}{2m}\sum_j (p_j - \tfrac{q}{c}A_j)^2 + q\phi.$$

Note that this reduces to our usual $P^2/2m$ when $\mathbf{A}$ and ϕ are both zero.

Based on the above, one's first guess for the quantum operator on $\mathcal{H}_{spin}$ is

$$H = \left[\frac{1}{2m}\sum_j \bigl(i\hbar\tfrac{\partial}{\partial x_j} - \tfrac{q}{c}A_j\bigr)^2 + q\phi\right] \otimes I. \tag{5.6.2}$$

Here the operator in brackets is really acting on $L^2(\mathbb{R}^3, \mathbb{R})$ and the I is the identity operator on $\mathcal{H}$. If we identify $\mathcal{H}_{spin}$ with length 2 column vectors of ordinary wave functions, then the above operator is a scalar matrix whose diagonal entries are the bracketed operator. The problem with this choice for H is that it does not differentiate among particles with different spin states: a spin-up particle and a spin-down particle will have the same behavior. Our Stern-Gerlach experiments show this to be false, this was the whole point from the beginning.

Pauli introduced a correction to the above Hamiltonian that incorporates the spin structure. The Pauli Hamiltonian is

$$H = \frac{1}{2m}\sum_j \Big((-i\hbar\tfrac{\partial}{\partial x_j} - \tfrac{q}{c}A_j)\sigma_j\Big)^2 + (q\phi)I \tag{5.6.3}$$

where the σ_i's are the Pauli spin matrices and I is the 2×2 identity matrix (here we are, as usual, identifying $\mathcal{H}_{spin}$ with column vectors of two ordinary wave functions). It will take us a while to answer the obvious question:"Why choose this particular form?" For the moment let us just accept the form and play with it a bit.

Quite generally, note that

$$\Big(\sum_j a_j\sigma_j\Big)^2 = \Big(\sum_j a_j^2\Big)I + \sum i[a_j, a_k]\sigma_l$$

where in the second sum the triple (j,k,l) runs over cyclic permutations of $1,2,3$. Noting that

$$\Big[i\hbar\tfrac{\partial}{\partial x_j} + \tfrac{q}{c}A_j, i\hbar\tfrac{\partial}{\partial x_k} + \tfrac{q}{c}A_k\Big] = \tfrac{i\hbar q}{c}\Big(\tfrac{\partial A_k}{\partial x_j} - \tfrac{\partial A_j}{\partial x_k}\Big),$$

we find that

$$\begin{aligned} H &= \Big[\frac{1}{2m}\sum_j(-i\hbar\tfrac{\partial}{\partial x_j} - \tfrac{q}{c}A_j)^2 + q\phi\Big]I - \tfrac{\hbar q}{2mc}\sum_j(\operatorname{curl}\mathbf{A})_j\sigma_j \\ &= \Big[\frac{1}{2m}\sum_j(-i\hbar\tfrac{\partial}{\partial x_j} - \tfrac{q}{c}A_j)^2 + q\phi\Big]I - \tfrac{\hbar q}{2mc}\sum_j B_j\sigma_j. \end{aligned}$$

The term $\sum_j B_j\sigma_j$ is usually written $\mathbf{B}\cdot\boldsymbol{\sigma}$ in the physics literature, where $\boldsymbol{\sigma}$ stands for the formal expression $\sigma_x\hat{\mathbf{i}} + \sigma_y\hat{\mathbf{j}} + \sigma_z\hat{\mathbf{k}}$. This term accounts for the interaction of the particle's spin structure with the magnetic field. My understanding of this term was that it was introduced on a somewhat *ad hoc* basis by Pauli, but was shown by him to lead to conclusions that agreed with experiment. Later Feynman observed that Pauli's term could be produced by writing the Hamiltonian in the form of (5.6.3), and this was recorded in Sakurai's textbook [**S1**]. Of course the form in (5.6.3) is still *ad hoc* in its own way; we will explore some motivation behind it in the next section.

Although we won't need it, the reader might be wondering about what the Pauli Hamiltonian looks like if you expand the squared terms. For the record, here it is:

$$\begin{aligned} H = \frac{1}{2m}\Big(-\hbar^2\nabla^2 + \tfrac{q^2}{c^2}(\mathbf{A}\cdot\mathbf{A}) + \tfrac{i\hbar q}{c}(\nabla\cdot\mathbf{A}) + \tfrac{2i\hbar q}{c}\mathbf{A}\cdot\operatorname{grad} + q\phi\Big)I \\ - \tfrac{\hbar q}{2mc}\sum_j(\operatorname{curl}\mathbf{A})_j\sigma_j. \end{aligned}$$

5.6.4. Rotational invariance and the Pauli Hamiltonian. The Pauli Hamiltonian came out of nowhere: one takes the naive Hamiltonian and then inserts the σ_i's into it, with little motivation for doing this. We can find some motivation, though, by thinking about rotational invariance. This is somewhat of a long topic, so bear with me as I back up for a moment.

Let us first ignore spin, so that our particles have wave functions $\psi\colon \mathbb{R}^3 \to \mathbb{R}$. We have an $SO(3)$-action on $L^2(\mathbb{R}^3, \mathbb{R})$ defined by

$$R.\psi = \psi \circ R^{-1}, \quad \text{or} \quad [R.\psi](\mathbf{x}) = \psi(R^{-1}\mathbf{x})$$

for $R \in SO(3)$. Note that the R^{-1} is necessary to get a left action. The symmetry of physics tells us that if H is our Hamiltonian then we would expect

$$e^{-\frac{i}{\hbar}Ht}[R.\psi] = R.[e^{-\frac{i}{\hbar}Ht}\psi]. \tag{5.6.5}$$

In other words, the time development of a rotated wave function should be the same as the time development of the original function, but rotated. Equation (5.6.5) is equivalent to

$$H[R.\psi] = R.[H\psi] \tag{5.6.6}$$

(for one direction, take the derivative with respect to t at $t = 0$).

Equation (5.6.6) is a guiding principle in determining the Hamiltonian H. Does $H = \frac{1}{2m}(P_x^2 + P_y^2 + P_z^2)$ have this property? Up to constants this is just the operator ∇^2. One could verify the property by direct computation, but it is a little less painful (and ultimately more useful) to break the computation up into pieces. Recall the chain complex

$$0 \to C^\infty(\mathbb{R}^3) \xrightarrow{grad} (\text{Vec. fields on } \mathbb{R}^3) \xrightarrow{curl} (\text{Vec. fields on } \mathbb{R}^3) \xrightarrow{div} C^\infty(\mathbb{R}^3) \to 0.$$

We have talked about the $SO(3)$-action on $C^\infty(\mathbb{R}^3)$, defined by $R.f = f \circ R^{-1}$. In the case of functions $F\colon \mathbb{R}^3 \to \mathbb{R}^3$ there is an $SO(3)$-action on both the domain and target, and these should both be taken into account when defining an $SO(3)$-action on functions. One defines

$$[R.F](\mathbf{x}) = R[F(R^{-1}\mathbf{x})].$$

In general, if S and T are sets with a left G-action then one defines a left G-action on $\operatorname{Hom}(S, T)$ by $(g.f)(s) = g[f(g^{-1}x)]$. This is all we are doing.

Now that we have $SO(3)$-actions on $C^\infty(\mathbb{R}^3)$ and on the space of vector fields $\mathbb{R}^3 \to \mathbb{R}^3$, it is natural to ask if the above operators preserve this action. Indeed, they do:

Proposition 5.6.7. *Let $f\colon \mathbb{R}^3 \to \mathbb{R}$ and $F\colon \mathbb{R}^3 \to \mathbb{R}^3$ be smooth. For any $R \in SO(3)$ one has*

$$\operatorname{grad}(R.f) = R.[\operatorname{grad} f], \quad \operatorname{curl}(R.F) = R.\operatorname{curl}(F), \qquad \operatorname{div}(R.F) = R.\operatorname{div}(F).$$

That is, the operators grad, curl, *and* div *all preserve the natural $SO(3)$-action. As a consequence, the composite operator $\nabla^2 = \operatorname{div} \circ \operatorname{grad}$ also preserves the $SO(3)$-action.*

Proof. This is completely clear if you think about the geometric properties of grad, curl, and div. For example, grad(f) is characterized by the property that for $x, u \in \mathbb{R}^3$ one has

$$f(x + u) \approx f(x) + (\operatorname{grad} f) \cdot u$$

as the first-order approximation. We then can write

$$\begin{aligned}
[R.f](x + u) = f(R^{-1}(x + u)) &= f(R^{-1}x + R^{-1}u) \\
&\approx f(R^{-1}x) + [(\operatorname{grad} f)(R^{-1}x)] \cdot (R^{-1}u) \\
&= f(R^{-1}x) + [R(\operatorname{grad} f)(R^{-1}x)] \cdot u
\end{aligned}$$

This shows that $\operatorname{grad}[R.f](x) = R\cdot[(\operatorname{grad} f)(R^{-1}x)]$, or more succinctly $\operatorname{grad}[R.f] = R.[\operatorname{grad} f]$.

The divergence is characterized by $(\operatorname{div} F)(x) = \frac{1}{V}\iiint_V (F\cdot\hat{n})\,dV$ for "infinitesimal" volumes V about x. Therefore

$$\begin{aligned}(\operatorname{div} R.F)(x) = \frac{1}{\text{vol}}\iiint\limits_V (RFR^{-1}\cdot\hat{n}_V)\,dV &= \frac{1}{\text{vol}}\iiint\limits_{R^{-1}V} (RF\cdot R\hat{n}_{R^{-1}V})\,dV\\ &= \frac{1}{\text{vol}}\iiint\limits_{R^{-1}V} (F\cdot\hat{n}_{R^{-1}V})\,dV\\ &= (\operatorname{div} F)(R^{-1}x).\end{aligned}$$

A similar analysis works for curl, using that $(\operatorname{curl} F)(x)\cdot u$ is the infinitesimal circulation of F, at x, counterclockwise about an axis directed along u. We leave the reader to work out the argument. □

Remark 5.6.8. ????

At this point we have shown that the Hamiltonian $\frac{1}{2m}(P_x^2 + P_y^2 + P_z^2)$ has the $SO(3)$-invariance property from (5.6.6). Even if we hadn't had an idea of what the Hamiltonian should be, if we went looking for differential operators with the invariance property it wouldn't be long before we found it.

Remark 5.6.9. IMPORTANT WARNING! In the rest of this section we will simplify formulas by ignoring most physical constants: not only things like $\hbar$ and c, but also mass and charge. This will let us better concentrate on the underlying mathematics, which is the main point for the moment. Occasionally we may reinsert some constants in order to make a pedagogical point.

Now let us add an electromagnetic field to the picture, but continue to neglect spin. Recall that the field is specified by the magnetic and electric potentials $\mathbf{A}$ and ϕ. Let us denote the Hamiltonian for this field by $H_{(\mathbf{A},\phi)}$, or just $H_{\mathbf{A}}$ for short. Note that we would *not* expect $e^{iH_At}(R.\psi) = R.[e^{iH_At}\psi]$, because a rotated wave function is going to look very different from the perspective of the fixed em-field. But if we rotate *both* the wave function and the field, the physics should be the same. This is expressed via the equations

$$e^{iH_{RA}t}[R.\psi] = R.[e^{iH_At}\psi], \qquad \text{or} \qquad H_{RA}[R.\psi] = R.[H_A\psi]$$

(recall our present policy of ignoring all physical constants).

We know from our experience in E-M theory that the Hamiltonian H_A is supposed to be obtained from the free-particle Hamiltonian by changing p_i to $p_i - qA_i$, and also adding a term $q\phi$. To deal with the first of these changes, let us introduce the formal symbol

$$\nabla_A = \left(\tfrac{\partial}{\partial x} - A_1\right)\hat{\mathbf{i}} + \left(\tfrac{\partial}{\partial y} - A_2\right)\hat{\mathbf{j}} + \left(\tfrac{\partial}{\partial z} - A_3\right)\hat{\mathbf{k}}$$

and define the operators

$$\operatorname{grad}_A f = \nabla_A f, \quad \operatorname{curl}_A F = \nabla_A \times F, \quad \operatorname{div}_A F = \nabla_A\cdot F.$$

These are operators

$$0 \to C^\infty(\mathbb{R}^3) \xrightarrow{\operatorname{grad}_A} (\text{Vec. flds on } \mathbb{R}^3) \xrightarrow{\operatorname{curl}_A} (\text{Vec. flds on } \mathbb{R}^3) \xrightarrow{\operatorname{div}_A} C^\infty(\mathbb{R}^3) \to 0.$$

Note that

$$\operatorname{grad}_A f = (\operatorname{grad} f) - f\mathbf{A},\ \operatorname{curl}_A \mathbf{F} = (\operatorname{curl} \mathbf{F}) - \mathbf{A} \times \mathbf{F},\ \operatorname{div}_A F = (\operatorname{div} \mathbf{F}) - \mathbf{A} \cdot \mathbf{F}.$$

Remark 5.6.10. Warning: The above operators do NOT form a chain complex. Indeed, one can compute

$$\begin{aligned}\operatorname{curl}_A \operatorname{grad}_A f &= \operatorname{curl}(\operatorname{grad} f - f\mathbf{A}) - \mathbf{A} \times (\operatorname{grad} f - f\mathbf{A})\\ &= -\operatorname{curl}(f\mathbf{A}) - \mathbf{A} \times (\operatorname{grad} f)\\ &= -[(\operatorname{grad} f) \times \mathbf{A} + f(\operatorname{curl} \mathbf{A})] - \mathbf{A} \times (\operatorname{grad} f)\\ &= -f(\operatorname{curl} \mathbf{A}).\end{aligned}$$

A similar computation shows that

$$\operatorname{div}_A \operatorname{curl}_A F = -(\operatorname{curl} \mathbf{A}) \cdot \mathbf{F}.$$

The appearance of $\operatorname{curl} \mathbf{A}$ in the expressions for both composites is not an accident. In fact, in the operators grad_A, curl_A, and div_A we are seeing the edge of a much bigger picture involving connections and curvature on vector bundles. This story will be developed in Chapter 6.

Although they do not form a chain complex, the new operators we have introduced do have the desired invariance properties:

Proposition 5.6.11. *For $R \in SO(3)$ one has*

$$\operatorname{grad}_{RA}(R.f) = R.(\operatorname{grad}_A f), \quad \operatorname{curl}_{RA}(R.F) = R.[\operatorname{curl}_A F],$$
$$\textit{and} \quad \operatorname{div}_{RA}(R.F) = R.[\operatorname{div}_A F].$$

Proof. These are all easy computations. For example,

$$\begin{aligned}\operatorname{grad}_{RA}(R.f) = \operatorname{grad}(R.f) - (R.f)\mathbf{A} &= R.[\operatorname{grad} f] - (f \circ R^{-1})\mathbf{A}\\ &= R.[\operatorname{grad} f] - (f\mathbf{A}) \circ R^{-1}\\ &= R.[\operatorname{grad} f] - R.(f\mathbf{A})\\ &= R.[\operatorname{grad}_A f].\end{aligned}$$

□

If we take our Hamiltonian to be $H_A = \operatorname{div}_A \circ \operatorname{grad}_A$, then it has the invariance property that we desire. This is appropriate if the electric potential ϕ is equal to zero. For a nonzero electric potential we define

$$H_A = (\operatorname{div}_A \circ \operatorname{grad}_A) + q\phi. \tag{5.6.12}$$

Note that the term $q\phi$ clearly has the correct invariance property, as

$$q[R.\phi] \cdot [R.\psi] = q(\phi \circ R^{-1})(\psi \circ R^{-1}) = (q\phi\psi) \circ R^{-1} = R.[q\phi\psi].$$

So $H_{(RA,R\phi)}[R.\psi] = R.[H_{(A,\phi)}\psi]$, which is what we wanted.

We are now ready for the final step of this long exploration of rotational invariance. The Hamiltonian we wrote down in (5.6.12) is perfectly fine for a particle with no internal spin structure: one whose wave function has the form $\psi\colon \mathbb{R}^3 \to \mathbb{R}$ and the $SO(3)$-action is therefore only acting on the domain. For a particle with spin the wave function instead has the form $\psi\colon \mathbb{R}^3 \to \mathcal{H}$. We have $SO(3)$-acting on the domain as usual, but now we have $SO(3)$ *almost* acting on the target—it acts on the associated projective space, and we know to study this by instead looking at a related S^3-action on $\mathcal{H}$. We then have S^3 acting on everything in sight (on $\mathcal{H}$ in the usual way, and on $\mathbb{R}^3$ via the projection $S^3 \to SO(3)$). So we should look for a Hamiltonian that has an S^3-invariance.

As we have discussed, the S^3-action on $L^2(\mathbb{R}^3, \mathcal{H})$ is given by

$$[R.\psi](x) = R[\psi(R^{-1}X)]$$

where $R \in S^3$. Such formulas will get slightly unwieldy, and to help with this we introduce the following notation: for $v \in \mathcal{H}$ we will write $R \cdot v$ and Rv interchangeably, but we will never write $R \cdot \psi$ for $R.\psi$. Note that under this scheme the string of symbols "$R \cdot \psi(x)$" only has one interpretation, namely $R[\psi(x)]$.

We look for a Hamiltonian H_A—a Hermitian operator on $L^2(\mathbb{R}^3, \mathcal{H})$—with the property that $H_{RA}(R\psi) = R[H_A(\psi)]$ for all $\psi \in L^2(\mathbb{R}^3, \mathcal{H})$ and all $R \in S^3$. Recall that $L^2(\mathbb{R}^3, \mathcal{H}) \cong L^2(\mathbb{R}^3, \mathbb{R}) \otimes \mathcal{H}$, and this is an isomorphism of S^3-representations if the tensor product is given the diagonal action of S^3. It is clear, then, that if we take $H = (\mathrm{div}_A \circ \mathrm{grad}_A + q\phi) \otimes \mathrm{id}$ then this has the required S^3-invariance property—simply because the term in parentheses did. This is the Hamiltonian of equation (5.6.2). It is a perfectly good, rotationally invariant Hamiltonian—but it describes particles whose "spin" structure does not interact with magnetic fields (if such things existed).

It seems reasonable to modify the Hamiltonian in the above paragraph by adding on one or more terms that capture the spin/field interaction. For the new Hamiltonian to have the S^3-invariance property, the stuff to be added on must itself have the invariance property. So our task is to search for more invariants; luckily they are not hard to find.

5.7. Spin and relativity: mathematical foundations

Dirac created a relativistic, quantum theory of the electron. To his surprise, electron spin came about as a *consequence* of imposing relativistic invariance. The other amazing consequence of Dirac's theory was the existence of antimatter. In this section we develop some of the mathematical foundations needed to understand Dirac's equation.

Recall that in special relativity the main group of symmetries is the Lorentz group $O(3,1)$. This is the group of linear automorphisms $\mathbb{R}^4 \to \mathbb{R}^4$ that preserve the Minkowski metric. We saw in Section 4.1 that $O(3,1)$ has four path components. The connected component of the identity is called the restricted Lorentz group, and denoted $SO^+(3,1)$.

Some of the biggest physical discoveries of the mid-twentieth century were that the laws of physics are not invariant under the full Lorentz group. In physics, the sign of the determinant is called the *parity*; in 1957 Wu and her collaborators discovered parity violation in radiocative beta decay. This showed that there was a fundamental difference between left and right in the universe. In 1964 the violation of time-reversal symmetry was found by Christensen, Cronin, Fitch, and Turlay in the decay of the neutral K-meson. To be precise, what they actually found was violation of "CP-symmetry", meaning the symmetry where one changes both charge and parity. But CPT-symmetry—where one changes charge, and parity, and the direction of time—is believed to be full symmetry of physical laws, and so violation of CP-symmetry implies violation of T-symmetry. We do not intend to give much discussion of these topics, only to whet the reader's appetite.

As far as anyone currently knows, the laws of physics *are* invariant under the group $SO^+(3,1)$. When physicists talk about "Lorentz invariance" this is what they mean. We will therefore mostly deal with this group, which we will call L for short.

Note that there is an evident inclusion $SO(3) \hookrightarrow L$. We will show below that this inclusion is a homotopy equivalence; therefore $\pi_1(L) = \mathbb{Z}/2$, and the universal cover $\tilde{L} \to L$ has degree 2. We will identify $\tilde{L}$ with the group $SL(2,\mathbb{C})$. The story from here closely parallels what we have already seen for $S^3 \to SO(3)$. The projective representations of L are intimately related to the honest representations of $\mathrm{SL}(2,\mathbb{C})$, and the latter are understood completely. This summarizes the main points of what will be covered in the present section.

We begin by constructing a group homomorphism $SL(2,\mathbb{C}) \to L$. Recall the Pauli spin matrices σ_1, σ_2, σ_3. To these we add $\sigma_0 = I$. Then these matrices form a basis for the vector space $M^h_{2\times 2}(\mathbb{C})$ of 2×2 Hermitian matrices. Define an isomorphism $\alpha\colon \mathbb{R}^4 \to M^h_{2\times 2}(\mathbb{C})$ by

$$\alpha(x_0,\ldots,x_3) = x_0\sigma_0 + x_1\sigma_1 + x_2\sigma_2 + x_3\sigma_3.$$

Notice that the Lorentz norm of $\mathbf{x}$ coincides with the determinant of $\alpha(\mathbf{x})$.

If $A \in SL(2,\mathbb{C})$ then $X \mapsto AXA^\dagger$ is a map $M^h_{2\times 2}(\mathbb{C}) \to M^h_{2\times 2}(\mathbb{C})$ that preserves the determinant. Under α this therefore corresponds to a Lorentz transformation. That is to say, define $\phi\colon SL(2,\mathbb{C}) \to O(3,1)$ by

$$\phi(A) = \big[\mathbf{x} \mapsto \alpha^{-1}(A\cdot\alpha(\mathbf{x})\cdot A^\dagger)\big].$$

Since $SL(2,\mathbb{C})$ is connected, the image must land inside of L; so we have actually defined $\phi\colon SL(2,\mathbb{C}) \to L$. It is easy to see that ϕ is a group homomorphism. A matrix A is in the kernel if and only if $A\sigma_i A^\dagger = \sigma_i$ for all i. When $i=0$ this says $AA^\dagger = I$, and then the other conditions are equivalent to $A\sigma_i = \sigma_i A$ for $i>0$. It is easy to see that this happens only when $A = \pm I$.

We claim that $SL(2,\mathbb{C}) \to L$ is surjective. To see this, let $\alpha, \beta \in \mathbb{R}$ be such that $\alpha^2 - \beta^2 = 1$ and let $A = \left[\begin{smallmatrix}\alpha & \beta\\ \beta & \alpha\end{smallmatrix}\right]$. The $X \mapsto AXA^\dagger$ fixes σ_2 and σ_3, and it sends σ_0 to $(\alpha^2+\beta^2)\sigma_0 + 2\alpha\beta\sigma_1$. The corresponding Lorentz transformation is an x-boost, and one readily checks that any x-boost in L (meaning one that preserves the direction of time) can be obtained in this way. The preservation of the time direction is reflected in the fact that $\alpha^2+\beta^2$ is always positive.

Next consider the two covering spaces

$$\begin{tikzcd}
\mathbb{Z}/2 \arrow[r, tail] \arrow[d, equal] & SL(2,\mathbb{C}) \arrow[r, two heads] & L \\
\mathbb{Z}/2 \arrow[r, tail] & SU_2 \arrow[r, two heads] \arrow[u, hook] & SU_2/\pm I. \arrow[u, hook]
\end{tikzcd} \tag{5.7.1}$$

Recall that $SU_2/\pm I \cong SO(3)$. The image of the composite $SU_2 \hookrightarrow SL(2,\mathbb{C}) \to L$ is easily checked to land in the subgroup of Lorentz transformations that fix the time coordinate, this subgroup being precisely $SO(3)$. With only a little trouble one checks that $SU_2/\pm I \to L$ maps the domain isomorphically onto this subgroup.

The Lorentz group L is generated by the x-boosts together with $SO(3)$, both of which have been shown to be in the image of ϕ. So ϕ is surjective.

At this point we have shown that $SL(2,\mathbb{C}) \to L$ is surjective with kernel $\{I, -I\}$. Recall that the inclusion $SU_2 \hookrightarrow SL(2,\mathbb{C})$ is a deformation retraction (use

Gram-Schmidt), and that $SU_2 \cong S^3$. In particular, $SL(2,\mathbb{C})$ is simply-connected; therefore $SL(2,\mathbb{C})$ is the universal cover of L. Consider the two covering spaces in (5.7.1), recalling that $SU_2/\pm I \cong SO(3)$. Since $SU_2 \hookrightarrow SL(2,\mathbb{C})$ is a homotopy equivalence, it follows from the two long exact sequences of homotopy groups (and the 5-Lemma) that $SO(3) \hookrightarrow L$ is a weak equivalence. (Of course it is therefore a homotopy equivalence as well, since both spaces may certainly be given the structure of CW-complexes).

At this point we have verified that $SL(2,\mathbb{C}) \twoheadrightarrow L$ is the universal cover, and that this map is weakly equivalent to $S^3 \to SO(3)$. Just as for the latter map, it follows that every projective representation of L lifts to an honest representation of $SL(2,\mathbb{C})$. Let us next consider what the representations of $SL(2,\mathbb{C})$ actually are.

Let $V = \mathbb{C}^2$ with the evident action of $SL(2,\mathbb{C})$. We have the symmetric powers $\operatorname{Sym}^k(V)$, with their induced action; these representations are irreducible because we already know they are irreducible when restricted to SU_2.

We also have the conjugate representation $\bar{V}$. This is $\mathbb{C}^2$ but where $X \in SL_2(\mathbb{C})$ acts by left multiplication by $\bar{X}$ (there are other descriptions of $\bar{V}$, but they are isomorphic to this one). Over SU_2 one has $V \cong \bar{V}$, but this is not true over $SL(2,\mathbb{C})$. To see this, note that a vector space isomorphism is simply a map $v \mapsto Qv$ where Q is an invertible 2×2 matrix. The condition that this be compatible with the $SL(2,\mathbb{C})$-actions is that $QXv = \bar{X}Qv$ for every X in $SL(2,\mathbb{C})$ and every $v \in \mathbb{C}^2$, or equivalently just that $QX = \bar{X}Q$ for every $X \in SL(2,\mathbb{C})$. Taking X to be a diagonal matrix $\left[\begin{smallmatrix} z & 0 \\ 0 & z^{-1} \end{smallmatrix}\right]$, one finds readily that the only possible Q is the zero matrix which is of course not invertible. (Over SU_2 one only has diagonal matrices as above where $z\bar{z} = 1$, and in this case one finds that $Q = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$ is allowed.)

Clearly $\bar{V}$ must be irreducible over $SL(2,\mathbb{C})$, since a decomposition would yield a corresponding decomposition for V by conjugating everything. Likewise, the symmetric products $\operatorname{Sym}^k(\bar{V})$ are also irreducible.

Generalizing this line of thought somewhat, we have the following:

THEOREM 5.7.2. *For $k, l \geq 0$ the representations $H_{k,l} = \operatorname{Sym}^k(V) \otimes \operatorname{Sym}^l(\bar{V})$ are all irreducible, and they are a complete list of irreducible representations for $SL(2,\mathbb{C})$.*

Note that the dimension of $H_{k,l}$ is $(k+1)(l+1)$. If v_1 and v_2 are formal variables, the representation $H_{k,l}$ may be regarded as having a basis consisting of monomials $v_1^a v_2^b \bar{v}_1^c \bar{v}_2^d$ where $a + b = k$ and $c + d = l$. A matrix $X \in SL(2,\mathbb{C})$ acts on such a monomial by the substitution $v_1 \mapsto x_{11}v_1 + x_{21}v_2$, $v_2 \mapsto x_{12}v_1 + x_{22}v_2$.

The smallest nontrivial, irreducible representations of $SL(2, C)$ are $H_{1,0}$ and $H_{0,1}$. They are both 2-dimensional. These are usually called the representations of **left- and right-handed spinors**. The representation $H_{1,1} = H_{1,0} \otimes H_{0,1}$ is called the **Dirac representation**, or the representation of **Dirac spinors**. Note that it is 4-dimensional.

The representations of L are precisely the representations of $SL(2,\mathbb{C})$ for which $-I$ acts as the identity. Among the irreducible representations these are the $H_{k,l}$ for which either k and l are both even, or k and l are both odd. Notice that $H_{1,1}$ is the smallest nontrivial irreducible representation of L.

5.8. The Dirac equation

CHAPTER 6

Gauge theory

6.1. Principal bundles

The examples in the previous two sections suggest that once quantum mechanics enters the picture the magnetic potential plays a more fundamental role than was evident from Maxwell's equations. In the next few sections our goal will be to develop a geometric setting where we can recast Maxwell's equations, having precisely the effect that the magnetic potential assumes a greater importance. This is the language of principal bundles and connections. It will turn out that the magnetic potential may be thought of as a connection on a principal $U(1)$-bundle, and the electromagnetic field may be obtained as the associated curvature.

We should admit that if our only goal was understanding electromagnetism, then pursuing these generalizations would probably not be worth the effort. The real payoff comes where the group $U(1)$ is replaced by some other Lie group G; in this case the analogs of Maxwell's equations are called the Yang-Mills equations, and the subject as a whole goes under the name *gauge theory*.

6.1.1. Principal bundles. Let G be a topological group. A **principal G-bundle** is a fiber bundle $\pi\colon P \to B$ together with a G-action $G \times P \to P$ such that

(1) $\pi(gx) = \pi(x)$ for all $g \in G$ and $x \in P$;
(2) For all $x \in P$, the map $G \to P_x$ given by $g \mapsto gx$ is a homeomorphism;
(3) For each $b \in B$ there exists an open set $b \in U \subseteq B$ and a G-equivariant homeomorphism

$$\begin{tikzcd}
\pi^{-1}(U) \arrow[rr, "\cong"] \arrow[dr] & & G \times U \arrow[dl] \\
& U. &
\end{tikzcd}$$

The way we have written this definition involves some redundancy: condition (3) implies both condition (2) and the fact that $P \to B$ is a fiber bundle. We have written it this way because the heirarchy of ideas seems more natural: a principal G-bundle is a fiber bundle carrying a G-action, where the fibers are homeomorphic to G, and which has an evident locally-trivial property.

Example 6.1.2. The usual projection $S^n \to \mathbb{R}P^n$ is a principal $\mathbb{Z}/2$-bundle, where S^n has the antipodal action. Likewise, if we write $S(\mathbb{C}^{n+1})$ for the unit sphere in $\mathbb{C}^{n+1}$ then the projection $S(\mathbb{C}^{n+1}) \to \mathbb{C}P^n$ is a prinicpal $U(1)$-bundle.

Suppose T is a right G-space and $\pi\colon P \to B$ is a principal G-bundle. Define

$$T \times_G P = (T \times P)/\sim$$

where the equivalence relation is generated by $(ug, x) \sim (u, gx)$ for $u \in T$, $g \in G$, and $x \in P$. We have a map $T \to B$ given by $[(u, x)] \mapsto \pi(x)$, and one can check that this is a fiber bundle with fiber T. This construction should be thought of as taking the bundle P and gluing T in as the fiber.

Example 6.1.3. Let $\mathbb{Z}/2$ act on $\mathbb{R} - 0$ by the sign action. Then there is an isomorphism of bundles over $\mathbb{R}P^n$

$$\begin{tikzcd}
(\mathbb{R} - 0) \times_{\mathbb{Z}/2} S^n \arrow[rr, "\cong"] \arrow[dr] & & \mathbb{R}^{n+1} - 0 \arrow[dl] \\
& \mathbb{R}P^n &
\end{tikzcd}$$

given by $[(\lambda, x) \mapsto \lambda x$.

Likewise, if $E = \{(L, x) \,|\, L \in \mathbb{R}P^n,\ x \in L\}$ then $E \to \mathbb{R}P^n$ is a rank 1 vector bundle. There is an isomorphims of bundles over $\mathbb{R}P^n$

$$\mathbb{R} \times_{\mathbb{Z}/2} S^n \xrightarrow{\cong} E$$

given by $[(\lambda, x)] \mapsto (\langle x \rangle, \lambda x)$.

Let V be a real vector space of dimension n. Define $\mathrm{Fr}(V)$ to be the subset of V^n consisting of tuples $(e_1, \dots, e_n)$ which are bases of V. We call $\mathrm{Fr}(V)$ the **space of frames** of V. Note that $\mathrm{Fr}(V)$ carries a natural left $GL(V)$-action given by $f.(e_1, \dots, e_n) = (f(e_1), \dots, f(e_n))$. However, $\mathrm{Fr}(V)$ also carries a natural right $\mathrm{GL}_n(\mathbb{R})$-action:

If $E \to B$ is a rank n vector bundle then we can look at the collection of frames in each fiber, and this gives a new bundle over B. Define

$$\mathrm{Fr}(E) = \{(b, e_1, \dots, e_n) \,|\, b \in B, (e_1, \dots, e_n) \in \mathrm{Fr}(E_b)\}.$$

One readily checks that the projection $\mathrm{Fr}(E) \to B$ is a principal $\mathrm{GL}_n(\mathbb{R})$-bundle.

At this point we have maps back and forth

$$\begin{tikzcd}
(\text{rank } n \text{ vector bundles over } X) \arrow[r, shift left, "\mathrm{Fr}"] & (\text{pr. } \mathrm{GL}_n(\mathbb{R})\text{-bundles over } X) \arrow[l, shift left, "\mathbb{R}^n \times_{\mathrm{GL}_n(\mathbb{R})} (-)"]
\end{tikzcd}$$

and one can check that these maps induce inverse bijections on the sets of isomorphism classes.

6.1.4. Reduction of the structure group. Let $E \to B$ be a rank n vector bundle and let $H \to \mathrm{GL}_n(\mathbb{R})$ be a continuous homomorphism of topological groups (in practice this will usually be an inclusion). We say that "the structure group of E can be reduced to H" if there exists a principal H-bundle $P \to B$ and an isomorphism of bundles $E \cong \mathbb{R}^n \times_H P$.

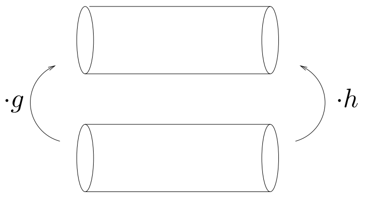
Example 6.1.5. Let $H = \mathrm{SL}_n(\mathbb{R})$. A vector bundle's structure group can be reduced to H if and only if the bundle is orientable in the usual sense.

6.1.6. Homotopy classification of principal bundles. Fix a topological group G. It is a theorem that one can produce a pointed space BG, called the classifying space for G, such that there are natural bijections

$$(\text{iso. classes of principal } G\text{-bundles over } X) \cong [X, BG]_*$$

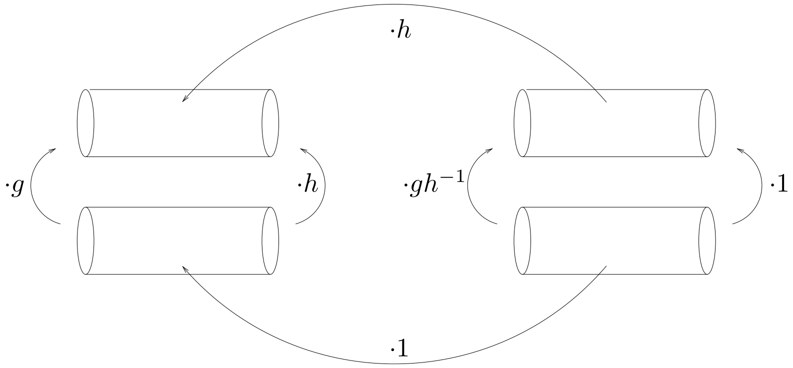
for any pointed space X. We won't need the full strength of this result, so for the moment we will be content with explaining a weaker version.

Let us explore principal G-bundles over S^1. In general, if $g, h \in G$ then we can construct a bundle $P(g,h)$ by gluing two copies of $G \times I$ together in the following way:



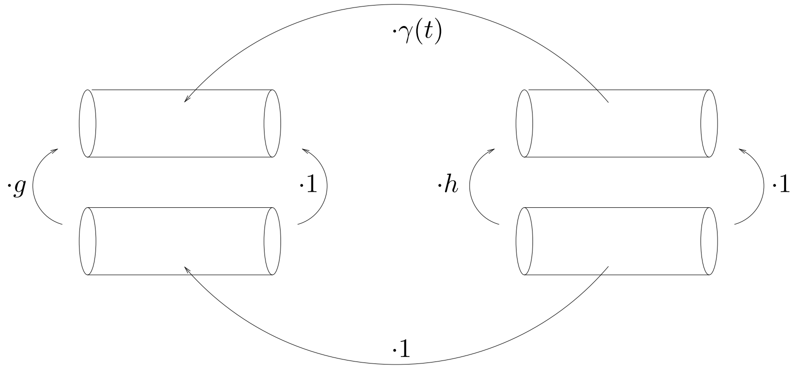
Here the two ends are glued together via *right* multiplication by g and h, respectively. It is important to use right multiplication so that there is still a left G-action on the identification space.

Since any principal G-bundle on S^1 must be trivializable on the top and bottom hemispheres of S^1 we can conclude that every principal G-bundle on S^1 is of the form $P(g,h)$ for some $g, h \in G$. This classification is somewhat overdetermined, however. For instance, it is easy to see that $P(g,h) \cong P(gh^{-1}, 1)$, via the following picture:



The picture is supposed to indicate an isomorphism that is the identity on the lower $G \times I$ and right multiplication by h on every fiber of the upper $G \times I$. One readily checks that this is indeed well-defined and compatible with the left G-action, and it is clearly an isomorphism.

A very similar picture shows that if g and h lie in the same path component of G then $P(g,1) \cong P(h,1)$. The assumption implies that $h^{-1}g$ is in the same path component as 1, so let $\gamma\colon I \to G$ be a path such that $\gamma(0) = h^{-1}g$ and $\gamma(1) = 1$. Consider the isomorphism



Here again, the isomorphism is just the identity on the lower $G \times I$. On the fiber over t in the upper $G \times I$, it is right multiplication by $\gamma(t)$. One readily checks that this works.

We conclude that we have a surjection

$$\pi_0(G) \twoheadrightarrow (\text{iso. classes of pr. } G\text{-bundles on } S^1)$$

sending the component of g to $P(g, 1)$. The reader can readily convince himself that this is a bijection.

The above analysis (even the pictures) readily generalizes to apply to principal G-bundles on a suspension ΣX. Let C_+X and C_-X denote the top and bottom cones forming ΣX.

(1) If $f\colon X \to G$ is any continuous map, define a bundle $P(f)$ by cluing $C_+X \times G$ to $C_-X \times G$ along their bases by $(x, g) \mapsto (x, gf(x))$. Check that $P(g)$ is a principal G-bundle over ΣX, and that every principal G-bundle on ΣX is isomorphic to one of this form.
(2) Let $x \in X$ be a chosen basepoint. Prove that $P(f) \cong P(f \cdot f(x)^{-1})$. Therefore every principal G-bundle on ΣX is isomorphic to one of the form $P(g)$ where $g\colon X \to G$ is a pointed map ($g(x) = 1$).
(3) Prove and if $f, g\colon X \to G$ are pointed maps which are homotopic relative to the basepoint, then $P(f) \cong P(g)$.

The above steps produce a surjection

$$[X, G]_* \twoheadrightarrow (\text{iso. classes of principal } G\text{-bundles over } \Sigma X).$$

With a little more work one can prove that this is a bijection.

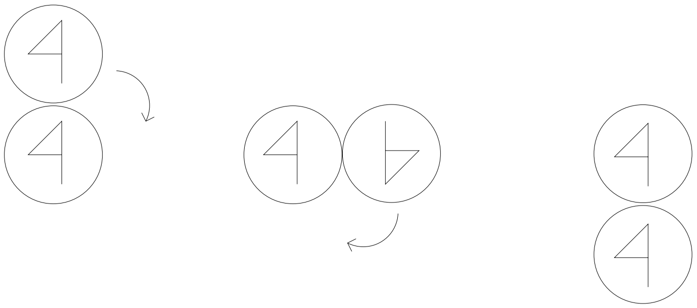
6.1.7. An extended example. Consider vector bundles over S^2, keeping in mind that S^2 is a suspension. The above techniques give us a sequence of bijections

$$(\text{rank 2 v.b. over } S^2) \cong (\text{principal } GL_2(\mathbb{R})\text{-bundles over } S^2) \cong [S^1, GL_2(\mathbb{R})]_*.$$

The inclusion $O(2) \hookrightarrow GL_2(\mathbb{R})$ is a homotopy equivalence (by Gram-Schmidt), and as a topological space $O(2) \cong S^1 \amalg S^1$. It follows that $\pi_1(GL_2(\mathbb{R}), I) \cong \mathbb{Z}$. So the different isomorphism classes of rank 2 vector bundles over S^2 can be classified by the elements of the set $\mathbb{Z}$.

This is all very nice, but can we use it in practice? Given a specific rank 2 bundle like the tangent bundle $TS^2 \to S^2$, can we determine the corresponding integer?

We claim that the integer corresponding to TS^2 is ± 2 (we are not going to worry about signs), and that this is demonstrated by the following procedure: take two quarters and place one directly above the other as shown in the figure below. Keep the bottom quarter fixed and roll the top quarter along its edge.

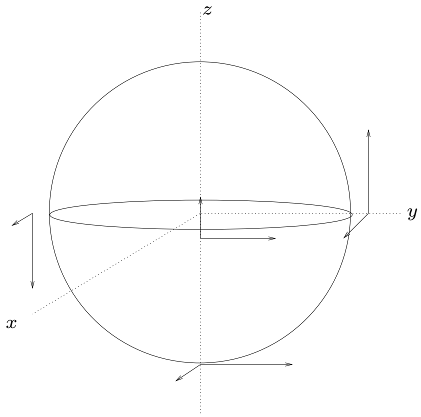


[Note: Rather than massacre George Washington I have replaced him with the the number 4].

As you watch the top quarter rotate, you will find that by the time it arrives at the point directly below its original position, it has rotated a complete 360 degrees. Keep rolling, and by the time it returns to its original position it will now have rotated through 360 degrees exactly *twice*. We claim that this is why TS^2 corresponds to ± 2 under the above bijections. [This demonstration is an old trick: I learned it from Haynes Miller when I was a graduate student.]

At this point you are probably wondering, “What the heck just happened?” Because it is not at all clear *why* the quarter demonstration has anything to do with what we’ve been talking about. So we now explain.

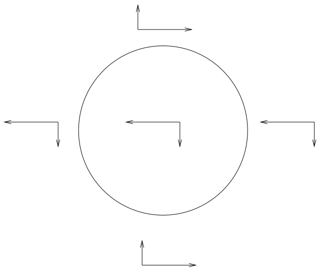
Start by choosing trivializations of TS^2 over the top and bottom hemispheres. This amounts to specifying a frame at each point, and one nice way to do this is to choose a frame at the south pole and then parallel transport it to each point by following the geodesic. For example (see the picture below) choose the $\hat{\mathbf{i}}, \hat{\mathbf{j}}$ frame at the south pole. Transporting this to the east pole gives the $\hat{\mathbf{i}}, \hat{\mathbf{k}}$ frame, while transporting it to the west pole gives the $\hat{\mathbf{i}}, -\hat{\mathbf{k}}$ frame. Transporting it to the “front” pole gives $\hat{\mathbf{k}}, \hat{\mathbf{j}}$. This process trivializes TS^2 over the bottom hemisphere.



Before doing the same thing over the top hemisphere, let's take advantage of the fact that we're doing topology: so we can be very flexible about what we mean by "hemisphere". Let us take the "bottom hemisphere" to be everything except a very small disk around the north pole. And let us have the "top hemisphere" just be that small disk.

Choose a frame at the north pole—we will choose the $\hat{\mathbf{i}}, -\hat{\mathbf{j}}$ frame because it nearly coincides with the frame already chosen at our new "east pole" (which is very near the north pole). We again transport this frame to other points on the small disk by moving along geodesics, but the disk is so small that one really doesn't have to think about this.

Now we need to look along the "equator" (really the boundary of our small disk) and see how the northern framing patches with the southern framing. This will give us the map $f\colon S^1 \to SO(2)$ for which $\mathrm{Fr}(TS^2) \cong P(f)$. Focusing on our small disk and looking down on it from above, we get a picture like the following:



Here the frame inside the disk represents the trivialization over the northern hemisphere (which we may as well assume is constant, with respect to our picture), and the framing outside the disk represents the southern trivialization. How do they relate? On the east pole they are identical, but as we move around the circle they start to differ by rotations. The map $f\colon S^1 \to SO(2)$ gives at each point the rotation that is necessary to make the inner and outer frames coincide, and clearly the S^1 will be wrapped around $SO(2)$ exactly twice.

Hopefully the connection with the quarter demonstration is clear: the portrait of Washington represents a framing, and one watches how it rotates as the outer quarter rotates. [What is not so clear, at least to me, is whether the demonstration is really anything more than a mnemonic for remembering the above argument. The connection between the two is a little tenuous—after all, it was important that it was two quarters rather than a quarter and a penny, and yet how did we *know* that we should use two quarters?]

6.2. Connections and curvature

In this section we get into the language of differential geometry. For me personally, this has not been an easy language to learn. ?????

6.2.1. Systems of parallel transport. Let $E \to B$ be a vector bundle of rank n. The following "definition" is not exactly rigorous, but it nevertheless captures some important intuition. By a **system of parallel transport** on the bundle E we mean an assignment which to every interal $[a,b]$ and every smooth path $\gamma\colon [a,b] \to B$ associates a linear map $\Phi_\gamma\colon E_{\gamma(0)} \to E_{\gamma(1)}$, such that the following properties hold:

(1) $\Phi_{\alpha\beta} = \Phi_\alpha \circ \Phi_\beta$ where $\alpha\beta$ is the usual juxtaposition of α and β (first do β then follow with α).
(2) $\Phi_{c_x} = \mathrm{Id}$ where c_x is the constant path at a point $x \in B$.
(3) $\Phi_\gamma = \Phi_{\bar{\gamma}}^{-1}$ where $\bar{\gamma}\colon [a,b] \to B$ is defined by $\bar{\gamma}(t) = \gamma(b - t + a)$.
(4) Φ_γ varies smoothly with γ.

It is the last condition which is difficult to make rigorous. People know how to do it, but it would take us far afield to try to write it down. Nevertheless, the *idea* of a system of parallel transport is clear enough.

If the bundle E is equipped with extra structure, we might ask that the parallel transport maps respect that structure. For instance, if E has a metric then we will ask that the maps Φ_γ be isometries.

Example 6.2.2. Consider the tangent bundle $TS^2 \to S^2$, with the usual metric. Given a tangent vector at x and a geodesic passing through x, one had an intuitive notion of what it means to transport the tangent vector along the geodesic in a way that keeps it parallel. If one wants to parallel transport the tangent vector along an arbitrary smooth path, one imagines approximating the path by a concantenation of small geodesics and transporting the vector along each piece.

Notice that systems of parallel transport are not homotopical notions. That is, if paths α and β are homotopic relative to the endpoints there is no reason for Φ_α and Φ_β to be equal. This can be seen explicitly in the example of parallel transport on S^2: different geodesics between the north and south poles clearly give rise to different parallel transports of tangent vectors.

Suppose that $\gamma\colon I \to B$ and $v \in E_{\gamma(0)}$. One gets a path $\tilde{\gamma}\colon I \to E$ that lifts γ by having

$$\tilde{\gamma}(t) = \Phi_{\gamma_{\leq t}}(v)$$

where $\gamma_{\leq t}$ is the restriction of γ to $[0,t]$ (recoordinatized to $[0,1]$ if necessary). We call such a $\tilde{\gamma}$ a **parallel** lifting of γ. More generally, if X is any smooth manifold and $X \to B$ a smooth map, a lifting $X \to E$ will be called parallel if it is parallel when restricted to any smooth path in X.

6.2.3. Affine connections. The idea now is that, rather than talk about systems of parallel transport, we will talk about something that is easier to define rigorously but that captures equivalent concepts. This is the notion of affine connection. To get the general idea, suppose that s is a section of a vector bundle $E \to B$, $x \in B$, and $v \in T_xB$. Choose a smooth path $\gamma\colon [-1,1] \to B$ such that $\gamma(0) = x$ and $\gamma'(0) = v$. If we had a system of parallel transport on E, we could attempt to define the derivative of s in the direction of v by a formula such as

$$D_v(s)_x = \lim_{t\to 0}\left[\frac{\Phi^{-1}_{\gamma|_{[0,t]}}(s(t)) - s(0)}{t}\right].$$

Of course one would need to prove that the limit exists and that it is independent of the choice of γ, but if we have the "right" definitions these arguments should be the familiar ones from calculus. The only real point here is that the Φ_γ maps are allowing one to move between different fibers of E: the vectors $s(t)$ and $s(0)$ live in different fibers, but using Φ we can push them into the same fiber.

The notion of affine connection is simply an axiomization of what one needs to be able to differentiate sections along tangent vectors. There are different ways one could encode this information, but the following is the most common:

Definition 6.2.4. *Let $E \to B$ be a smooth vector bundle with B a smooth manifold. An* ***affine connection*** *on E is an assignment*

$$\Gamma(TB) \times \Gamma(E) \to \Gamma(E), \qquad (\mathfrak{X}, s) \mapsto D_{\mathfrak{X}}s$$

having the following properties:

(1) $D_{a_1\mathfrak{X}_1+a_2\mathfrak{X}_2}(s) = a_1D_{\mathfrak{X}_1}(s) + a_2D_{\mathfrak{X}_2}(s)$, for any $a_1, a_2 \in C^\infty(B)$, $\mathfrak{X}_1, \mathfrak{X}_2 \in \Gamma(TB)$.

(2) $D_{\mathfrak{X}}(s+t) = D_{\mathfrak{X}}(s) + D_{\mathfrak{X}}(t)$ for any $s, t \in \Gamma(E)$.

(3) $D_{\mathfrak{X}}(f \cdot s) = (\partial_{\mathfrak{X}}f) \cdot s + f \cdot D_{\mathfrak{X}}(s)$ for any smooth map $f\colon B \to \mathbb{R}$.

Basic facts about connections

(1) For $b \in B$, the value $[D_{\mathfrak{X}}s](b)$ only depends on the restriction of $\mathfrak{X}$ and s to a neighborhood of b. In other words, given $\mathfrak{X}'$ and s' such that $\mathfrak{X}'|_U = \mathfrak{X}|_U$ and $s|_U = s'|_U$ for some neighborhood U of b, then $[D_{\mathfrak{X}}s](b) = [D_{\mathfrak{X}'}s'](b)$.
(2) Even more, if $\mathfrak{X}(b) = \mathfrak{X}'(b)$ then $[D_{\mathfrak{X}}s](b) = [D_{\mathfrak{X}'}s](b)$. If v is any tangent vector to B at b, we will write $(D_vs)(b)$ to denote $[D_{\mathfrak{X}}s](b)$ for any vector field $\mathfrak{X}$ satisfying $\mathfrak{X}(b) = v$.
(3) If $\gamma\colon [-1,1] \to B$ is a smooth curve such that $\gamma(0) = b$ and $\gamma'(0) = v$, then $[D_vs](b)$ only depends on the values of s along γ. In other words, if s and s' are two sections such that $s \circ \gamma = s' \circ \gamma$, then $[D_vs](b) = [D_vs'](b)$.
(4) Suppose that $E' \hookrightarrow E$ is a subbundle which admits a bundle retraction $r\colon E \to E'$. Then the connection D on E induces a connection D' on E' by the formula
$$D'_v(s) = r[D_v(s)].$$
(5) If D and D' are two connections on E and $t \in [0,1]$, then $tD + (1-t)D'$ is again a connection on E.

(6) Every bundle admits a connection.

Less basic facts about connections

(1) If D and D' are any two connections on E, then $D - D'$ lies in $\Omega^1(B; \underline{\mathrm{End}}(E))$. If $A \in \Omega^1(B; \underline{\mathrm{End}}(E))$ then $D + A$ is a new connection. In short, if D is a fixed connection on E then the space of all connections may be identified with $D + \Omega^1(B; \underline{\mathrm{End}}(E))$.

    The space $\Omega^1(B; \underline{\mathrm{End}}(E))$ may be regarded as a parameter space for all connections on E. Note that this space is contractible (it is an infinite-dimensional vector space). Elements of $\Omega^1(B; \underline{\mathrm{End}}(E))$ will be called **connection potentials**.

(2) If E is trivial and $s_1, \dots, s_n$ is a basis of sections, then one may define a connection of E by the formula
$$D_v\Bigl(\sum a_i s_i\Bigr) = \sum_i \partial_v(a_i)s_i.$$
In other words, one simply differentiates the s-coordinates of a given section. This is the unique connection satisfying $D_v(s_i) = 0$ for all v and i. We call it the **standard flat connection** on E. However, note that it depends on the choice of basis $s_1, \dots, s_n$; the use of the adjective "standard" can therefore be a little disorienting.

(3) If E and F are bundles with connections D^E and D^F, then a map of bundles $h\colon E \to F$ is **compatible with the connections** if $h(D^E_v s) = D^F_v(hs)$ for all tangent vectors TB and sections s.

(4) Given connections D^E on E and D^F on F, there is a standard induced connection on $E \otimes F$ given by the formula
$$D^{E\otimes F}_v(s \otimes t) = D^E_v(s) \otimes t + s \otimes D^F_v(t).$$
There is also a standard induced connection on $\underline{\mathrm{Hom}}(E, F)$, defines as follows. If α is a section of $\underline{\mathrm{Hom}}(E, F)$ and s is a section of E, then define $D_v(\alpha)$ to be the section given by
$$(D^{\underline{\mathrm{Hom}}(E,F)}_v \alpha)(s) = D^F_v(\alpha(s)) - \alpha(D^E_v s).$$
Note that this is the unique connection such that the evaluation map of bundles $\underline{\mathrm{Hom}}(E, F) \otimes E \to F$ becomes compatible with the induced connections.

(5) Given local sections $s_1, \dots, s_n$ of E forming a basis, one can write
$$D_v(s_i) = \sum_j \omega_{ij}(v)s_j$$
for uniquely-defined coefficients $\omega_{ij}(v)$. The ω_{ij} may be regarded as 1-forms, and the matrix $\Omega = (\omega_{ij})$ is called the **connection matrix for D with respect to the local basis** $s_1, \dots, s_n$. Note that the 1-forms are only defined on the neighborhood where the s_i's are defined.

    If $e_1, \dots, e_n$ is another local basis of sections for E, defined on a neighborhood overlapping that domain of the s_i's, then we may write $e_i = \sum_j g_{ij}s_j$ for certain real-valued functions g_{ij}. One may check the identity
$$\Omega^e = (dg)g^{-1} + g\Omega^s g^{-1}$$

where Ω^e and Ω^s are the connection matrices for D with respect to the two local bases.

Facts about Curvature

(6) Given a connection D on a smooth bundle $E \to B$, there is an associated 2-form $R^D \in \Omega^2(B; \underline{\mathrm{End}}(E))$.

(7) If E is trivial and D is a standard flat connection with respect to some choice of basis for E, then $R^D = 0$. If in addition A is a connection potential then

$$R^{D+A} = dA + A \wedge A.$$

(8) If Ω is the connection matrix for D with respect to some basis of sections $s_1, \dots, s_n$, then

$$R^D = d\Omega - \Omega \wedge \Omega.$$

de Rham theory with coefficients in a bundle

(9) If V is any finite-dimensional real vector space, then there is a natural isomorphism $\Omega^p(B; V) \cong \Omega^p(B) \otimes V$. In particular we have maps $d\colon \Omega^p(B;V) \to \Omega^{p+1}(B;V)$ obtained by tensoring the usual deRham differential with the identity on V. Note that $d^2 = 0$ as usual.

(10) Now suppose that $E \to B$ is a smooth bundle with connection D. We may define a map $d_D \colon \Omega^0(B;E) \to \Omega^1(B;E)$ by the formula

$$d_D(s) = [v \mapsto D_v(s)].$$

Note that the information in d_D is really the same information as in D, just reorganized.

We extend d_D to maps $\Omega^p(B;E) \to \Omega^{p+1}(B;E)$ by the Leibniz rule. Elements of $\Omega^p(B;E)$ are linear combinations of elements of the form $\omega \otimes s$ where $\omega \in \Omega^p(B)$ and $s \in \Gamma E$, and we define

$$d_D(\omega \otimes s) = (d\omega) \otimes s + (-1)^p \omega \wedge d_D(s).$$

(11) One computes that $d_D^2(\omega) = R_D \wedge \omega$, for any $\omega \in \Omega^p(B;E)$.

(12) $d_{\mathrm{End}(E)}(R_D) = 0$. This is called the Bianchi identity.

(13) If D is the standard flat connection on a trivial bundle, then d_{D+A} is ????

(14) $\mathrm{tr}(d_D(\sigma)) = d(\mathrm{tr}\,\sigma)$ for $\sigma \in \Gamma(\mathrm{End}(E))$.

6.2.5. A quick and dirty introduction to Lie theory.

Definition 6.2.6. *A **Lie group** is a topological group G which is also a smooth manifold, where the multiplication and inverse maps are smooth.*

The theory of Lie groups is of course very well-developed and beautiful, but here we will only cover the basics for what we will need. In particular, it will suffice for us to restrict our attention entirely to matrix groups: that is, we will assume G is a subgroup of $\mathrm{GL}_n(\mathbb{R})$.

One sets $\mathfrak{g} = T_1(G)$, the tangent space at the identity. Since $G \hookrightarrow M_n(\mathbb{R}) \cong \mathbb{R}^{n^2}$, we have $\mathfrak{g} = T_1 G \hookrightarrow T_1(M_n(\mathbb{R})) = M_n(\mathbb{R})$. Here is a useful fact that works for matrix groups:

Given $A \in \mathfrak{g}$, the path $t \mapsto e^{tA}$ is a path in G whose derivative at $t = 0$ is A. This fact allows us to give a precise description for $\mathfrak{g}$ in any specific example.

Example 6.2.7. Let $G = SO(n)$. If $A \in \mathfrak{g}$ then $e^{tA} \in SO(n)$, which is true if and only if $e^{tA}(e^{tA})^T = I$. Expanding these exponentials as powers gives $I + t(A + A^T) + t^2(???) + \cdots = I$, and since this holds for all t we must have $A + A^T = 0$ (differentiate both sides at $t = 0$). It's not hard to convince oneself that this is actually an *equivalent* condition to $A \in \mathfrak{g}$. That is,

$$\mathfrak{so}_n = T_1(SO(n)) = \{A \in M_n(\mathbb{R}) | A + A^T = 0\}.$$

Example 6.2.8. As an even simpler (but still important) example, consider $G = U(1) = S^1 \subseteq \mathbb{C}$. One has $x \in T_1U(1) \iff e^x \in S^1$, or $T_1(U(1)) = i\mathbb{R}$.

The adjoint representation. If $g \in G$ then consider the conjugation map $C_g \colon G \to G$ given by $h \mapsto ghg^{-1}$. Its derivative at the identity is denoted $\mathrm{Ad}(g)$:

$$\mathrm{Ad}(g) = (DC_g)\colon T_1G \to T_1G.$$

As an exercise, one should check that if $v \in \mathfrak{g}$ then $\mathrm{Ad}_{g_1}(\mathrm{Ad}_{g_2} v) = \mathrm{Ad}_{g_1 g_2}(v)$. So we have a representation of G on $\mathfrak{g}$, called the **adjoint representation**. For matrix groups $Ad(g)$ is simply conjugation by g, because if $A \in \mathfrak{g}$, then

$$\begin{aligned} Ad(g)(A) = \frac{d}{dt}\bigg|_{t=0} \big[C_g(e^{tA})\big] = \frac{d}{dt}\bigg|_{t=0} \big[g(e^{tA})g^{-1}\big] &= \frac{d}{dt}\bigg|_{t=0} [I + g(tA)g^{-1} + \cdots] \\ &= gAg^{-1}. \end{aligned}$$

In particular, all matrix conjugates of an element of $\mathfrak{g}$ are again in $\mathfrak{g}$. This useful fact will be used many times in the context of bundles and connections.

The Lie bracket. Since $Ad\colon G \to \mathrm{End}(\mathfrak{g})$ is smooth, we can take its derivative at the identity and get $D(Ad)\colon \mathfrak{g} \to T_I \,\mathrm{End}(\mathfrak{g}) = \mathrm{End}(\mathfrak{g})$. We usually write $[A, B]$ for $D(Ad)(A)(B)$. This can be shown to define a Lie bracket on $\mathfrak{g}$. In the case of matrix groups $[-, -]$ is just the commutator:

$$\begin{aligned} [A, B] = \frac{d}{dt}\bigg|_{t=0} \big[Ad(e^{tA})(B)\big] &= \frac{d}{dt}\bigg|_{t=0} \big[e^{tA}Be^{-tA}\big] \\ &= \frac{d}{dt}\bigg|_{t=0} \big[B + t(AB - BA) + \cdots\big] \\ &= AB - BA. \end{aligned}$$

6.2.9. Endomorphisms of G-bundles. Suppose $E \to B$ is a vector bundle where $E \cong V \times_G P$ for a (left) principal G-bundle $\pi : P \to B$ and a (right) G-representation $\rho : G \to GL(V)$. We will call such $E \to B$ a *G-bundle*.

For any such G-bundle, there exist open sets $U_\alpha \subseteq B$ and "G-trivializations" $\phi_\alpha : E|_{U_\alpha} \overset{\cong}{\to} V \times U_\alpha$ satisfying the following compatibility condition on $U_{\alpha\beta} = U_\alpha \cap U_\beta$:

$$V \times U_{\alpha\beta} \xrightarrow{\phi_\alpha^{-1}} E|_{U_{\alpha\beta}} \xrightarrow{\phi_\beta} V \times U_{\alpha\beta} \text{ is of the form } (v, x) \mapsto (v \cdot g_{\alpha\beta}(x), x)$$

where $g_{\alpha\beta} : U_{\alpha\beta} \to G$ is some smooth function.

Alternatively, we can choose trivializations for $P \to B$ and then take the product with the fiber V to get G-trivializations for $E \to B$.

Definition 6.2.10. *The G-endomorphisms of E is the image of the canonical map* $\mathrm{End}(P) \to \mathrm{End}(E)$.

That is, a G-endomorphism of E is an endomorphism that acts on each fiber as multiplication by some group element. More precisely, we want a collection of maps $\Phi_\alpha : U_\alpha \to \mathrm{End}(V)$ with the property that for each $x \in U_{\alpha\beta}$, there is a commutative diagram on fibers over x

Not sure whether this is a good defn, or even exactly what End(P) is

$$\begin{tikzcd}
V \arrow[r, "\Phi_\alpha(x)"] \arrow[d, "\cdot g_{\alpha\beta}(x)"'] & V \arrow[d, "\cdot g_{\alpha\beta}(x)"] \\
V \arrow[r, "\Phi_\beta(x)"] & V
\end{tikzcd}$$

Note that $\Phi_\alpha(x) \in Im(\rho)$ iff $\Phi_\beta(x) \in Im(\rho)$, that is, $\Phi_\alpha(x)$ corresponds to multiplication by an element of G iff $\Phi_\beta(x)$ does.

Moreover, $\mathrm{End}(P)$ is in one-to-one correspondence with $\mathrm{Map}(B, G)$ since $f : P \to P$ is an endomorphism (over B) iff $f(p) = p \cdot h(\pi(p))$ for a continuous map $h : B \to G$.

Example 6.2.11. Consider the trivial rank two vector bundle $I \times \mathbb{R}^2 = E$ over the unit interval $I = B$. If we equip this bundle with a metric, it has the structure of an $SO(2)$-bundle. The set of all endomorphisms $\mathrm{End}(E)$ is $C^\infty(B, M_2(\mathbb{R}))$, but the set of $SO(2)$-endomorphisms $\mathrm{End}_{SO(2)}(E)$ is the proper subset $C^\infty(B, SO(2))$.

We can also define $\mathfrak{g}$-endomorphisms for the Lie algebra $\mathfrak{g}$ associated to a Lie group G. These are endomorphisms of E such that, relative to a system of G-trivializations $\{U_\alpha\}$, the maps on each fiber act as elements of $\mathfrak{g}$. More explicitly, given a representation $G \xrightarrow{\rho} GL(V) \hookrightarrow \mathrm{End}(V)$ inducing $\mathfrak{g} \xrightarrow{D\rho} T_I GL(V) \hookrightarrow T_I \mathrm{End}(V) = \mathrm{End}(V)$, a $\mathfrak{g}$-endomorphism is a collection of maps $\Phi_\alpha : U_\alpha \to \mathrm{End}(V)$ such that for each $x \in U_{\alpha\beta}$ the following square of fibers commutes:

$$\begin{tikzcd}
V \arrow[r, "\Phi_\alpha(x)"] \arrow[d, "\cdot g_{\alpha\beta}(x)"'] & V \arrow[d, "\cdot g_{\alpha\beta}(x)"] \\
V \arrow[r, "\Phi_\beta(x)"] & V
\end{tikzcd}$$

As before, $\Phi_\alpha(x) \in Im(D\rho)$ iff $\Phi_\beta(x) \in Im(D\rho)$, since all G-conjugates of an element of $\mathfrak{g}$ are again in $\mathfrak{g}$.

6.2.12. Relation to connections. Let $E \to B$ be a smooth G-bundle and let $D_{(-)}(-)$ be a covariant derivative.

Definition 6.2.13. *Given a G-trivialization of E determined by linearly independent sections $e_1, \ldots, e_n$, D is a G-covariant derivative provided that for each vector field $\mathfrak{X}$ and each section $s = \sum_j s^j e_j$,*

$$D_{\mathfrak{X}}(s) = \sum_j (\partial_{\mathfrak{X}} s^j) e_j + A_{\mathfrak{X}}(s)$$

where A is a 1-form with values in $\underline{\mathrm{End}}(E)$ such that each $A_{\mathfrak{X}}$ is a $\mathfrak{g}$-endomorphism. (This happens to be independent of the choice of trivializing sections.)

Example 6.2.14. Consider $\mathbb{C} \times B \to B$ as a $U(1)$-bundle. (Recall that $T_I U(1) = i\mathbb{R}$.) Any $\mathfrak{u}(1)$-connection will be of the form $D_\mathfrak{X}(s) = [(\partial_\mathfrak{X} s_1) + i(\partial_\mathfrak{X} s_2)] + iT_\mathfrak{X}(s_1 + is_2)$ where $s = s_1 + is_2$ is a section with real-valued s_i and $T_\mathfrak{X}$ is a 1-form on B.

Example 6.2.15. Let $B = \mathbb{R}^4$ in the previous example. Since the magnetic potential $\mathfrak{f}$ is also a 1-form (because $d\mathfrak{f} = 0$), we can use it to define a $\mathfrak{u}(1)$-connection on this trivial bundle.

Example 6.2.16. Less trivially, a magnetic monopole generates a field with noncontractible base space $\mathbb{R}^3 - 0$. Although we don't have a global magnetic potential in this case, we do get a $\mathfrak{u}(1)$-connection.

6.2.17. Principal connections. Let G be a Lie group with associated Lie algebra $\mathfrak{g}$. Right translation by any $g \in G$ induces a map on tangent spaces $DR_g : T_I G \to T_g G$. For each $A \in \mathfrak{g}$, letting g vary over all elements in G defines a vector field $V_g(A) = DR_g(A)$ on G.

Similarly, for any principal G-bundle $P \to B$, the fiber is by definition homeomorphic to G, so for any $x \in P$, the right multiplication map $j_x : G \to P, g \mapsto g \cdot x$ induces a vector field $A \mapsto Dj_x(A) = V_x(A)$ on the fiber over x. Taken together over all fibers, we get a "vertical" vector field $A \mapsto V(A)$ on all of P.

Definition 6.2.18. *A principal connection on a principal G-bundle $P \to B$ assigns to each $x \in P$ and each $v \in T_x P$ an $\omega_x(v) \in \mathfrak{g}$ such that*

(1) $\omega_x(V_x(A)) = A$

(2) $\omega_{gx}(DL_g(v)) = Ad(g)(\omega_x(v))$, where L_g denotes left translation by $g \in G$.

We can think of this ω as an element of $\Omega^1(P; \mathfrak{g}) = \Gamma(T^*P \otimes \mathfrak{g})$.

6.3. Yang-Mills theory

Recall that we have now seen three equivalent versions of electromagnetism. In the first one studies two functions $\mathbf{E}, \mathbf{B} \colon \mathbb{R}^4 \to \mathbb{R}$ which satisfies the four Maxwell equations (for a fixed choice of ρ and $\mathbf{J}$). Sometimes one uses a vector potential $\mathbf{A}$ with the property that $\nabla \times \mathbf{A} = \mathbf{B}$, and an electric potential ϕ such that...

In the second version we start with a manifold M with a nondegenerate metric and a fixed 1-form $\mathfrak{J}$. One then studies 2-forms $\mathfrak{F}$ with the property that $d\mathfrak{F} = 0$ and $*d*\mathfrak{F} = \mathfrak{J}$. The former equation allows one to write $\mathfrak{F} = d\mathcal{A}$ for some 1-form $\mathcal{A}$ that is not uniquely specified.

Finally, in the third version of the theory one starts with a manifold M having a nondegenerate metric, together with a complex line bundle $E \to M$ equipped with a $\mathfrak{u}(1)$-connection D. Then the electromagnetic 2-form is the curvature associated to D,

$$\mathfrak{F} = R^D \in \Omega^2(M; i\mathbb{R}) \subseteq \Omega^2(M; \underline{\mathrm{End}}(E)).$$

What are the analogs of Maxwell's equations here? They are

$$d_D \mathfrak{F} = 0 \quad \text{(called the Bianchi identity)} \quad \text{and} \quad * d_D * \mathfrak{F} = \mathfrak{J}.$$

Here $\mathfrak{J}$ is regarded as an $\underline{\mathrm{End}}(E)$-valued 1-form, but taking its values in the subbundle $i\mathbb{R} \subseteq \underline{\mathrm{End}}(E)$.

...

Lemma 6.3.1. *Let $E \to B$ be a smooth vector bundle with connection D. Assume $\mu \in \Omega^p(B; \underline{\mathrm{End}}(E))$.*

(a) $\operatorname{tr}(d_D\mu) = d(\operatorname{tr}\mu)$ *as elements of* $\Omega^{p+1}(B)$.

(b) $\int_B \operatorname{tr}(\mu \wedge *\omega) = \int_B \operatorname{tr}(\omega \wedge *\mu)$ *for* $\omega \in \Omega^{B-p}(B; \underline{\operatorname{End}}(E))$.

(c) $\int_B \operatorname{tr}(d_D\mu \wedge \omega) = (-1)^{p+1}\int_B \operatorname{tr}(\mu \wedge d_D\omega)$ *for any* $\omega \in \Omega^{B-p-1}(B; \underline{\operatorname{End}}(E))$.

(d) *If* $D = D^0 + A$ *for a vector potential* A, *then* $d_D(\omega) = (d_{D^0}\omega) + A\wedge\omega - (-1)^p\omega\wedge A$.

(e) $\operatorname{tr}(\mu\wedge\omega) = (-1)^{pq}\operatorname{tr}(\omega\wedge\mu)$ *for any* $\omega \in \Omega^q(B; \underline{\operatorname{End}}(E))$.

For convenience we will assume that D^0 is flat, so that $\mathcal{F}_A = dA + A\wedge A$. Then we observe that

$$\begin{aligned}\mathcal{F}_{A+\delta A} &= d(A+\delta A) + (A+\delta A)\wedge(A+\delta A)\\ &= (dA + A\wedge A) + d(\delta A) + (\delta A)\wedge A + A\wedge(\delta A) + (\delta A\wedge\delta A)\\ &= \mathcal{F}_A + \big[(\delta A)\wedge A + A\wedge(\delta A)\big] + (\delta A\wedge\delta A)\end{aligned}$$

and therefore

$$\delta_A\mathcal{F} = d(\delta A) + (\delta A)\wedge A + A\wedge(\delta A) = d_D(\delta A)$$

where we have used Lemma 6.3.1(d) for the last equality. Now we compute

$$\begin{aligned}\delta_A S = \frac{1}{2}\delta_A\int_B \operatorname{tr}(\mathcal{F}\wedge *\mathcal{F}) &= \frac{1}{2}\int_B \operatorname{tr}(\delta_A\mathcal{F}\wedge *\mathcal{F}_A + \mathcal{F}_A\wedge *\delta_A\mathcal{F})\\ &= \int_B \operatorname{tr}(\delta_A\mathcal{F}\wedge *\mathcal{F}_A) \quad \text{(using Lemma 6.3.1(b))}\\ &= \int_B \operatorname{tr}(d_D(\delta A)\wedge *\mathcal{F}_A)\\ &= \int_B \operatorname{tr}(\delta A\wedge d_D(*\mathcal{F}_A)) \quad \text{(using Lemma 6.3.1(c))}.\end{aligned}$$

The only way this last expression will vanish for every choice of δA is if $d_D(*\mathcal{F}) = 0$. So we have that A is an extreme point of S if and only if $d_D(*\mathcal{F}_A) = 0$.

An important property of the Yang-Mills Lagrangian is that it is gauge invariant. Suppose that $E \to B$ is a G-bundle, and let $\mathscr{G}$ be the group of G-automorphisms. If D is an affine connection of E and $g \in \mathscr{G}$, then we get a new affine connection D^g via the formula

$$D^g_v(s) = g\cdot D_v(g^{-1}s).$$

This is the evident thing to do: apply g^{-1}, differentiate using D, then apply g to put us back to where we were.

If $D = D^0 + A$, let us write $D^g = D^0 + A^g$. Then $A \mapsto A^g$ gives a $\mathscr{G}$-action on the space $\Omega^1(B; \underline{\operatorname{End}}(E))$ of connection potentials. One readily verifies that

$$\begin{aligned}[\mathcal{F}_{D^g}]_{X,Y,s} &= D^g_X D^g_Y s - D^g_Y D^g_X s - D^g_{[X,Y]}(s)\\ &= (gD_Xg^{-1})(gD_Yg^{-1})s - (gD_Yg^{-1})(gD_Xg^{-1})s - gD_{[X,Y]}g^{-1}(s)\\ &= [g(\mathcal{F}_D)g^{-1}]_{X,Y,s}.\end{aligned}$$

In other words, $\mathcal{F}_{A^g} = g\mathcal{F}_A g^{-1}$. Therefore

$$\begin{aligned} L(A^g) = \frac{1}{2}\operatorname{tr}(g\mathcal{F}_A g^{-1} \wedge *g\mathcal{F}_A g^{-1}) &= \frac{1}{2}\operatorname{tr}(g(\mathcal{F}_A \wedge *\mathcal{F}_A)g^{-1}) \\ &= \frac{1}{2}\operatorname{tr}(\mathcal{F}_A \wedge *\mathcal{F}_A) = L(A). \end{aligned}$$

6.4. Digression on characteristic classes and Chern-Simons forms

In the previous section we set up an action functional on the space of connections on a given bundle $E \to B$. More precisely, we chose a fixed connetion D^0 and then identified the space of all connections with the contractible space $D^0 + \Omega^1(B; \underline{\mathrm{End}}(E))$. In this way we could define the action on the space of connection potentials $\Omega^1(B; \underline{\mathrm{End}}(E))$ rather than on the space of connections. But it amounts to the same thing.

As soon as one has the idea to do all this, it becomes apparent that there are other actions one might consider. Given a connection potential A, we get a curvature 2-form $\mathcal{F}_A$. How can we get a real number out of this? If we had a form whose dimension is $\dim B$ then we could integrate it, and we could make such a form by taking powers of $\mathcal{F}_A$—at least if $\dim B$ is even. So let's do this. Assume $\dim B = 2n$ and consider the action defined by

$$S(A) = \int_B \operatorname{tr}(\mathcal{F}_A \wedge \mathcal{F}_A \wedge \cdots \wedge \mathcal{F}_A) \qquad (n \text{ wedge products}).$$

Following the same steps as in the last section we compute that

$$\begin{aligned} \delta_A S &= \int_B \operatorname{tr}(\delta_A(\mathcal{F} \wedge \mathcal{F} \wedge \cdots \wedge \mathcal{F})) \\ &= \int_B \operatorname{tr}\Big(\delta_A\mathcal{F} \wedge \mathcal{F}^{\wedge(n-1)} + \mathcal{F} \wedge \delta_A\mathcal{F} \wedge \mathcal{F}^{\wedge(n-2)} + \cdots + \mathcal{F}^{\wedge(n-1)} \wedge \delta_A\mathcal{F}\Big) \\ &= n\int_B \operatorname{tr}(\delta_A\mathcal{F} \wedge \mathcal{F}^{\wedge(n-1)}) \qquad \text{(using the cyclic property of the trace)} \\ &= n\int_B \operatorname{tr}(d_D(\delta A) \wedge \mathcal{F}^{\wedge(n-1)}) \qquad (\text{since } \delta_A\mathcal{F} = d_D(\delta A)) \\ &= n\int_B \operatorname{tr}(\delta_A \wedge d_D(\mathcal{F}^{\wedge(n-1)})) \qquad \text{(by Lemma 6.3.1(d))} \\ &= 0. \end{aligned}$$

The last equality follows because $d_D(\mathcal{F}^{\wedge(n-1)}) = 0$, a consequence of the Leibniz rule and the Bianchi identity: for instance, when $n = 2$ one has

$$d_D(\mathcal{F} \wedge \mathcal{F}) = d_D(\mathcal{F}) \wedge \mathcal{F} + \mathcal{F} \wedge d_D(\mathcal{F}) = 0 + 0 = 0$$

and the analagous argument works for higher n.

Since we have proven that $\delta_A S = 0$ no matter what A is, it follows that S is constant on the space of connection potentials. In other words, S is an invariant of just the original bundle $E \to B$ (not depending on the connection). Of course for all we know right now it might be the trivial invariant, the same for all bundles—but we will soon see that this is not the case.

Looking carefully at the above argument, one sees right away that the integration over B was really not used at all. Or rather, it was only used so that $S(A)$ was an actual real number—but other than that it played no part in the argument. ???

To make the next part cleaner we will revert back to writing things in terms of the connection D rather than the connection potential A. Define $\theta_k(D) = \mathrm{tr}(\mathcal{F}_D \wedge \mathcal{F}_D \wedge \cdots \wedge \mathcal{F}_D)$ (k factors in the wedge). Then $\theta_k(D) \in \Omega^{2k}(B)$ and

$$d_D(\theta_k(D)) = d\Big(\mathrm{tr}(\mathcal{F}_D^{\wedge(k)})\Big) = \mathrm{tr}\Big(d_D\big(\mathcal{F}_D^{\wedge(k)}\big)\Big) = \mathrm{tr}(0) = 0.$$

So $\theta_k(D)$ is a deRham cocycle and therefore represents an element of $H^{2k}(B;\mathbb{R})$. Note that in the case $2k = \dim B$ then we could use integration over B to translate the top-dimensional form into a real number (assuming B is connected).

We claim that the cohomology class of $\theta_k(D)$ does not depend on the choice of D. This is an easy consequence of the naturality of our construction, together with the fact that the space of connections is affine. To see this, let D' be another connection and consider the bundle $E \times I \to B \times I$. We can concoct a connection on this new bundle by the formula

$$\nabla = tD + (1-t)D'.$$

That is to say, on the fiber over (b,t) we use the connection $tD + (1-t)D'$. (???) We therefore get a deRham cocycle $\theta_k(\nabla) \in C^{2k}(B \times I;\mathbb{R})$. If $i_0, i_1 \colon B \hookrightarrow B \times I$ denote the evident two inclusions, one readily checks that $i_0^*(\theta_k(\nabla)) = \theta_k(D)$ and $i_1^*(\theta_k(\nabla)) = \theta_k(D')$. Indeed,

$$i_0^*(\theta_k(\nabla)) = i_0^*\Big(\mathrm{tr}(\mathcal{F}_\nabla)^{\wedge(k)}\Big) = \mathrm{tr}\Big(\big(i_0^*(\mathcal{F}_\nabla)\big)^{\wedge(k)}\Big)$$

and so it is only a matter of checking that $i_0^*(\mathcal{F}_\nabla) = \mathcal{F}_D$. But this is clear enough from the definition of the curvature form.

Since i_0^* and i_1^* induce the same map on cohomology, we conclude that $\theta_k(D)$ and $\theta_k(D')$ represent the same cohomology class. This proves independence on the choice of connection.

Because $\theta_k(D)$, thought of as a cohomology class, does not depend on the choice of connection, it is therefore a topological invariant of the bundle $E \to B$. We will write $\theta_k(E)$ from now on. Of course it is still possible that this is the trivial invariant, but this is not the case:

Proposition 6.4.1. *Expanding coefficients from $H^{2k}(B;\mathbb{R})$ to $H^{2k}(B;\mathbb{C})$, we have*

$$\theta_k(E) = k! \cdot (-2\pi i)^k \cdot \mathrm{ch}_k(E)$$

where $\mathrm{ch}_k(E)$ is the kth term of the Chern character of E (lying in $H^{2k}(B;\mathbb{Q})$).

Remark 6.4.2. There are different choices for what one might mean by the "Chern character", differing from each other by a sign in odd dimensions. The key decision is whether the first Chern class of the tautological line bundle $L \to \mathbb{C}P^\infty$ is the standard generator of $H^2(\mathbb{C}P^\infty)$ or its negative. Topologists usually choose the former, whereas geometers choose the latter (becaue geometers prefer to have the first Chern class of the dual bundle L^*—called $\mathcal{O}(1)$ in their language—be the generator). Since our purpose in these notes is to understand work of the geometers, it seems sensible for us to adopt their conventions. If one were to use the topological conventions instead, the formula in Proposition 6.4.1 should have the minus sign removed from the $2\pi i$.

Proof of Proposition 6.4.1. The class $\mathrm{ch}_k(E)$ is completely characterized by naturality together with the following two properties:

(1) $\mathrm{ch}_k(E \oplus F) = \mathrm{ch}_k(E) + \mathrm{ch}_k(F)$, and
(2) $\mathrm{ch}_k(L) = \frac{1}{k!} c_1(L)^k$ if L is a line bundle.

It will therefore suffice to prove that $\theta_k(E \oplus F) = \theta_k(E) + \theta_k(F)$ and $\theta_k(L) = (-2\pi i)^k c_1(L)^k$ for line bundles L. The additivity statement is fairly easy, the key being that if A and B are square matrices and M is the block matrix

$$M = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix},$$

then $\mathrm{tr}(M^k) = \mathrm{tr}(A^k) + \mathrm{tr}(B^k)$. Note that if E and F are bundles with connections D_E and D_F, then $D_E \oplus D_F$ gives a connection on $E \oplus F$. The curvature form $R_{E\oplus F}$ is readily checked to be $R_E \oplus R_F$, by which we mean the 2-form whose values are the block sums of endomorphisms

$$\begin{bmatrix} R_E & 0 \\ 0 & R_F \end{bmatrix}.$$

At this point it is just a matter of chasing through the definitions to see that $\theta_k(E \oplus F) = \theta_k(E) + \theta_k(F)$.

For the calculation of $\theta_1(L)$ when L is a line bundle, it suffices by naturality to check this when L is the universal line bundle over $\mathbb{C}P^\infty$. But then applying naturality one more time, it suffices to check it when L is the tautological bundle over $\mathbb{C}P^1$. We use the connection on L induced by the standard embedding of L into the trivial rank 2 bundle (see Example A.1.12). Recall that the local connection matrices for this connection are

$$\Omega_U = \frac{\bar{z}\, dz}{z\bar{z} + 1} \qquad \text{and} \qquad \Omega_V = \frac{\bar{w}\, dw}{w\bar{w} + 1}.$$

See Example A.1.12 for notation; since these are 1×1 matrices we omit the matrix notation. The curvature form is given by

$$R_U = d\Omega_U - \Omega_U \wedge \Omega_U = d\Omega_U$$

and similarly for $R_V = d\Omega_V$. Let us now compute:

$$R_U = d\Omega_U = d\Big(\frac{\bar{z}}{z\bar{z}+1}\Big) \wedge dz = \left[\frac{d\bar{z}}{z\bar{z}+1} - \frac{\bar{z}}{(z\bar{z}+1)^2}(\bar{z}dz + zd\bar{z})\right] \wedge dz$$
$$= \frac{-1}{(z\bar{z}+1)^2}\, dz\, d\bar{z}.$$

The same computation shows that $R_V = \frac{-1}{(w\bar{w}+1)^2}\, dw\, d\bar{w}$, and if one wants one can readily check that these two forms agree on $U \cap V$ and hence patch together to give a global 2-form $R \in \Omega^2(\mathbb{C}P^1; \mathbb{C})$ (this is a nice exercise!)

The class $\theta_1(L)$ is the cohomology class in $H^2(\mathbb{C}P^1; \mathbb{C})$ represented by R (which is clearly a deRham cocycle). Therefore $\theta_1(L)$ is a complex multiple of $c_1(L)$, and we need to determine what the multiple is. To do this, recall that in the conventions of geometers $c_1(L)$ is the generator of $H^2(\mathbb{C}P^1; \mathbb{Z})$ satisfying $c_1(L) \cap [\mathbb{C}P^1] = -1$ (where the fundamental class of $\mathbb{C}P^1$ is the one determined by the complex orientation). See Remark 6.4.2 for more information about this. So let us compute:

$$\begin{aligned}
R \cap [\mathbb{C}P^1] = \int_{\mathbb{C}P^1} R = \int_U R_U &= \int_{\mathbb{C}} \frac{-1}{(z\bar{z}+1)^2}\, dz\, d\bar{z} \\
&= \int_{\mathbb{R}^2} \frac{-1}{(x^2+y^2+1)^2}(dx+idy)\wedge(dx-idy) \\
&= \int_{\mathbb{R}^2} \frac{2i}{(x^2+y^2+1)^2} dxdy \\
&= 2i \int_0^{2\pi}\int_0^{\infty} \frac{1}{(r^2+1)^2} r\, dr\, d\theta \\
&= 4\pi i \int_0^{\infty} \frac{1}{(r^2+1)^2} r\, dr \\
&= 4\pi i \cdot -\frac{1}{2}\cdot \frac{1}{r^2+1}\bigg|_0^{\infty} = 2\pi i.
\end{aligned}$$

We conclude that $\theta_1(L) = R = -2\pi i \cdot c_1(L)$ as elements of $H^2(\mathbb{C}P^1;\mathbb{Z})$.

We have now proven that $\theta_1(\mathcal{L}) = -2\pi i c_1(\mathcal{L})$ for any line bundle $\mathcal{L} \to B$. To complete the proof let us consider $\theta_k(\mathcal{L})$. Note that if V is a one-dimensional vector space then $\mathrm{Hom}(V,V)$ is *canonically* isomorphic to $\mathbb{C}$. It follows that for any line bundle $\mathcal{L}$, the bundle $\mathrm{End}(\mathcal{L})$ is trivial. The curvature form of $\mathcal{L}$ lies in $\Omega^2(B;\underline{\mathrm{End}}(\mathcal{L}))$, but we may identify this with $\Omega^2(B;\mathbb{C})$. The wedge product $R\wedge\cdots\wedge R$ of $\underline{\mathrm{End}}(\mathcal{L})$-valued forms becomes just the ordinary wedge product of $\mathbb{C}$-valued forms, and taking "trace" does nothing at all (i.e., the trace of a 1×1 matrix is just the matrix entry). We conclude that

$$\theta_k(\mathcal{L}) = \mathrm{tr}(R_{\mathcal{L}}\wedge\cdots\wedge R_{\mathcal{L}}) = (R_{\mathcal{L}})^k = \theta_1(\mathcal{L})^k = (-2\pi i)^k \cdot c_1(\mathcal{L})^k.$$

This completes the proof. □

Corollary 6.4.3. *For any k, the class $\frac{1}{(-2\pi i)^k}\cdot\theta_k(E)$ lies in the image of $H^{2k}(B;\mathbb{Z})$ in $H^{2k}(B;\mathbb{C})$.*

Proof. This follows from the basic topological fact that for any bundle E the class $k!\cdot\mathrm{ch}_k(E)$ lies in $H^{2k}(B;\mathbb{Z})$. The proof of this is as follows. Formally factor the total Chern class $c(E)$ as $c(E) = (1+x_1)(1+x_2)\cdots(1+x_n)$ where n is the rank of E. Then the kth Chern class of E is the kth elementary symmetric function in the x_i's. The Chern character of E is

$$\mathrm{ch}(E) = \sum_i e^{x_i} = \sum_i \Big(1 + x_i + \frac{x_i^2}{2} + \frac{x_i^3}{3!} + \cdots\Big).$$

Therefore $\mathrm{ch}_k(E) = \frac{1}{k!}\cdot(x_1^k+\cdots+x_n^k)$. The power sum $x_1^k+\cdots+x_n^k$ is a symmetric polynomial in the x_i's, and it therefore can be written as an integral-coefficient polynomial in the elementary symmetric functions. In other words, $k!\cdot\mathrm{ch}_k(E)$ is an integral-coefficient polynomial in the Chern classes of E. Since the Chern classes lie in $H^*(B;\mathbb{Z})$, so does $k!\cdot\mathrm{ch}_k(E)$. □

6.4.4. The Chern-Weil approach to characteristic classes. The following is somewhat of an aside, but it seems worthwhile to explain how the above results on characteristic classes fit in with the classical Chern-Weil approach.

Let $X = (x_{ij})$ be an $n \times n$ matrix of indeterminates over a field k of characteristic zero. A polynomial $P \in k[x_{ij}]$ is called *invariant* if $P(QXQ^{-1}) = P(X)$ for every invertible matrix Q with entries in k. Two evident examples are $P(X) = \mathrm{tr}(X)$ and $P(X) = \det(X)$. Some less obvious examples are the sequences

$$\mathrm{tr}(X),\ \mathrm{tr}(X^2),\ \mathrm{tr}(X^3),\ \ldots \qquad \text{and} \qquad \mathrm{tr}(X),\ \mathrm{tr}(\wedge^2 X),\ \mathrm{tr}(\wedge^3 X),\ \ldots$$

Let $\mathrm{InvP}[n]$ denote the subring of $k[x_{ij}]$ consisting of the invariant polynomials.

Let $P \in \mathrm{InvP}_{\mathbb{C}}[n]$, and for convenience assume that P is homogeneous of degree k. The polynomial P gives a characteristic class for rank n complex bundles in the following way. For a bundle $E \to B$, choose a connection D and let Ω_U be the local connection matrices over various open sets U covering B. One has the resulting curvature matrices $R_U = (d\Omega_U - \Omega_U \wedge \Omega_U)^T$, and these are compatible in the sense that $R_V = \alpha R_U \alpha^{-1}$ on $U \cap V$, where α is the transition function for E. If we apply P to the previous formula then we get $P(R_V) = P(R_U)$, and hence these forms patch together to give a global form $P(R) \in \Omega^{2k}(B; \mathbb{C})$. With a little work one can see that these are deRham cocycles, and that the cohomology class $[P(R)]$ does not depend on the choice of connection or local trivialization. In this way we obtain characteristic classes. (Note that it is important that the matrix entries of R_U are 2-forms, and hence commute with each other; this guarantees that $P(R_U)$ has an unambiguous interpretation).

Once one stumbles on this idea, the obvious next step is to determine the complete ring of invariant polynomials. It turns out this is not hard. If $\lambda_1, \ldots, \lambda_n$ denote the eigenvalues of X in some algebraically closed field extension of $k(x_{ij})$, every symmetric function in the λ_i's is an invariant polynomial (because conjugating X will not change the eigenvalues). That is, we have a map

$$\phi\colon k[\lambda_1, \ldots, \lambda_n]^{\Sigma_n} \to \mathrm{InvP}[n].$$

Note that this map sends the power sum $\lambda_1^r + \cdots + \lambda_n^r$ to $\mathrm{tr}(X^r)$, whereas it sends the rth elementary symmetric function $\sigma_r(\lambda_1, \ldots, \lambda_n)$ to $\mathrm{tr}(\wedge^r X)$.

Proposition 6.4.5. *The map ϕ is an isomorphism, and so* $\mathrm{InvP}[n]$ *is a polynomial ring over k in n variables. One has*

$$\mathrm{InvP}[n] = k[\mathrm{tr}(X), \mathrm{tr}(X^2), \ldots, \mathrm{tr}(X^n)] = k[\mathrm{tr}(X), \mathrm{tr}(\wedge^2 X), \ldots, \mathrm{tr}(\wedge^n X)].$$

Note that the generators $\mathrm{tr}(\wedge^r X)$ are in some sense more fundamental then the generators $\mathrm{tr}(X^r)$, because of the connection with the ring of invariants $k[\lambda_1, \ldots, \lambda_n]^{\Sigma_n}$. When one works over $\mathbb{Z}$ rather than a field, this ring of invariants is generated by the elementary symmetric functions but is *not* generated by the power sums: that is, the power sums are integral polynomial expressions in the elementary symmetric functions, but the elementary symmetric functions are only *rational* polynomial expressions in the power sums.

The rth Chern class $c_r(E)$ of a bundle $E \to B$ is defined to be the characteristic class associated to the invariant polynomial

$$P_r(X) = \frac{1}{(-2\pi i)^r}\,\mathrm{tr}(\wedge^r X) = \frac{1}{(-2\pi i)^r} \cdot \sigma_r(\lambda_1, \ldots, \lambda_n).$$

The normalization constant is there to make this definition coincide with the topological Chern classes which lie in integral cohomology. Let us write $s_r(E)$ for the characteristic class associated to the invariant polynomial

$$S_r(X) = \frac{1}{(-2\pi i)^r}\operatorname{tr}(X^r) = \frac{1}{(-2\pi i)^r}\cdot(\lambda_1^r + \cdots + \lambda_n^r).$$

It is easy to determine the relationship between the s_r classes and the Chern classes, by writing the power sums as polynomial expressions in the elementary symmetric functions. For instance,

$$\lambda_1^2 + \cdots + \lambda_n^2 = \sigma_1^2 - 2\sigma_2$$

and therefore

$$s_2(E) = c_1(E)^2 - 2c_2(E).$$

Note that we have $s_1(E) = c_1(E)$. As an exercise, the reader may work out that $s_3(E) = c_1(E)^3 - 3c_2(E)c_1(E) + 3c_3(E)$.

The class $\theta_r(E)$ defined in the last section is the characteristic class associated to the invariant polynomial $\operatorname{tr}(X^r)$. That is to say, $s_r(E) = \frac{1}{(-2\pi i)^r}\theta_r(E)$. This follows directly from Proposition 6.4.1.

6.4.6. Bundles with flat connection and Chern-Simons forms. Now suppose that our bundle $E \to B$ admits a flat connection D. The curvature form of such a connection is identically zero (by definition), and therefore $\theta_k(D) = 0$. It then follows that $\theta_k(D+A)$ represents the zero cohomology class, for any connection potential A. So $\theta_k(D+A)$ is the coboundary of some $(2k-1)$-form on B. With a little work it turns out that one can write down an *explicit* $(2k-1)$-form that does the job. These are the so-called **Chern-Simons forms**, which can be thought of as secondary characteristic classes for flat bundles. We will derive these next.

Consider the 1-parameter family of connection potentials $A_s = sA$ for $s \in [0,1]$. The $\mathcal{F}_{A_s} = dA_s + A_s \wedge A_s = s(dA) + s^2(A \wedge A)$. We observe that

$$\begin{aligned}
\theta_k(A) = \theta_k(A) - \theta_k(0) &= \int_0^1 \frac{d}{ds}\operatorname{tr}(\mathcal{F}_s \wedge \cdots \mathcal{F}_s)\,ds \\
&= k\int_0^1 \operatorname{tr}\Big(\frac{d\mathcal{F}_s}{ds} \wedge \mathcal{F}_s^{\wedge(k-1)}\Big)\,ds \quad \text{(cyclic property)} \\
&= k\int_0^1 \operatorname{tr}\Big((dA + 2s(A\wedge A)) \wedge (sdA + s^2 A \wedge A)^{\wedge(k-1)}\Big)ds \\
&= k\int_0^1 s^{k-1}\operatorname{tr}\Big((dA + 2sA\wedge A) \wedge (dA + sA \wedge A)^{\wedge(k-1)}\Big)ds.
\end{aligned}$$

Before attempting the general case let us look carefully at $k = 2$. Inside the integral we are looking at the trace of

$$(dA \wedge dA) + 2s(A \wedge A \wedge dA) + s(dA \wedge A \wedge A) + 2s^3(A \wedge A \wedge A \wedge A).$$

The last of these terms has zero trace, because A is a 1-form and therefore $\operatorname{tr}((A \wedge A \wedge A) \wedge A) = -\operatorname{tr}(A \wedge (A \wedge A \wedge A))$. In general, any even wedge power of A will be traceless. The middle two terms can be combined using the cyclic property of the trace, and so we are looking at the trace of

$$(dA \wedge dA) + 3s(dA \wedge A \wedge A).$$

The cyclic property also implies that when k is even one has

$$\operatorname{tr}(d(A^{\wedge(k)})) = k \cdot \operatorname{tr}(dA \wedge A^{\wedge(k-1)}).$$

We leave this as an exercise. Using this, we see that the trace inside our integral coincides with

$$\operatorname{tr}(d((A\wedge dA)+s(A\wedge A\wedge A)))=d(\operatorname{tr}(A\wedge dA+sA\wedge A\wedge A)).$$

So finally we have that

$$\begin{aligned}\theta_2(A)&=2\cdot\int_0^1 s\cdot d(\operatorname{tr}(A\wedge dA+sA\wedge A\wedge A))\\&=d\biggl(\operatorname{tr}\biggl(\int_0^1 2s(A\wedge dA)+2s^2(A\wedge A\wedge A)\,ds\biggr)\biggr)\\&=d\biggl(\operatorname{tr}\biggl(s^2(A\wedge dA)+\frac{2}{3}s^3(A\wedge A\wedge A)\biggr]_{s=0}^{s=1}\biggr)\biggr)\\&=d\biggl(\operatorname{tr}\biggl(A\wedge dA+\frac{2}{3}A\wedge A\wedge A\biggr)\biggr).\end{aligned}$$

We will write

$$CS_3(A)=A\wedge dA+\tfrac{2}{3}A\wedge A\wedge A\in\Omega^3(B;\underline{\operatorname{End}}(E))$$

and likewise

$$cs_3(A)=\operatorname{tr}(CS_3(A))\in\Omega^3(B).$$

These are called Chern-Simons 3-forms. Note that $cs_3(A)$ is almost certainly *not* a deRham cocycle; its coboundary is $\theta_2(A)$, which need not be zero.

The above procedure can be carried out for any k, although the bookkeeping becomes qute messy. The end result is the following. Define

$$cs_k(A)=\operatorname{tr}\Bigl(\int_0^1 A\wedge(\mathcal{F}_s)^{\wedge(k-1)}\,ds\Bigr)\in\Omega^{2k-1}(B).$$

PROPOSITION 6.4.7. *$\theta_k(A)=d(cs_k(A))$ for any connection potential A.*

PROOF. This is an unpleasant computation that we leave to the reader. (In these notes we will only ever use the case $k=2$). □

Part 3

Quantum field theory and topology

CHAPTER 7

Quantum field theory and the Jones polynomial

7.1. A first look at Chern-Simons theory

Let G be a compact Lie group, let B be a 3-manifold, and let $E \to B$ be a trivial G-bundle. Let D be the standard flat connection on the bundle. Let $\mathscr{G}$ be the group of gauge transformations, which for a trivial bundle is just $\mathrm{Map}(B, G)$. Let $\mathscr{G}_0 \subseteq \mathscr{G}$ be the connected component of the identity.

For a vector potential A we can consider the Chern-Simons 3-form $CS_3(A) = A \wedge dA + \frac{2}{3} A \wedge A \wedge A$. The trace of this 3-form lies in $\Omega^3(B)$, and so we can integrate it to get a real number. Define

$$L(A) = \mathrm{tr}(A \wedge dA + \tfrac{2}{3} A \wedge A \wedge A) \in \Omega^3(B)$$

and

$$S(A) = \int_B L(A) \in \mathbb{R}.$$

It turns out that this "action" functional is not invariant under gauge transformations. But it *almost* is; we will prove below that if $g \in \mathscr{G}$ then $S(A^g) = S(A) + 8\pi^2 n$ for some $n \in \mathbb{Z}$ (the $8\pi^2$ comes from the integrality properties of the second Chern class). Because of this, the expression $e^{iS(A)/4\pi}$ *is* invariant under gauge transformations.

Of course $S(A)$ is not a topological invariant of the situation, it depends very much on the choice of vector potential A. Witten had the idea of producing topological invariants by taking all vector potentials at once and averaging them together via a Feynman-type integral. This involves looking at integrals of the form

$$\int_{\mathcal{A}} e^{\frac{ikS(A)}{4\pi}} DA \qquad \text{and} \qquad \int_{\mathcal{A}} e^{\frac{ikS(A)}{4\pi}} \boxed{????} DA.$$

Here $\mathcal{A} = \Omega^1(B; \mathfrak{g})$ is the space of all vector potentials, and $k \in \mathbb{Z}$ is a fixed integer called the **level**. This integer k plays a role similar to $\frac{1}{\hbar}$ in the quantum theory. In particular, as k gets large then the integral becomes wildly oscillatory and is dominated by the critical points of S. The mystery box in the second integral denotes some "observable" that we might be interested in.

At the moment, none of the above makes much sense. And even when we're done it still won't make much sense, because the integrals will not be rigorously defined. But eventually we will at least have an idea of what the basic issues are, and what the above integrals are trying to capture.

7.1.1. Behavior under gauge transformations.

Proposition 7.1.2. *If $g \in \mathscr{G}_0$ then $S(A^g) = S(A)$.*

PROOF. Let $s \mapsto g_s$ denote a smooth path $I \to \mathscr{G}$ such that $g_0 = 1$ and $g_1 = g$. Define $A_s = A^{g_s} = g_s A g_s^{-1} + g_s d(g_s^{-1})$. We will prove that $\frac{d}{ds}\Big|_{s=0} S(A_s) = 0$. ????

Set $T = \frac{d}{ds}\Big|_{s=0} g_s$. Differentiating the equation $g_s \cdot g_s^{-1} = 1$ gives

$$T \cdot 1 + 1 \cdot \frac{d}{ds}\Big|_{s=0} g_s^{-1} = 0,$$

so $\frac{d}{ds}\Big|_{s=0} g_s^{-1} = -T$. We conclude that

$$\frac{d}{ds}\Big|_{s=0} A_s = TA - AT + \cancel{T \cdot d(\mathrm{id})} + \mathrm{id} \cdot d(-T)$$

□

PROPOSITION 7.1.3. *For any $g \in \mathscr{G}$, $S(A^g) - S(A) \in 8\pi^2\mathbb{Z}$.*

PROOF. Set $A_0 = A$, $A_1 = A^g$, and $A_s = sA^g + (1-s)A$. Then A_s defines a vector potential on the bundle $E \times I \to B \times I$ in the evident way; that is, we can regard A_s as defining an element of $\Omega^1(B \times I; \underline{\mathrm{End}}(E \times I))$.

Now let us make a bundle $\tilde{E} \to S^1 \times B$ by gluing $E \times \{0\}$ to $E \times \{1\}$ via g. Our vector potential A_s induces a corresponding potential $\tilde{A} \in \Omega^1(B \times S^1; \underline{\mathrm{End}}(\tilde{E})$. Let $\tilde{F}$ be the curvature form corresponding to $\tilde{A}$. Then we have

$$\mathrm{tr}(\tilde{F} \wedge \tilde{F}) = \theta_2(\tilde{A}) = 2! \cdot (2\pi i)^2 c_2(\tilde{E}) = -8\pi^2 c_2(\tilde{E}).$$

But $c_2(\tilde{E}) \in H^4(B \times S^1; \mathbb{Z})$, and so $\int_{B \times S^1} c_2(\tilde{E}) \in \mathbb{Z}$. We conclude that

$$\int_{B \times S^1} \mathrm{tr}(\tilde{F} \wedge \tilde{F}) \in 8\pi^2\mathbb{Z}.$$

But

$$\begin{aligned}\int_{B \times S^1} \mathrm{tr}(\tilde{F} \wedge \tilde{F}) = \int_{B \times I} \mathrm{tr}(\tilde{F} \wedge \tilde{F}) &= \int_{B \times I} d(cs_3(AI)) \\ &= \int_B cs_3(A_1) - \int_B cs_3(A_0) \\ &= S(A_1) - S(A_0) \\ &= S(A^g) - S(A).\end{aligned}$$

□

7.1.4. Wilson loops and observables. Let $\gamma \colon I \to B$ be a closed path, and let $x = \gamma(0)$. The affine connection $D + A$ gives a system of parallel transport on our bundle, which gives us a linear map $P_\gamma^A \colon E_x \to E_x$. Now let y be another point on γ, and imagine going around the same loop but based at y. Let γ' denote this new path. So if α denotes the portion of the loop from x to y we can write $\gamma\alpha = \alpha\gamma'$. So

$$P_\gamma^A \circ P_\alpha^A = P_\alpha^A \circ P_{\gamma'}^A$$

or

$$P_{\gamma'}^A = (P_\alpha^A)^{-1} \circ P_\gamma^A \circ P_\alpha^A.$$

In particular, we conclude from this that $\mathrm{tr}(P_\gamma^A) = \mathrm{tr}(P_{\gamma'}^A)$. So this trace only depends on the loop, not on a starting point. We will write $W(\gamma, A)$ for this trace. Physicists call it a "Wilson loop".

If our bundle is a G-bundle and both D and A are G-connections, then P_γ^A will be multiplication by an element g of G that is well-defined up to conjugacy. If R is a finite-dimensional G-representation then we can look at

$$W_R(\gamma, A) = \operatorname{tr}_R(P_\gamma^A),$$

by which we mean the trace of g acting on R. This does not depend on the choice of g in its conjugacy class, or on a starting point for the loop γ.

Now fix a collection of loops $\gamma_1, \ldots, \gamma_k$ inside of B, together with chosen G-representations $R_1, \ldots, R_k$. One should imagine the γ_i's as a collection of knots in B, possibly linked together in complicated ways. From this data Witten proposed to form a physicists' expectation value:

$$\langle W_{R_1}(\gamma_1) \cdots W_{R_k}(\gamma_k) \rangle = \frac{\int_{\mathcal{A}} W_{R_1}(\gamma_1, A) \cdots W_{R_k}(\gamma_k, A) e^{\frac{ikS(A)}{4\pi}}\, DA}{\int_{\mathcal{A}} e^{\frac{ikS(A)}{4\pi}}\, DA}.$$

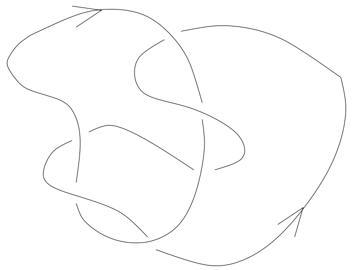
Provided one can make sense of these Feynman integrals, one obtains invariants for collections of knots inside of B.

There are two special cases that we will focus on in the upcoming sections. The simplest case is when $G = U(1)$ and all the R_i's are the standard representation of $U(1)$ on $\mathbb{C}$. When A is a $U(1)$-connection the Chern-Simons form simplifies, because the $A \wedge A \wedge A$ term vanishes (because $U(1)$ is abelian, and therefore $A \wedge A$ vanishes). In this case the Feynman integrals, once suitably interpreted, recover the classical linking numbers of the knots. We will review these invariants in the next section.

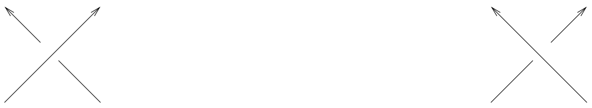
The next interesting case is $G = SU(2)$, where we take all the R_i's to be the standard representation of $SU(2)$ on $\mathbb{C}^2$. In this case the knot invariants are supposed to recover some form of the Jones polynomial.

7.2. Linking numbers and the writhe

Suppose we are given two oriented knots in $\mathbb{R}^3$. Pick a generic plane and project the knots onto the plane to form a "knot diagram" such as the one shown below:



Let us divide the crossings in the knot diagram into two types, called positive and negative, according to the scheme shown here:



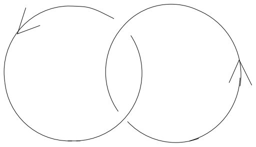
If we look at a knot crossing in such a way that the orientations of both strands move from our chins to our foreheads, then a positive crossing is one in which the top strand crosses from left to right; a negative crossing is one where the top strand crosses from right to left. The knot diagram above has six crossings, three positive and three negative.

Define the **linking number** of the two oriented knots by the formula

$$L(\alpha,\beta)=\frac{\#(\text{positive crossings})-\#(\text{negative crossings})}{2}.$$

The number of positive crossings plus the number of negative crossings is the total number of crossings, which must be even. From this it is clear that the linking number is always an integer. It is not too hard to show that this is a topological invariant of the linked knots.

Example 7.2.1. Given a pair of linked rings as shown below



the linking number is -1. For the knots depicted at the beginning of this section, the linking number is zero (which demonstrates that the linking number is not an absolute invariant for when two knots can be separated or not).

There are several alternative descriptions of the linking number which are also useful. Pick an oriented surface D whose boundary is $\partial D=\alpha$ and which intersects β transversely. Every point of intersection can be given a sign ± 1 determined in an evident way by how the orientations match up. The the linking number is the total number of points in $D\cap\beta$, where the points are counted with sign.

The linking number was first introduced by Gauss, who came to it through his study of electromagnetism. Imagine that the knot β is a wire with a steady electric current of unit strength flowing through it. This current generates a magnetic field, and so one can consider the magnetic circulation around α: that is, consider the line integral $\int_\alpha \mathbf{B}\cdot\mathbf{ds}$. We claim that (up to units) this is just the linking number. In fact this is a direct consequence of Maxwell's equations. We are in a static situation where both $\mathbf{E}$ and $\mathbf{B}$ are not changing with time, so

$$\int_\alpha \mathbf{B}\cdot\mathbf{ds}=\iint\limits_D \mathbf{J}\cdot\hat{\mathbf{n}}\,dS$$

where we are leaving out physical constants and where D is any disk whose boundary is α. Since the current has unit strength, it is clear that the double integral on the right is just counting the number of times the wire crosses the surface of D, with different crossing directions counted as ± 1. This is the linking number.

Homological description of the linking number. From the standpoint of algebraic topology the easiest description is as follows. If K is any knot in S^3, then the homology gropu $H_1(S^3-K)$ is isomorphic to $\mathbb{Z}$, with a generator given by a small loop that circles around the "string" of the knot:

????

This computation is usually done in basic algebraic topology texts as a simple Mayer-Vietoris argument. Note that an orientation of the knot K can be used to fix a generator of $H_1(S^3-K)$, e.g. by the right-hand-rule where the thumb points along the oriented knot and the fingers curl around the knot specifying the generator.

If L is another oriented knot in S^3 that is disjoint from K, then L specifies a 1-cycle in $H_1(S^3-K)$. It is therefore an integral multiple of our chosen generator, and this integer is clearly the linking number: essentially one is breaking L up into "small loops around K" and "other stuff that doesn't matter", and adding up the small loops with their correct signs.

Integral description of the linking number. One can also obtain the linking number as the homological degree of a certain map. Define the so-called "Gauss map" $\Phi(\alpha,\beta)\colon S^1\times S^1\to S^2$ by the formula

$$(s,t)\mapsto \frac{\alpha(s)-\beta(t)}{|\alpha(s)-\beta(t)|}.$$

The degree can be computed by choosing a generic point $p\in S^2$ and counting the points in the preimage of p (where each point has a ±1 multiplicity). The point p is a unit vector in $\mathbb{R}^2$, so consider the normal plane to p and project the knots onto this plane. Points in the preimage of p correspond to crossings in the knot diagram where the α strand is on top. One checks that the sign of the crossing coresponds to whether $\Phi(\alpha,\beta)$ is orientation-preserving or reversing near the corresponding point of its preimage.

There are a number of ways of writing an integral formula for the linking number. The easiest comes directly from the description as the degree of $\Phi(\alpha,\beta)$. If we temporarily let d denote this degree, then we can write

$$\begin{aligned}4\pi=\Phi(\alpha,\beta)^*(\mathrm{vol}_{S^2}\cap[S^2]) &= \Phi(\alpha,\beta)^*\big(\mathrm{vol}_{S^2}\cap\tfrac{1}{d}\Phi(\alpha,\beta)_*([S^1\times S^1])\big)\\ &=\tfrac{1}{d}\cdot\Phi(\alpha,\beta)^*(\mathrm{vol}_{S^2})\cap[S^1\times S^1]\\ &=\tfrac{1}{d}\cdot\int_{S^1\times S^1}\Phi(\alpha,\beta)^*(\mathrm{vol}_{S^2}).\end{aligned}$$

Since d is the linking number, we obtain the integral formula

$$L(\alpha,\beta)=\frac{1}{4\pi}\int_{S^1\times S^1}\Phi(\alpha,\beta)^*(\mathrm{vol}).$$

If $\mathbf{r_1}$ and $\mathbf{r_2}$ denote parameterizations of the two knots (we called these α and β above), we claim the above integral can also be written as

$$\begin{aligned}L(K,L)&=\frac{1}{4\pi}\iint\frac{1}{|\mathbf{r_1}-\mathbf{r_2}|^3}\,(\mathbf{r_1}-\mathbf{r_2})\cdot(d\mathbf{r_1}\times d\mathbf{r_2})\\ &=\frac{1}{4\pi}\iint_{s,t}\frac{1}{|\mathbf{r_1}(s)-\mathbf{r_2}(t)|^3}\,\big[\mathbf{r_1}(s)-\mathbf{r_2}(t)\big]\cdot\Big(\frac{d\mathbf{r_1}}{ds}\times\frac{d\mathbf{r_2}}{dt}\Big)\end{aligned}\tag{7.2.2}$$

(the first integral is really just shorthand for the second). To verify this claim we have to compute $\Phi(\alpha,\beta)^*(\mathrm{vol}_{S^2})$. It's easier if we factor $\Phi(\alpha,\beta)$ into two maps

$$S^1\times S^1\xrightarrow{g}\mathbb{R}^3-0\xrightarrow{h}S^2$$

where $g(s,t)=\mathbf{r_1}(s)-\mathbf{r_2}(t)$ and $h(v)=\frac{v}{|v|}$. The messy part is computing $h^*(\mathrm{vol})$.

If $A\colon I \to \mathbb{R}^3 - 0$ is a smooth path, an easy computation shows that

$$\frac{d}{ds}\left(\frac{A(s)}{|A(s)|}\right) = \frac{1}{|A(s)|^3}\Big[\big(A(s)\cdot A(s)\big)A'(s) - \big(A(s)\cdot A'(s)\big)A(s)\Big].$$

If we write $A(0) = p$ and $A'(0) = u$, then we have computed

$$(Dh)_p(u) = \frac{d}{ds}\bigg|_{s=0}\left(\frac{A(s)}{|A(s)|}\right) = \frac{1}{|p|^3}\Big[(p\cdot p)u - (p\cdot u)p\Big] = \frac{1}{|p|}\left[u - \Big(\frac{p}{|p|}\cdot u\Big)\frac{p}{|p|}\right].$$

Notice that up to scalar multiple this is the projection of u onto the plane perpendicular to p.

Regarding S^2 as embedded in $\mathbb{R}^3$ in the usual way, the volume form on S^2 is defined so that for two tangent vectors u and v at a point p, $\mathrm{vol}_p(u, v)$ is almost just the norm of the cross product $u \times v$. The "almost" comes from the fact that one needs to add signs in places, since after all vol is supposed to be an alternating tensor. The slick way to do this is via the formula

$$\mathrm{vol}_p(u, v) = p\cdot(u\times v).$$

It follows that if $p \in \mathbb{R}^3 - 0$ and $u, v \in \mathbb{R}^3$ are tangent vectors at p, then

$$\begin{aligned} h^*(\mathrm{vol})_p(u,v) &= \frac{p}{|p|}\cdot\Big[(Dh)_p(u)\times(Dh)_p(v)\Big] \\ &= \frac{p}{|p|^3}\cdot\left[\left[u - \Big(\frac{p}{|p|}\cdot u\Big)\frac{p}{|p|}\right]\times\left[v - \Big(\frac{p}{|p|}\cdot v\Big)\frac{p}{|p|}\right]\right] \\ &= \frac{1}{|p|^3}\Big[p\cdot(u\times v)\Big]. \end{aligned}$$

The derivative of g is much easier to compute, and one has that at any point the unit tangent vectors of the two S^1's map to $\frac{d\mathbf{r_1}}{ds}$ and $\frac{d\mathbf{r_2}}{dt}$. We therefore have computed that, at any point of $S^1 \times S^1$, $\Phi(\alpha,\beta)^*(\mathrm{vol})$ evaluated on the two evident unit tangent vectors gives the number

$$\frac{1}{|\mathbf{r_1}(s) - \mathbf{r_2}(t)|^3}\big[\mathbf{r_1}(s) - \mathbf{r_2}(t)\big]\cdot\Big(\frac{d\mathbf{r_1}}{ds}\times\frac{d\mathbf{r_2}}{dt}\Big).$$

This completes our derivation of the integral formula (7.2.2).

Physicists like to write the integral from (7.2.2) in coordinates, and it's worth doing this here so that it doesn't seem as scary when coming across it in a physics paper. For this we need the Levi-Civita ε symbol, which is used for writing out the coordinates of a cross product. The Levi-Civita symbol consists of numbers $\varepsilon_{ijk} \in \{1, -1\}$, and is defined by $\varepsilon_{123} = 1$ and the criterion that ε be antisymmetric. The reader may check that using this symbol one can write

$$(u\times v)_i = \sum \varepsilon_{ijk}u_jv_k.$$

As another simple example, if A is a 3×3 matrix then $\det A = \varepsilon_{ijk}a_{1i}a_{2j}a_{3k}$.

If we now write $\mathbf{r_1}(s) = x(s) = (x_1(s), x_2(s), x_3(s))$ and $\mathbf{r_2}(t) = y(t) = (y_1(t), y_2(t), y_3(t))$, formula (7.2.2) becomes

$$L(K, L) = \frac{1}{4\pi}\iint\frac{1}{|x-y|^3}(x_i - y_i)\epsilon_{ijk}dx_jdy_k.$$

Linking numbers for links. Suppose now that K and L are oriented links in $\mathbb{R}^3$ that do not intersect (where a *link* is like a knot but can have many components). We define the linking number of K and L by

$$L(K,L)=\sum_{i,j}L(K_i,L_j)$$

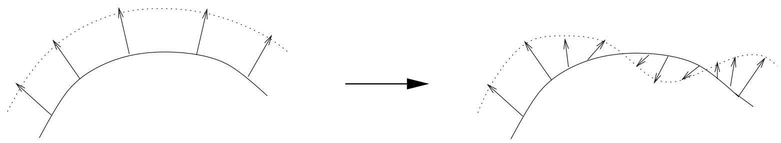
where the K_i's and L_j's are the different components of K and L.

Suppose that K and L are knots. Let $\tilde{K}$ be a link consisting of m copies of K that are slightly perturbed to be disjoint from each other. Likewise, let $\tilde{L}$ be a link consisting of r slightly perturbed copies of L. Note that $L(\tilde{K},\tilde{L})=mr\cdot L(K,L)$.

7.2.3. The writhe. For evident reasons the above definitions do not generalize to define a "self-linking number" for a single oriented knot α. Our experience with intersection theory suggests that a self-linking number, if it exists, should be the ordinary linking number of α and a knot obtained by α via a "small movement". The trouble is that there is no canonical choice of how to perform such a movement, and different movements will lead to different linking numbers (this will be explained further below).

The solution is to equip our knot with extra data, in this case a **framing**—a nonzero section of the normal bundle of the knot. One can picture a framing as a collection of arrows that point perpendicularly away from the knot at every point. Framed knots are also called "ribbons", for the evident reason. The idea is that the framing specifies a "small movement" of α. If α is an oriented, framed knot then we may define an invariant $\mathrm{wr}(\alpha)$, called the **writhe** of α, to be the linking number of α and a knot obtained by moving α a tiny bit in the direction specified by the framing. A few remarks about this invariant:

- By changing the framing on a given knot, one can make the writhe equal to any integer. To see this, imagine cutting out a small portion of the framed knot and then "twisting" the framing through 360 degrees:



  If one thinks of the framing as modelling a ribbon, this is literally putting a full twist into the ribbon. It is not hard to see that the writhe of the new framed knot differs by ±1 from the original writhe (depending on the direction the framing was twisted), and so putting t twists in the framing will produce a difference of $\pm t$.
- Given a knot diagram, one can equip it with the so-called "blackboard framing" that comes directly out of the paper (or the blackboard) at the onlooker. For an oriented knot in the blackboard framing one can check that

$$\mathrm{wr}(\alpha)=\#(\text{positive crossings})-\#(\text{negative crossings}).$$

- The definition of the writhe can readily be applied to framed links as well as framed knots. Let L be an oriented, framed link, and let $K_1,\ldots,K_n$

denote the components. Then one checks that

$$\mathrm{wr}(L) = \sum_{i \neq j} L(K_i, K_j) + \sum_i \mathrm{wr}(K_i).$$

The symbiosis of linking numbers and writhes in this formula makes it clear that the writhe plays the role of a “self-linking number”.

7.2.4. Linking numbers and the Hopf invariant. Recall that $\pi_3(S^2) \cong \mathbb{Z}$. The Hopf invariant is simply the name of a map $\pi_3(S^2) \to \mathbb{Z}$ giving the isomorphism. For $f\colon S^3 \to S^2$ we now give four different descriptions of the Hopf invariant $H(f)$:

Version 1. Form the mapping cone of f, whereby a 4-cell is attached to S^2 via the map f. That is, $Cf = S^2 \cup_f e^4$. Then $H^*(Cf)$ is equal to $\mathbb{Z}$ in degrees 0, 2, and 4, and chosen orientations of S^2 and S^4 yield generators x and y of $H^2(Cf)$ and $H^4(Cf)$. Then x^2 is an integral multiple of y, and the integer in question is the Hopf invariant of the map f:

$$x^2 = H(f) \cdot y.$$

From this description it is not particularly clear that H is additive. But it is clear that $H(f) = 1$ when f is the Hopf map, because the cone of the Hopf map is $\mathbb{C}P^2$.

Verrsion 2. Let $x \in C^2(S^2)$ be a singular cocycle representing a chosen generator for $H^2(S^2)$. The $x \cup x$ is probably not zero, but it is certainly zero in cohomology: so choose $z \in C^3(S^2)$ such that $x \cup x = \delta z$.

Consider $f^*(x) \in C^2(S^3)$. This is also zero in cohomology, so $f^*(x) = \delta a$ for some $a \in C^1(S^3)$. One computes that $\delta(a \cup \delta a) = \delta a \cup \delta a = f^*(x \cup x) = f^*(z)$, and so $(a \cup \delta a) - f^*(z)$ is a cocycle. If $[y] \in H^3(S^3)$ is a chosen generator, we therefore have

$$[(a \cup \delta a) - f^*(z)] = H(f) \cdot [y]$$

for some integer $H(f)$.

One readily checks that the integer $H(f)$ does not depend on the choice of z or the choice of a. For latter, if $f^*(x) = \delta(a')$ then $a - a'$ is a 1-cocycle and so $a - a' = \delta(w)$ for some $w \in C^1(S^3)$. We then have

$$(a \cup \delta a - f^*(z)) - (a' \cup \delta a' - f^*(z)) = (a - a') \cup \delta a = \delta w \cup \delta a = \delta(w \cup \delta a),$$

and thus $[a \cup \delta a - f^*(z)] = [a' \cup \delta a' - f^*(z)]$.

The proof that this definition agrees with that of version 1 is purely homotopy theoretic; it would be a bit of a diversion to give it here, but we include it as an appendix.

Verrsion 3. Choose a point p in S^2 such that the fiber $f^{-1}(p)$ is connected. Choose a surface D in S^3 with $\partial D = f^{-1}(p)$ (this is always possible). Orient D so that if $\{u, v\}$ is an oriented basis for a tangent plane then $\{u, v, u \times v\}$ is an oriented basis on S^3. The map $f|_D\colon D \to S^3$ factors through $D/\partial D$ to give a map

$$D/\partial D \to S^2.$$

Since $D/\partial D$ is an oriented 2-sphere, we can obtain an integer by taking the degree of this map. This is the Hopf invariant $H(f)$.

Verrsion 4. For each p in S^2 consider the preimage $f^{-1}(p)$ in S^3. For generically chosen p this will give a link in S^3. Note that $f^{-1}(p)$ cannot intersect $f^{-1}(q)$ for $p \neq q$. From this it follows that the linking number $L(f^{-1}(p), f^{-1}(q))$ will not depend on the choice of p and q—moving q around, the preimage twists and contorts itself but since it never crosses $f^{-1}(p)$ the linking number will never change. This linking number is the Hopf invariant of f:

$$H(f) = L(f^{-1}(p), f^{-1}(q)) \quad \text{for generically chosen } p \text{ and } q \text{ in } S^2.$$

Version 5. Assume that f is smooth. Then $f^*(\mathrm{vol}_{S^2})$ is a closed 2-form on S^3, and since $H^2(S^3) = 0$ this means that $f^*(\mathrm{vol}_{S^2}) = dA$ for some $A \in \Omega^1(S^3; \mathbb{R})$. Then define

$$H(f) = \int_{S^3} A \wedge dA \in \mathbb{R}.$$

From this perspective it is not so clear that $H(f)$ is an integer!

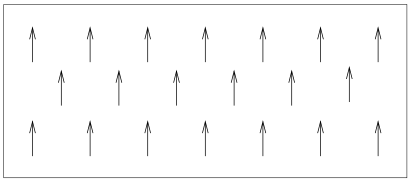
Note that this definition is a direct analog of the one from Version 2, except applied to de Rham cohomology rather than singular cohomology. In the de Rham theory one has $\mathrm{vol}_{S^2} \cup \mathrm{vol}_{S^2} = 0$ on the nose, not just in cohomology, and so one can avoid the choice of z that was needed for the singular theory. The fact that the above integral is independent of the choice of $A \in \Omega^1(S^3; \mathbb{R})$ follows just as in Version 2. We can also phrase the argument slightly differently: if $dA = dA'$ then

$$\int_{S^3} A \wedge dA - \int_{S^3} A' \wedge dA' = \int_{S^3} (A - A') \wedge dA = \int_{S^3} d\Big((A - A') \wedge A\Big) = 0$$

where the last equality is by Stokes's Theorem.

7.3. Polyakov's paper

Witten's paper relies on a paper by Polyakov [**P**] on linking. John Baez has posted some notes that explains some of the things going on in Polyakov's paper. The examples given involve the study of magnets. In any piece of iron each atom has a specific spin, i.e. for each atom one can associate a specific unit vector in a particular direction. If all of the spins of the atoms in a piece of iron line up then you will have a magnet.



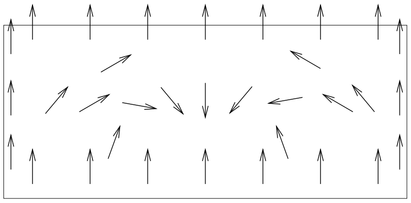
In two dimensions this can be modeled by a map

$$I^2 \longrightarrow S^2$$

where each point $x \in I^2$ is mapped to its spin vector. It will be useful to assume all of the spins on the boundary are lined up. So we can think of our map as

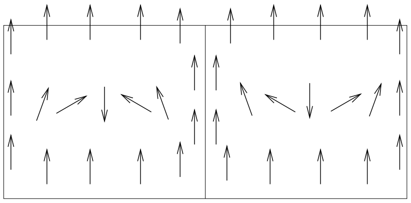
$$I^2/\partial I^2 \longrightarrow S^2$$

i.e. $S^2 \to S^2$. These maps are classified up to homotopy by their degree. What would maps of degree k look like for a given k? A degree 0 map would be the case in which each of the spins of elements in I^2 was lined up. A degree one map, in this model, woud look something like



This is somewhat hard to draw, but the idea is that as moving inward from the boundary the arrows rotate downwards, toward the center, until at the center they are pointing directly down. A map of degree -1 map would be the same kind of thing, but the spins would rotate outward. For a degree $k > 0$ map we would have the arrows rotating uniformly so that by the time they reached the center they had completed $k/2$ complete rotations.

Now, as topologists we know that a degree 0 map is the same (up to homotopy) as the sum of a degree 1 map and a degree -1 map. This sum can be depicted as follows:



Physicists look at this picture and think that it reminds them of a particle and an antiparticle, which can annihilate each other thereby leaving nothing. When thought of this way, the maps $S^2 \to S^2$ are called something like "topological solitions".

The states of our spin system are maps $S^2 \to S^2$. So what are the paths between states? We can draw such a path as something like this:

[Insert picture of a path between states]

Here we imagine the third (vertical) dimension as time with the states being equal on the top and bottom and an electromagnetic disturbance happening sometime in the middle. Thus, paths between states are given by maps

$$I^3/\partial I^3 = S^3 \to S^2$$

In terms of the physics, given $f \in Map(S^3, S^2)$ we associate an $S(f)$-action. Physically acceptable f are f such that $\delta S(f) = 0$. Polyakov defines a new action by adding a new term

$$S_{new}(f, A) = S_{old}(f) + \frac{\theta}{16\pi^2} \int_{S^3} A \wedge dA$$

where $S_{old}(f)$ is the action we have previously defined. This new term measures the linking between the particle and antiparticle,

[Diagram of interlinking particles]

Here states are maps $S^2 \longrightarrow S^3$ along with a 1-form on S^3 such that $dA = f^*(Vol_{S^2})$. Observe that the extreme points of S_{old} are equal to the extreme points of S_{new}, which means the new added term doesn't effect the "basic physics." Quantum mechanics gives us

$$\int e^{\frac{i}{\hbar}S_{new}(F,A)}$$

We now have phase changes in the exponentials. In other words

[Insert Two diagrams of intertwined particles]

Polyakov's point is to introduce these phase changes which give rise to probability computations which in turn should give rise to something good.

7.4. The Jones Polynomial

For every oriented link of a knot, L, we associate a rational function $V_L(q)$ such that the following holds

1. $V_0 = 1$
2. $V_{K \coprod L} = V_K \cdot V_L$
3.

[Insert Skein relation]

The last relation is called the Skein relation. The rules above allows one to define V for any knot.

Example 7.4.1. Consider the trefoil knot.

[Insert Skein relation for link of trefoil]

We can then repeat this process to compute V_L completely.

The hard thing to see is that these processes always gives the same answer. Witten wanted to give some explanation as to why this worked.

Let $G = SU(n)$, $R_i = \mathbb{C}^n$ be the standard representation, and M be a closed oriented framed 3 manifold

[Insert picture of manifold with loops]

We define

$$Z_{\gamma_1,\ldots,\gamma_r}(M) = \int_A w(\gamma_1, A)\cdots w(\gamma_r, A)e^{ki/4\pi}DA$$

If M has a boundary then we can associate to it a hilbert space $\mathcal{H}_{\partial M}$ and $Z_{\gamma_1,\ldots,\gamma_r}(M) \in \mathcal{H}_{\partial M}$. We have the following fact

$$\mathcal{H}_{\partial M} = \begin{cases} 1 \text{ dimensional} & \text{if } \partial M = S^2 \text{ with no points} \\ 2 \text{ dimensional} & \text{if } \partial M = S^2 \text{ 4 marked points} \end{cases}$$

[Insert picture of $M = S^3$ with loops]

If we focus in on a specific crossing and cut M across the crossing

[Insert picture of two manifolds with boundary; left, M_L, and right, M_R.]

Now $Z(M_L) \in \mathcal{H}_{\partial M_L}$ and $Z(M_R) \in \mathcal{H}_{\partial M_R}$. The orientation on M gives us the identification $\mathcal{H}_{\partial M_R} \cong \mathcal{H}^*_{\partial M_L}$ and

$$Z(M) = \langle Z(M_L), Z(M_R)\rangle = Z(M_R)(Z(M_L))$$

Consider the following

[Insert pictures of various intertwinings of loops on boundary, M_R, $M_{R'}$, and $M_{R''}$]

Now $Z(M_R), Z(M_{R'}), Z(M_{R''}) \in \mathcal{H}^*_{\partial M_L}$ and $\mathcal{H}^*_{\partial M_L}$ is a 2 dimensional space, which implies $Z(M_R), Z(M_{R'}), Z(M_{R''})$ are linearly dependent. Thus we have a relation

$$\alpha Z(M_R) + \beta Z(M_{R'}) + \gamma Z(M_{R''}) = 0$$

for $\alpha, \beta, \gamma \in \mathbb{C}$, which gives rise to the skein relation

$$\alpha Z(M) + \beta Z(M') + \gamma Z(M'') = 0$$

Hence for any given link in M one can perform the skein relation over and over again and see that $Z(M)$ is a rational function in α, β and γ.

CHAPTER 8

Quantum field theory

8.1. Introduction to functional integrals

8.1.1. A differentiation problem. Let B be a symmetric $n \times n$ matrix, and consider the function $Q\colon \mathbb{R}^n \to \mathbb{R}$ given by $Q(\mathbf{x}) = e^{\mathbf{x}^T B\mathbf{x}}$. Given a sequence $i_1, \dots, i_k$ of not-necessarily-distinct numbers from $\{1, \dots, n\}$, we wish to determine a formula for

$$\frac{\partial^k}{\partial x_{i_1} \cdots \partial x_{i_k}} \Big(e^{\mathbf{x}^T B\mathbf{x}} \Big).$$

For convenience we will just write $\partial_j = \frac{\partial}{\partial x_j}$.

First note that

$$\begin{aligned}\partial_j(e^{\mathbf{x}^T B\mathbf{x}}) &= e^{\mathbf{x}^T B\mathbf{x}} \cdot \Big[\sum_k x_k B_{k,j} + \sum_k B_{j,k} x_k\Big] \\ &= e^{\mathbf{x}^T B\mathbf{x}} \cdot \Big[2\sum_k B_{j,k} x_k\Big] = e^{\mathbf{x}^T B\mathbf{x}} \cdot \Big[2(B\mathbf{x})_j\Big].\end{aligned}$$

Here we have used that B is symmetric, and we are writing $(B\mathbf{x})_j$ for $(B\mathbf{x})_{j,1}$. For two derivatives we have

$$\partial_k \partial_j(e^{\mathbf{x}^T B\mathbf{x}}) = e^{\mathbf{x}^T B\mathbf{x}} \cdot \Big[4(B\mathbf{x})_k(B\mathbf{x})_j + 2B_{j,k}\Big]$$

and for three derivatives the formula becomes

$$\partial_l \partial_k \partial_j(e^{\mathbf{x}^T B\mathbf{x}}) = e^{\mathbf{x}^T B\mathbf{x}} \cdot \Big[8(B\mathbf{x})_k(B\mathbf{x})_j(B\mathbf{x})_l + 4B_{j,k}(B\mathbf{x})_j + 4B_{k,l}(B\mathbf{x})_j + 4(B\mathbf{x})_k B_{j,l}\Big].$$

It is not hard to see the pattern here. First note that the powers of 2 are easy to predict: one should imagine that every time a "B" appears it comes with a "2" attached. In general we will have

$$\partial_{i_1} \cdots \partial_{i_m}(e^{\mathbf{x}^T B\mathbf{x}}) = e^{\mathbf{x}^T B\mathbf{x}} \cdot \Big[????\Big]$$

where inside the brackets is a sum in which each term has an appropriate power of 2 and a product of things that look like either $(B\mathbf{x})_i$ or $B_{r,s}$. To keep track of the terms one can imagine graphs whose vertices are $i_1, i_2, \dots, i_m$ and where every vertex is attached to at most one edge (and where there are no loops). These are simply graphs where some vertices are paired together are some vertices are free. [Note: If the set of numbers $i_1, \dots, i_m$ has repetitions it is important that the vertex set also have repetitions]. One readily checks that the general formula is

$$\partial_{i_1} \cdots \partial_{i_m}(e^{\mathbf{x}^T B\mathbf{x}}) = e^{\mathbf{x}^T B\mathbf{x}} \cdot \sum_{\Lambda} \Big[2^{\#f(\Lambda) + \#\mathcal{E}(\Lambda)} \cdot \prod_{i \in f(\Lambda)} (B\mathbf{x})_i \cdot \prod_{e \in \mathcal{E}(\Lambda)} B_{e_0, e_1}\Big]$$

where the sum runs over all graphs Λ. Here $\mathcal{E}(\Lambda)$ denotes the edge set of Λ, e_0 and e_1 are the vertices of the edge e, and $f(\Lambda)$ is the set of free vertices of Λ (the vertices that are not attached to an edge).

Let's try an example. To compute $\partial_1^2\partial_2\partial_3(e^{\mathbf{x}^TB\mathbf{x}})$ we first make a list of the appropriate graphs. The vertex set will be $\{1a, 1b, 2, 3\}$ (we write $1a$ and $1b$ to distinguish the two different vertices corresponding to the index 1). There is exactly one graph with no edges, $\binom{4}{2} = 6$ graphs with one edge, and for two edges we have the following three graphs:

?????

We can immediately write down the following formula:

$$\begin{aligned}\partial_{i_1}\cdots\partial_{i_m}(e^{\mathbf{x}^TB\mathbf{x}}) =& e^{\mathbf{x}^TB\mathbf{x}}\cdot\Big[16(B\mathbf{x})_1(B\mathbf{x})_1(B\mathbf{x})_2(B\mathbf{x})_3 + 8B_{1,1}(B\mathbf{x})_2(B\mathbf{x})_3\\ &+ 8B_{1,2}(B\mathbf{x})_1(B\mathbf{x})_3 + 8B_{1,3}(B\mathbf{x})_1(B\mathbf{x})_2 + 8B_{2,3}(B\mathbf{x})_1(B\mathbf{x})_1\\ &+ 4B_{1,1}B_{2,3} + 4B_{1,2}B_{1,3} + 4B_{1,2}B_{1,3}\Big]\end{aligned}$$

(note that the last two terms are identical: we have not combined them to accentuate that they came from two different graphs).

From our general formula we immediately deduce the following consequence, known as Wick's Theorem in the physics literature:

Proposition 8.1.2 (Wick's Theorem). *If m is odd then $\partial_{i_1}\cdots\partial_{i_m}(e^{\mathbf{x}^TB\mathbf{x}})\Big|_{\mathbf{x}=0} = 0$. If m is even then*

$$\partial_{i_1}\cdots\partial_{i_m}(e^{\mathbf{x}^TB\mathbf{x}})\Big|_{\mathbf{x}=0} = 2^{\frac{m}{2}}\sum_{\Lambda}\Bigg[\prod_{e\in\mathcal{E}(\Lambda)} B_{e_0,e_1}\Bigg]$$

where the graphs in the sum run over all complete pairings of the vertex set $\{i_1,\dots,i_m\}$ (that is, graphs where every vertex belongs to exactly one edge).

Believe it or not, the graphs that we are seeing in Wick's Theorem are the beginning of Feynman diagrams!

Exercise 8.1.3. When $m = 2r$, prove that the number of graphs Λ appearing in the sum from Wick's Theorem is equal to $\frac{1}{2^r}\cdot\frac{(2r)!}{r!}$. Prove that this number equals $(m-1)(m-3)(m-5)\cdots 3\cdot 1$. So when $m = 8$ there are 105 terms in the sum.

8.1.4. Some elementary functional integrals. We want to eventually understand something about integrals of the form

$$Z = \int_{\mathcal{A}} e^{\frac{i}{\hbar}S(\gamma)}\mathcal{D}\gamma$$

where $\mathcal{A}$ is some infinite-dimensional space like a space of paths. Sometimes we will have an "observable" $T\colon \mathcal{A}\to\mathbb{R}$ and we would also like to calculate the expectation value

$$\langle T\rangle = \frac{1}{Z}\int_{\mathcal{A}} T(\gamma)e^{\frac{i}{\hbar}S(\gamma)}\mathcal{D}\gamma.$$

Quantum field theory is largely about computing these kinds of integrals.

In this section we will explore analagous integrals where $\mathcal{A} = \mathbb{R}^n$. We will assume that $S, T\colon \mathbb{R}^n\to\mathbb{R}$ are polynomial functions. It will suffice to assume that T has the form $T(\mathbf{x}) = x_{i_1}x_{i_2}\cdots x_{i_m}$ for not-necessarily-distinct indices $i_1,\dots,i_m$.

To get started, let us recall how these computations are done for the basic Gaussian case where $S(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ and A is positive-definite. There we know that $Z = \sqrt{\frac{(2\pi\hbar i)^n}{\det A}}$. The trick for finding $\langle x_{i_1} \cdots x_{i_m}\rangle$ is to realize that

$$\int_{\mathbb{R}^n} x_{i_1} \cdots x_{i_m} e^{\frac{i}{\hbar}S(x)} dx = \frac{\partial^m}{\partial j_{i_1} \cdots \partial j_{i_m}} \int_{\mathbb{R}^n} e^{\frac{i}{\hbar}S(x)+j_1x_1+j_2x_2+\cdots+j_nx_n}\, dx.$$

We know by completing the square that

$$\int_{\mathbb{R}^n} e^{\frac{i}{\hbar}S(\mathbf{x})+J\mathbf{x}}\, d\mathbf{x} = Z \cdot e^{-\frac{\hbar}{2i}J^TA^{-1}J}$$

for any $n \times 1$ matrix J. So

$$\langle x_{i_1} \cdots x_{i_m}\rangle = \frac{\partial^m}{\partial j_{i_1} \cdots \partial j_{i_m}} e^{-\frac{\hbar}{2i}J^TA^{-1}J}\Bigg|_{J=0}.$$

Now consider a more general action functional, of the form

$$S(\mathbf{x}) = S(0) + \mathbf{x}^T A\mathbf{x} + \text{higher order terms}$$

where A is a symmetric, positive-definite $n \times n$ matrix. We might as well assume $S(0) = 0$, since otherwise this just adds a scalar coefficient to the integral. Write $M(\mathbf{x})$ for the higher-order-terms, so that we have $S(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} + M(\mathbf{x})$. Then

$$e^{\frac{i}{\hbar}S(\mathbf{x})} = e^{\frac{i}{\hbar}\mathbf{x}^TA\mathbf{x}} \cdot e^{\frac{i}{\hbar}M(\mathbf{x})} = e^{\frac{i}{\hbar}\mathbf{x}^TA\mathbf{x}} \cdot \Big[1 + \frac{i}{\hbar}M(\mathbf{x}) - \frac{1}{2\hbar^2}M(\mathbf{x})^2 - \frac{i}{6\hbar^3}M(\mathbf{x})^3 + \cdots\Big]$$

The expression in brackets is just a power series in the x_j's, therefore we know how to integrate it term by term against the quadratic $e^{\frac{i}{\hbar}\mathbf{x}^TA\mathbf{x}}$. Let us write Z_A and $\langle T\rangle_A$ for the partition function and expectation values computed with respect to $S(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$. Then we observe that

$$Z = Z_A \cdot \Big[1 + \frac{i}{\hbar}\langle M(\mathbf{x})\rangle_A - \frac{1}{2\hbar^2}\langle M(\mathbf{x})^2\rangle_A - \frac{i}{6\hbar^3}\langle M(\mathbf{x})^3\rangle_A + \cdots\Big].$$

As an example, let us assume that we only have cubic terms in $M(\mathbf{x})$. We can write

$$M(\mathbf{x}) = \sum_{i,j,k} C_{ijk}x_ix_jx_k$$

where C is symmetric in the three indices: $C_{ijk} = C_{\sigma(i)\sigma(j)\sigma(k)}$ for any permutation σ of three elements. Wick's Theorem tells us that $\langle M(\mathbf{x})\rangle_A = 0$ (because 3 is an odd number), and likewise for all the odd powers of $M(\mathbf{x})$.

Part 4

Appendices

APPENDIX A

Background on differential geometry

A.1. Introduction to connections

Recall the definition of an affine connection:

Definition A.1.1. *Let $E \to B$ be a smooth vector bundle with B a smooth manifold. An **affine connection** on E is an assignment*

$$\Gamma(TB) \times \Gamma(E) \to \Gamma(E), \qquad (\mathfrak{X}, s) \mapsto D_{\mathfrak{X}} s$$

having the following properties:

(1) $D_{a_1\mathfrak{X}_1+a_2\mathfrak{X}_2}(s) = a_1 D_{\mathfrak{X}_1}(s) + a_2 D_{\mathfrak{X}_2}(s)$, for any $a_1, a_2 \in C^\infty(B)$, $\mathfrak{X}_1, \mathfrak{X}_2 \in \Gamma(TB)$.

(2) $D_{\mathfrak{X}}(s+t) = D_{\mathfrak{X}}(s) + D_{\mathfrak{X}}(t)$ for any $s, t \in \Gamma(E)$.

(3) $D_{\mathfrak{X}}(f \cdot s) = (\partial_{\mathfrak{X}} f) \cdot s + f \cdot D_{\mathfrak{X}}(s)$ for any smooth map $f\colon B \to \mathbb{R}$.

A.1.2. Local properties of connections.

Proposition A.1.3. *$(D_{\mathfrak{X}} s)(b)$ only depends on $\mathfrak{X}(b)$ and the values of s along a curve, defined in a neighborhood of b, with $\mathfrak{X}(b)$ as tangent vector.*

Proof. We first argue that if $\mathfrak{X}$ vanishes in a neighborhood of b, then $[D_{\mathfrak{X}} s](b) = 0$. Let $U' \subset U$ be a neighborhood of b such that $\overline{U'} \subseteq U$. Choose a function $f\colon B \to \mathbb{R}$ such that $f(U') = 0$ and $f(B - U) = 1$. Then $\mathfrak{X} = f \cdot \mathfrak{X}$, and so $D_{\mathfrak{X}} s = D_{f\mathfrak{X}} s = f \cdot (D_{\mathfrak{X}} s)$. Evaluating at b and using $f(b) = 0$, we get $[D_{\mathfrak{X}} s](b) = 0$.

If $\mathfrak{X}$ and $\mathfrak{X}'$ are vector fields that agree within a neighborhood of b, then $D_{\mathfrak{X}} s - D_{\mathfrak{X}'} s = D_{\mathfrak{X}-\mathfrak{X}'} s$ and by the previous pargraph this vanishes at b. We might say that $[D_{\mathfrak{X}} s](b)$ only depends on the local properties of $\mathfrak{X}$ near b.

The same argument shows that if s and s' are sections that agree within a neighborhood of b, then $[D_{\mathfrak{X}} s](b) = [D_{\mathfrak{X}'} s](b)$.

Now assume that $\mathfrak{X}$ is a vector field such that $\mathfrak{X}(b) = 0$. Let $e_1, \dots, e_n$ be sections of TB that give a local basis at b, and write $\mathfrak{X} = \sum_j f_j e_j$ within a neighborhood of b. Note that $f_j(b) = 0$, for all j. But then

$$[D_{\mathfrak{X}} s](b) = [D_{\sum f_j e_j} s](b) = \sum_j f_j(b) \cdot [D_{e_j} s](b) = 0.$$

Again, linearity in the $\mathfrak{X}$ variable now shows that if $\mathfrak{X}(b) = \mathfrak{X}'(b)$ then $[D_{\mathfrak{X}} s](b) = [D_{\mathfrak{X}'} s](b)$. If $v = \mathfrak{X}(b)$, we will for the rest of the proof write $[D_v s](b)$ instead of $[D_{\mathfrak{X}} s](b)$.

Finally, suppose that $\gamma\colon [-1, 1] \to B$ is a smooth curve such that $\gamma(0) = b$ and $\gamma'(0) = v$. Suppose that $s \circ \gamma$ vanishes in a neighborhood of b. Let $e_1, \dots, e_k$ be a basis of local sections for E near b, are write $s(x) = \sum_j f_j(x)e_j(x)$ where the f_j are smooth, real-valued functions (defined in a neighborhood of b). Then

$$[D_v s](b) = \sum_j \Big[(\partial_v f_j)(b)e_j(b) + f_j(b)[D_v e_j](b)\Big].$$

But each f_j vanishes along γ is a neighborhood of b, so both $f_j(b)$ and $\partial_v f_j(b)$ will be zero. Hence $[D_v s](b) = 0$.

As usual, the above paragraph implies that if s and s' are two sections that agree along γ then $[D_v s](b) = [D_v s'](b)$. □

Remark A.1.4. Because of the above proposition, the affine connection D can be thought of as giving much more than just the pairing $\Gamma(TB) \otimes \Gamma E \to \Gamma E$. In fact it gives us a way of making sense of $D_v s(b)$ whenever v is a tangent vector at b and s is a section defined on some smooth curve through b with tangent vector v: one chooses a vector field $\mathfrak{X}$ with $\mathfrak{X}(b) = v$ and a global section $S\colon B \to E$ that extends s (it is always possible to do both) and defines $D_v s(b)$ to be $[D_{\mathfrak{X}} S](b)$. The above proposition shows that this construction is independent of the choices. Our main use of this will be when s is a section defined on some neighborhood of b.

A.1.5. Connections on trivial bundles. Assume that $s_1, \dots, s_n$ is a trivializing basis of sections for a bundle $E \to B$. Given a connection D, then we can write

$$[D_v s_j](b) = w_{j1}(v)_b s_1(b) + w_{j2}(v)_b s_2(b) + \cdots + w_{jn}(v)_b s_n(b)$$

for uniquely-defined real numbers $w_{jk}(v)_b$. We will usually write this more succinctly as

$$Ds_j = w_{j1}s_1 + \cdots + w_{jn}s_n. \tag{A.1.6}$$

Note that each w_{jk} is a 1-form on B: it takes a point $b \in B$ and a tangent vector v at that point, and outputs a real number.

Define Ω to be the matrix of 1-forms (ω_{jk}). This is called the **connection matrix** for D with respect to the basis $s_1, \dots, s_n$.

One can in fact check that any matrix of 1-forms defines an affine connection on E, via the formulas above. More precisely, given Ω we define an affine connection D as follows. Given a section $\chi = \sum_j \chi_j s_j$, we define

$$D_v \chi = \sum_j (\partial_v \chi_j)s_j + \sum_{j,k} \chi_j w_{jk}(v) s_k.$$

Note that this is just a matter of starting with (A.1.6) and forcing the Leibniz Rule. One readily checks that D is indeed an affine connection.

In the case when Ω is the zero matrix, the associated connection is called the **standard flat connection** on E. To be completely clear, this is the connection given by

$$D_{\mathfrak{X}}(\chi) = \sum_j (\partial_{\mathfrak{X}} \chi_j)s_j.$$

Note that this depends on the chosen trivialization, however—so the use of the word "standard" can seem misleading. This is the unique connection satisfying $D_{\mathfrak{X}}(s_j) = 0$ for all j and $\mathfrak{X}$.

The upshot of this discussion is that we know all possible connections on a rank n trivial bundle. Once you pick a trivialization, they correspond to elements of $\Omega^1(B;\mathbb{R}^{n^2})$ (note that this can be canonically identified with the space of $n\times n$ matrices whose entries are in $\Omega^1(B;\mathbb{R})$). We can think of $\Omega^1(B;\mathbb{R}^{n^2})$ as representing the "space of all connections" on the bundle E.

As always in linear algebra, it is useful to know what happens when one changes basis:

Proposition A.1.7. *Let $e_1,\dots,e_n$ and $s_1,\dots,s_n$ be two trivializing bases for E. Given a connection D on E, let Ω^e and Ω^s be the connection matrices with respect to the e-basis and s-basis, respectively. Write $e_j=\sum\alpha_{ij}s_j$, where the α_{ij} are smooth functions $B\to\mathbb{R}$. Then*

$$\Omega^e=(d\alpha)\alpha^{-1}+\alpha\Omega^s\alpha^{-1}. \tag{A.1.8}$$

In differential geometry half the battle is figuring out what the statements actually say. In the above proposition, α is an $n\times n$ matrix of C^∞ functions. Then α^{-1} is another $n\times n$ matrix of C^∞ functions, which can for instance be obtained from α by the usual formula using the classical adjoint. Note that $(\alpha^{-1})_{jk}\neq(\alpha_{jk})^{-1}$, so one must really be careful to pay attention to context when parentheses are dropped.

The symbol Ω^s is an $n\times n$ matrix of 1-forms, and $\alpha\Omega^s\alpha^{-1}$ denotes the matrix of 1-forms whose (j,k) entry is $\sum_{p,q}\alpha_{jp}\omega^s_{pq}(\alpha^{-1})_{qk}$. Likewise, $d\alpha$ must mean the matrix of 1-forms obtained by applying d to each entry in α. Then $(d\alpha)\alpha^{-1}$ is the matrix whose (j,k)-entry is $\sum_p d\alpha_{jp}\cdot(\alpha^{-1})_{pk}$. So the proposition says that

$$\omega^e_{jk}=\sum_p d\alpha_{jp}\cdot(\alpha^{-1})_{pk}+\sum_{p,q}\alpha_{jp}\omega^s_{pq}(\alpha^{-1})_{qk}.$$

Proof of Proposition A.1.7. It will be convenient to use Einstein summation convention in this proof. So we will write $e_j=\alpha_{jk}s_k$, and inversely $s_u=(\alpha^{-1})_{uv}e_v$. We now just calculate:

$$\begin{aligned}\omega^e_{jm}(v)e_m=D_ve_j=D_v(\alpha_{jk}s_k)&=(\partial_v\alpha_{jk})\cdot s_k+\alpha_{jk}D_vs_k\\&=(\partial_v\alpha_{jk})\cdot s_k+\alpha_{jk}\omega^s_{kl}(v)s_l\\&=(\partial_v\alpha_{jk})\cdot(\alpha^{-1})_{km}e_m+\alpha_{jk}\omega^s_{kl}(v)(\alpha^{-1})_{lm}e_m.\end{aligned}$$

We see immediately that $\omega^e_{jm}=d\alpha_{jk}\cdot(\alpha^{-1})_{km}+\alpha_{jk}\omega^s_{kl}(\alpha^{-1})_{lm}$, which is what we wanted. □

The transformation law of (A.1.8) is very important in the theory of connections, as we will begin to see in the next section.

A.1.9. Constructing connections on general bundles. Here are some of the most useful ways for constructing connections on a bundle $E\to B$:

(1) Pick local trivializations for the bundle, choose connections on each piece, and make sure they patch together correctly.
(2) Pick local trivializations and connections on each piece, then "average them" together using a partition of unity to make a global connection.
(3) Realize the bundle as a retract of some other bundle $\mathcal{E}\to B$ that already has a connection on it.
(4) Start with a connection on E that one already has, and alter it to make a new connection on E.

For the method of (1), imagine that U and V are open sets in B over which the bundle is trivial. Let $u_1, \dots, u_n$ and $v_1, \dots, v_n$ be a basis of sections defined on U and V, respectively. A connection on $E|_U \to U$ is specified by a connection matrix Ω^U of 1-forms on U, and likewise a connection on $E|_V \to V$ is specified by a matrix Ω^V of 1-forms on V.

On the overlap $U \cap V$ we may write $v_i = \sum \alpha_{ij} u_j$. Based on Proposition A.1.7, the connections on $E|_U$ and $E|_V$ patch together to give a connection on $E|_{U \cup V}$ if we have $\Omega^V = (d\alpha)\alpha^{-1} + \alpha\Omega^U\alpha^{-1}$ on $U \cap V$.

In general, one has a trivializing cover $\{U_\sigma\}$, a basis of sections s^σ defined on each element of the cover, and matrices Ω^σ of 1-forms on U_σ. These patch together to give a connection on the bundle E if the compatibility condition (A.1.8) is satisfied for every pair of indices σ and σ'. Of course this may be a lot to check in practice, and constructing connections in this way is of somewhat limited use. It is also not immediately clear whether the construction is possible for any bundle.

Method (2) is based on the observation that if D and D' are connections on a bundle then so is $tD + (1-t)D'$, for any $t \in \mathbb{R}$. This is a very easy check which we leave to the reader. Note that although we haven't yet introduced a "space" of connections, this observation suggests that once we do have such a space it will be contractible (or even more, convex in some sense).

Pick a trivializing cover $\{U_\sigma\}$ for the bundle $E \to B$, and pick connections D_σ for each $E|_{U_\sigma} \to U_\sigma$. Choose a partition of unity $\{\psi_\sigma\}$ suppordinate to this cover, and define $D = \sum_\sigma \psi_\sigma D_\sigma$. One readily checks that D is a connection on the bundle E. We have therefore proven:

Proposition A.1.10. *Every bundle admits a connection.*

Although the above construction works, it would not be at all pleasant to use in practice. A more accessible construction is given by the next method.

The method of (3) hinges on the following simple result:

Proposition A.1.11. *Suppose that $E \hookrightarrow \mathcal{E}$ is a subbundle which admits a bundle retraction $r\colon \mathcal{E} \to E$. Then a connection $\mathcal{D}$ on $\mathcal{E}$ induces a connection D on E by the formula*

$$D_v(s) = r[\mathcal{D}_v(s)].$$

Proof. Easy. □

For example, suppose that $\mathcal{E}$ is a bundle with a metric on it, and $E \subseteq \mathcal{E}$ is a subbundle. Then orthogonal projection gives a retraction $\mathcal{E} \to E$, and therefore any connection on $\mathcal{E}$ induces one on E. As any bundle over a compact CW-complex can be embedded in a trivial bundle (and such a trivial bundle can of course be given a metric), this method also guarantees the existence of connections on bundles over such spaces.

To demonstrate the efficacy of this method we offer the following example:

Example A.1.12. Consider the tautological line bundle $L \to \mathbb{C}P^1$. This embeds into a rank 2 trivial bundle in the usual way: points of L are pairs (l, x) where l is a line in $\mathbb{C}^2$ and x is a point on that line, and this representation exactly gives an embedding $L \hookrightarrow \mathbb{C}P^1 \times \mathbb{C}^2$.

The bundle L is trivialized by the open cover $\{U, V\}$ where U is the set of points $[z : w]$ with $w \neq 0$ and V is the set of points $[z : w]$ with $z \neq 0$. Let z be the standard complex coordinate on $\mathbb{C}P^1$, so that really z is an abbreviation for the point $[z : 1]$; likewise, use w as an abbreviation for the point $[1 : w]$. Note that on the intersection of these two coordinate patches we have $z = \frac{1}{w}$.

Trivializing sections over each piece of the cover are given by

$$s_U(z) = (z, 1) \quad \text{and} \quad s_V(w) = (1, w).$$

Note that on $U \cap V$ we have $s_V = \frac{1}{z} s_U$. Let $\alpha(z) = \frac{1}{z}$. Then by (A.1.8) one way to specify a connection on L would be to give 1-forms (really 1×1 matrices of 1-forms) Ω^U and Ω^V subject to the following transformation condition:

(A.1.13) $\Omega^V = (d\alpha)\alpha^{-1} + \alpha\Omega^U\alpha^{-1} = -\frac{1}{z^2} dz \cdot z + \Omega^U = -\frac{1}{z}\, dz + \Omega^U.$

Perhaps it is not so clear how to choose such a Ω^U and Ω^V, however.

Let us instead use method (3) to induce a connection on L via the embedding $L \hookrightarrow \mathbb{C}P^1 \times \mathbb{C}^2$. We give $\mathbb{C}P^1 \times \mathbb{C}^2$ the standard flat connection with respect to the standard basis of $\mathbb{C}^2$. Recall that this means that if $s = (s_1, s_2)$ is a section of this bundle then

$$[D_v s](b) = \Big((\partial_v s_1)(b),\, (\partial_v s_2)(b)\Big).$$

We also give $\mathbb{C}P^1 \times \mathbb{C}^2$ the standard Hermitian metric.

Recall that for a unit vector $u \in \mathbb{C}^2$, projection onto the line spanned by u is given by $x \mapsto (x \cdot u)u$. So if $(a, b) \in \mathbb{C}^2$ is an arbitrary nonzero vector, projection onto the line spanned by (a, b) is given by

$$(z, w) \mapsto \left[(z, w) \cdot \frac{1}{\sqrt{a\bar{a} + b\bar{b}}}(a, b)\right] \frac{1}{\sqrt{a\bar{a} + b\bar{b}}}(a, b) = \frac{1}{a\bar{a} + b\bar{b}}\Big[(z, w) \cdot (a, b)\Big](a, b).$$

The induced connection on L is therefore given as follows. For the section s_U we have

$$[D_v s_U](z) = \left[\frac{1}{z\bar{z} + 1}(\partial_v z, 0) \cdot (z, 1)\right](z, 1) = \left[\frac{(\partial_v z)\bar{z}}{z\bar{z} + 1}\right](z, 1)$$

and so $\Omega^U = \frac{\bar{z}\, dz}{z\bar{z}+1}$ for this connection. Likewise,

$$[D_v s_V](w) = \left[\frac{1}{w\bar{w} + 1}(0, \partial_v w) \cdot (1, w)\right](1, w) = \left[\frac{(\partial_v w)\bar{w}}{w\bar{w} + 1}\right](w, 1)$$

and so $\Omega^V = \frac{\bar{w}\, dw}{w\bar{w}+1}$.

Because D is a connection on L, the patching equation (A.1.13) is automatically satisfied. However, it is worth checking this by hand just to be assured that everything is working as it should: use that $w = \frac{1}{z}$, and therefore $dw = -\frac{1}{z^2}\, dz$. (Special thanks to Gustavo Granja for a useful conversation about this example.)

Example A.1.14. As a second example, consider the tangent bundle $TS^2 \to S^2$. The usual embedding $S^2 \in \mathbb{R}^3$ gives an embedding to TS^2 into the rank 3 trivial bundle. This induces a connection on TS^2. If s is a section of TS^2, write $s(p) = (s_1(p), s_2(p), s_3(p))$ for $p \in \mathbb{S}^2$ (using the embedding $TS^2 \hookrightarrow S^1 \times \mathbb{R}^3$). The connection on TS^2 is then given by

$$[D_v s](p) = \pi_p(\partial_v s_1(p), \partial_v s_2(p), \partial_v s_3(p))$$

where π_p denotes projection from $\mathbb{R}^3$ onto T_pS^2. This projection is given by the formula $\pi_p(w) = w - (w \cdot p)p$, and using this we compute that if $p = (x, y, z)$ then

$$\begin{aligned}[D_v s](p) = \big[(1-x^2)\partial_v s_1 - xy\partial_v s_2 - xz\partial_v s_3,\ -xy\partial_v s_1 + (1-y^2)\partial_v s_2 - yz\partial_v s_3,\\ - xz\partial_v s_1 - yz\partial_v s_2 + (1-z^2)\partial_v s_3\big].\end{aligned}$$

The usefulness of this description is somewhat unclear, as it does not make use of intrinsic coordinate systems on S^2 or TS^2 (both of which are two-dimensional rather than three-dimensional).

Finally, we turn to method (4). The basic thing to notice is that if D and D' are two affine connections on $E \to B$, then $D - D'$ is $C^\infty(B)$-linear in the section variable. In other words, $(D-D')_{\mathfrak{X}}(fs) = f \cdot (D-D')_{\mathfrak{X}} s$ for all smooth functions $f\colon B \to \mathbb{R}$. This is easy to see, as

$$\begin{aligned}(D-D')_{\mathfrak{X}}(f \cdot s) &= [(\partial_{\mathfrak{X}} f) \cdot s + f \cdot D_{\mathfrak{X}}(s)] - [(\partial_{\mathfrak{X}} f) \cdot s + f \cdot D'_{\mathfrak{X}}(s)]\\ &= f \cdot (D_{\mathfrak{X}} - D'_{\mathfrak{X}})(s).\end{aligned}$$

To paraphrase the above, $(D - D')_{\mathfrak{X}} \colon \Gamma(E) \to \Gamma(E)$ is $C^\infty(B)$-linear. But it is a general fact of bundle theory that any such $C^\infty(B)$-linear map must come from a map of bundles. So we may regard $(D-D')_{\mathfrak{X}}$ as a map $E \to E$, or as a section of $\underline{\mathrm{End}}(E)$. The assignment

$$\Gamma(TB) \to \Gamma(\underline{\mathrm{End}}(E)), \qquad \mathfrak{X} \mapsto (D-D')_{\mathfrak{X}}$$

is also $C^\infty(B)$-linear (because both D and D' are $C^\infty(B)$-linear in the $\mathfrak{X}$-variable), and so we actually have a bundle map $TB \to \underline{\mathrm{End}}(E)$. Call this bundle map A, and regard it as an element of $\Gamma(\underline{\mathrm{Hom}}(TB, \underline{\mathrm{End}}(E)))$. Finally, note the isomorphism

$$\underline{\mathrm{Hom}}(TB, \underline{\mathrm{End}}(E)) \cong \underline{\mathrm{Hom}}(TB, \mathbb{R}) \otimes \underline{\mathrm{End}}(E) = T^*B \otimes \underline{\mathrm{End}}(E)$$

so that the sections of this bundle are the 1-forms $\Omega^1(B; \underline{\mathrm{End}}(E))$. We will regard A as belonging to this space.

The content of the above paragraph is perhaps overly obtuse, as we are not actually doing anything substantial—we are just reorganizing information. If v is a tangent vector to B at b then A_v is the endomorphism $E_b \to E_b$ given as follows: for $p \in E_b$, take any section s such that $s(p) = b$ and send p to $D_v s(b) - D'_v s(b)$. While it seems that this depends on the choice of s, it actually does not. This process describes our 1-form $A \in \Omega^1(B; \underline{\mathrm{End}}(E))$.

Elements of $\Omega^1(B; \underline{\mathrm{End}}(E))$ will be called **connection potentials**.

Proposition A.1.15. *Fix a connection D^0 on $E \to B$. Then the set of affine connections on E is in bijective correspondence with the vector space $\Omega^1(B; \underline{\mathrm{End}}(E))$. To better illustrate the bijection it is useful to write*

$$\{\textit{space of affine connections on } E\} = D^0 + \{\textit{space of connection potentials.}\}$$

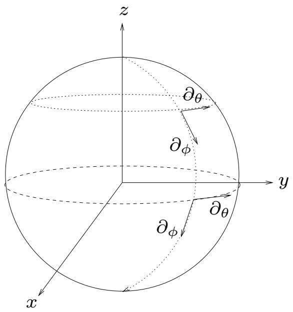
Proof. Above we showed the $\subseteq$ inclusion: given any affine connection D, then $D - D^0$ is a connection potential. For the reverse direction, check that if $A \in \Omega^1(B; \underline{\mathrm{End}}(E))$ then $D_{\mathfrak{X}}(s) = D^0_{\mathfrak{X}}(s) + A_{\mathfrak{X}}(s)$ is an affine connection. This is easy, and left to the reader. □

Note, as a corollary, that the space of connections on E is contractible—because $\Omega^1(B; \underline{\mathrm{End}}(E))$ is contractible, being a real vector space.

Remark A.1.16. We have seen an incarnation of $\Omega^1(B;\underline{\mathrm{End}}(E))$ once before, where it was a little simpler. If E is trivial of rank n, then choosing a basis of sections for E allows us to identify $\underline{\mathrm{End}}(E)$ with the trivial bundle of rank n^2. Then $\Omega^1(B;\underline{\mathrm{End}}(E)) = \Omega^1(B;\mathbb{R}^{n^2})$, and this is precisely the space of all connection matrices as discussed in Section A.1.5.

To get a sense of the above method for constructing affine connections, let us consider an example.

Example A.1.17. Let $B = S^2$ and $E = TS^2$ be the tangent bundle. Let θ and ϕ denote the usual spherical coordinates on S^2, which give local coordinates everywhere except the north and south poles. Write ∂_θ and ∂_ϕ for the associated vector fields. These are depicted below:



A section of TS^2 can be written (away from the poles) as

$$s(\theta,\phi) = s_1(\theta,\phi)\cdot\partial_\theta + s_2(\theta,\phi)\cdot\partial_\phi$$

An arbitrary affine connection on TS^2 will then look like

$$(D_{\partial_\theta}s) = \left(\frac{\partial s_1}{\partial\theta}\partial_\theta + \frac{\partial s_2}{\partial\theta}\partial_\phi\right) + (Bs_1(\theta,\phi)+Cs_2(\theta,\phi))\partial_\theta + (Es_1(\theta,\phi)+Fs_2(\theta,\phi))\partial_\phi$$

$$(D_{\partial_\phi}s) = \left(\frac{\partial s_1}{\partial\phi}\partial_\theta + \frac{\partial s_2}{\partial\phi}\partial_\phi\right) + (Gs_1(\theta,\phi)+Hs_2(\theta,\phi))\partial_\theta + (Is_1(\theta,\phi)+Js_2(\theta,\phi))\partial_\phi$$

where the variables B through J denote smooth functions of θ and ϕ. Note that the first terms in parentheses denote the standard flat connection with respect to the basis ∂_θ, ∂_ϕ, whereas the later terms represent the connection potential A. Here

$$A_{\partial_\theta} = \begin{bmatrix} B & C \\ E & F \end{bmatrix} \qquad \text{and} \qquad A_{\partial_\phi} = \begin{bmatrix} G & H \\ I & J \end{bmatrix},$$

where the matrices are interpreted as elements of $\mathrm{End}(TS^2)$ defined away from the poles, with respect to the basis ∂_θ, ∂_ϕ.

This is essentially all that we really know at this point, all that our general machinery tells us so far. But to get a sense of how to put this information to use, let's ask a natural question. The bundle TS^2 comes to us with a metric on it, induced by the embedding $S^2 \hookrightarrow \mathbb{R}^3$. What is necessary for the above affine connection to be compatible with the metric? That is, we want

$$\partial_v(s\cdot t) = (D_v s)\cdot t + s\cdot(D_v t) \tag{A.1.18}$$

for all sections s and t of TS^2.

To recall how the metric works, we need to convert to Cartesian coordinates. Recall that

$$x = \cos\theta\sin\theta, \quad y = \sin\theta\sin\phi, \quad z = \cos\phi.$$

Differentiating, we then find that

$$\partial_\theta = [\sin\theta\sin\phi, \cos\theta\sin\theta, 0], \qquad \partial_\phi = [\cos\theta\cos\phi, \sin\theta\cos\phi, -\sin\phi].$$

So

$$\partial_\theta\cdot\partial_\theta = \sin^2\phi, \quad \partial_\theta\cdot\partial_\phi = 0, \quad \partial_\phi\cdot\partial_\phi = 1.$$

(Note that this conforms to our picture: the vectors ∂_ϕ always have the same size, but the vectors ∂_θ get smaller as they approach the poles).

If we have sections $s = s_1\partial_\theta + s_2\partial_\phi$ and $t = t_1\partial_\theta + t_2\partial_\phi$, then $s\cdot t = s_1t_1\sin^2\phi + s_2t_2$. Thus,

$$\partial_{\partial_\theta}(s\cdot t) = \frac{\partial s_1}{\partial\theta}t_1\sin^2\phi + s_1\frac{\partial t_1}{\partial\theta}\sin^2\phi + \frac{\partial s_2}{\partial\theta}t_2 + s_2\frac{\partial t_2}{\partial\theta}.$$

We also compute that

$$(D_{\partial_\theta}s)\cdot t = \Big(\frac{\partial s_1}{\partial\theta} + Bs_1 + Cs_2\Big)t_1\sin^2\phi + \Big(\frac{\partial s_2}{\partial\theta} + Es_1 + Fs_2\Big)t_2$$

and

$$s\cdot(D_{\partial_\theta}t) = s_1\Big(\frac{\partial t_1}{\partial\theta} + Bt_1 + Ct_2\Big)\sin^2\phi + s_2\Big(\frac{\partial t_2}{\partial\theta} + Et_1 + Ft_2\Big).$$

To guarantee (A.1.18) for all choices of s_1, s_2, t_1, t_2, we need $B = F = 0$, and $E = -C\sin^2\phi$.

A similar computation of (A.1.18) for $v = \partial_\phi$ shows that $J = 0$, $I = -H\sin^2\phi$, and $G = \cot\phi$.

So there exists a 2-parameter family of connections that are compatible with the metric. They look like

$$(D_{\partial_\theta}s) = \left(\frac{\partial s_1}{\partial\theta}\partial_\theta + \frac{\partial s_2}{\partial\theta}\partial_\phi\right) + Cs_2\partial_\theta - Cs_1\sin^2\phi\cdot\partial_\phi$$

$$(D_{\partial_\phi}s) = \left(\frac{\partial s_1}{\partial\phi}\partial_\theta + \frac{\partial s_2}{\partial\phi}\partial_\phi\right) + (s_1\cot\phi + Hs_2)\partial_\theta - Hs_1\sin^2\phi\cdot\partial_\phi$$

where C and H are arbitrary smooth functions of θ and ϕ.

A.2. Curvature

Given a connection D on a smooth bundle $E \to B$, one can define an object called the associated *curvature*. This may be described very loosely as follows. Let $b \in B$ and let v and w be two tangent vectors at b. Choose a coordinate system on the manifold where moving each of the first two coordinates gives curves with tangent vectors v and w. Let s be a vector in the fiber E_b. Using the curves of our coordinate system, parallel transport the vector s in a “rectangle”: first in the v-direction, then in the w-direction, then backwards in the w-direction, then backwards in the v-direction. If one does this infinitesimally (whatever that means), one gets a new vector in E_b. It turns out this construction is linear in s, thereby giving a linear transformation $R_{v,w}(b)\colon E_b \to E_b$. It also turns out to be alternating tensorial in v and w—and so gives a 2-form in $\Omega^2(B; \underline{\mathrm{End}}(E))$. This entity is the curvature 2-form corresponding to the bundle with connection (E, D).

There are different approaches to constructing this 2-form rigorously. All of the methods involve cooking up some strange combination of symbols which turns out to be tensorial in all of the variables, but this tensorial nature is always just something that falls out of a computation—it never seems intuitive from the beginning.

A.2.1. Method via the connection matrices. Suppose $U \subseteq B$ is an open set over which E is trivial, and let $s_1, \dots, s_n$ be a basis of local sections for E over U. We have seen that the connection on $E|_U$ is determined by a matrix Ω^s, and that if $e_1, \dots, e_n$ is another basis then the corresponding matrix Ω^e is related by the equation

$$\Omega^e = (d\alpha)\alpha^{-1} + \alpha\Omega^s\alpha^{-1}$$

where $e_i = \sum \alpha_{ij} s_j$.

Consider the expression $K^s = d\Omega^s - \Omega^s \wedge \Omega^s$. First of all, what does this mean? The first term, $d\Omega^s$, can only mean the $n \times n$ matrix of 2-forms obtained by applying d to each entry of Ω^s. The term $\Omega^s \wedge \Omega^s$ must then also be an $n \times n$ matrix of 2-forms, and it seems reasonable to imagine that it is given by $(\Omega^s \wedge \Omega^s)_{ij} = \sum_k \Omega^s_{ik} \wedge \Omega^s_{kj}$.

We need to determine the relation betwen K^e and K^s. We first compute that

$$\begin{aligned} d\Omega^e &= d((d\alpha)\alpha^{-1} + \alpha\Omega^s\alpha^{-1}) \\ &= -(d\alpha)d(\alpha^{-1}) + (d\alpha)\Omega^s\alpha^{-1} + \alpha(d\Omega^s)\alpha^{-1} - \alpha\Omega^s d(\alpha^{-1}). \end{aligned}$$

Applying d to $\alpha \cdot \alpha^{-1} = I$ gives $(d\alpha)\alpha^{-1} + \alpha d(\alpha^{-1}) = 0$, or $d(\alpha^{-1}) = -\alpha^{-1}(d\alpha)\alpha^{-1}$. Using this, we have that

$$\Omega^e = (d\alpha)\alpha^{-1}(d\alpha)\alpha^{-1} + (d\alpha)\Omega^s\alpha^{-1} + \alpha(d\Omega^s)\alpha^{-1} + \alpha\Omega^s\alpha^{-1}(d\alpha)\alpha^{-1}.$$

(If this looks truly awful to you, don't worry—you are not alone, and things will get better soon).

We next compute that

$$\begin{aligned} \Omega^e \wedge \Omega^e &= ((d\alpha)\alpha^{-1} + \alpha\Omega^s\alpha^{-1}) \wedge ((d\alpha)\alpha^{-1} + \alpha\Omega^s\alpha^{-1}) \\ &= (d\alpha)\alpha^{-1}(d\alpha)\alpha^{-1} + (d\alpha)\Omega^s\alpha^{-1} + \alpha\Omega^s\alpha^{-1}(d\alpha)\alpha^{-1} + \alpha(\Omega^s \wedge \Omega^s)\alpha^{-1}. \end{aligned}$$

We find that there is some remarkable cancellation when we form $d\Omega^e - \Omega^e \wedge \Omega^e$, and we get

$$K^e = \alpha\Omega^s\alpha^{-1} - \alpha(\Omega^s \wedge \Omega^s)\alpha^{-1} = \alpha K^s \alpha^{-1}. \tag{A.2.2}$$

So when changing basis from s to e, the entity K^s changes in a very simple way. This is the magic behind curvature!

Now let $\{U_i\}$ be a collection of open sets in B that trivialize the bundle, let $s^i_1, \dots, s^i_n$ be a basis of sections over each U_i, and let Ω^i be the connection matrix for D with respect to this basis. Let $K^i = d\Omega^i - \Omega_i \wedge \Omega_i$, which can either be thought of as a matrix of 2-forms on U_i or as a 2-form on U_i with values in $M_{n\times n}(\mathbb{R})$. The latter is a little better for our present purposes, because what we want to do is to patch these local 2-forms together to make a global 2-form. Of course they *don't* patch together, for then we would have $K_i|_{U_i \cap U_j} = K_j|_{U_i \cap U_j}$ and instead we have the relation (A.2.2). But what (A.2.2) tells us is that if we regard K^i as a 2-form on U_i with values in $\underline{\text{End}}(E)$ (which is the same as $M_{n\times n}(\mathbb{R})$ on U_i), then the forms patch together. So we find the existence of a global 2-form

$$K \in \Omega^2(B; \underline{\text{End}}(E)).$$

This K is called the **curvature 2-form** of the pair (E, D).

Of course this entire discussion begs a question: how did we decide to look at the expression $d\Omega - \Omega \wedge \Omega$ in the first place? I don't have a good answer to this, except to say that if one wrote down all possible ways of using Ω to make a matrix of 2-forms (there aren't many) then one would fairly quickly notice that this particular combination results in lots of cancellation and a simple transformation rule.

REMARK A.2.3. WARNING: In some texts you will see the formula $K = d\Omega + \Omega \wedge \Omega$ instead of the one we used. This seems confusing at first, for how could they both be correct? The reason comes down to a subtle difference in notation. Recall that our convention was

$$Ds_i = \sum \omega_{ij} s_j$$

and that $\Omega = (\omega_{ij})$. One could just as well have decided to write

$$Ds_i = \sum \tilde{\omega}_{ji} s_j$$

and have $\tilde{\Omega} = (\tilde{\omega}_{ji})$. Note that $\omega_{ij} = \tilde{\omega}_{ji}$, or $\Omega = \tilde{\Omega}^T$. Our curvature formula is then

$$\begin{aligned} K_{pq} = d\omega_{pq} - \sum_k \omega_{pk}\omega_{kq} = d\tilde{\omega}_{qp} - \sum_k \tilde{\omega}_{kp}\tilde{\omega}_{qk} &= d\tilde{\omega}_{qp} + \sum_k \tilde{\omega}_{qk}\tilde{\omega}_{kp} \\ &= [d\tilde{\Omega} + \tilde{\Omega} \wedge \tilde{\Omega}]_{qp}. \end{aligned}$$

Note that the sign change is due to the $\tilde{\omega}_{ij}$'s being 1-forms, which therefore anticommute. One could also write this key step as $[\Omega \wedge \Omega]^T = -\Omega^T \wedge \Omega^T$.

If we were adopting the indexing convention of the $\tilde{w}$'s we would define $\tilde{K} = d\tilde{\Omega} + \tilde{\Omega} \wedge \tilde{\Omega}$, and we would have to occasionally remember that this matrix is the transpose of the matrix K that some other people use. Differential geometry is filled with these kinds of differing indexing conventions, often leading to strange sign changes. Once you've come to expect them they are somewhat less frustrating.

REMARK A.2.4. Imagine that we have local coordinates $x^1, \ldots, x^r$ defined on U. Then any 1-form on U may be written as $\sum_i f_i dx^i$ where the f_i are smooth functions. In particular, we may do this simultaneously for every entry in the matrix Ω. What results is an expression

$$\Omega = \Omega_1 dx^1 + \Omega_2 dx^2 + \cdots + \Omega_r dx^r$$

where each Ω_i is an $n \times n$ matrix of smooth functions. One readily computes that

$$d\Omega = \sum_{i,j} (\partial_i \Omega_j) dx^i dx^j = \sum_{i<j} (\partial_i \Omega_j - \partial_j \Omega_i)\, dx^i dx^j$$

where $\partial_i \Omega_j$ represents the matrix obtained by applying $\frac{\partial}{\partial x_i}$ to every entry of Ω_j. Likewise, one finds that

$$\Omega \wedge \Omega = \sum_{i,j} \Omega_i \Omega_j dx^i dx^j = \sum_{i<j} (\Omega_i \Omega_j - \Omega_j \Omega_i)\, dx^i dx^j = \sum_{i<j} [\Omega_i, \Omega_j]\, dx^i dx^j.$$

So

$$K = d\Omega - \Omega \wedge \Omega = \sum_{i<j} \Big[(\partial_i \Omega_j - \partial_j \Omega_i) - [\Omega_i, \Omega_j] \Big]\, dx^i dx^j.$$

For the moment, the above expression for K is just something of a curiosity. Instead of focusing on each 2-form K_{pq}, we are focusing on the $dx^i dx^j$ components of all those 2-forms together: and that is what $(\partial_i \Omega_j - \partial_j \Omega_i) - [\Omega_i, \Omega_j]$ tells us. In essence, we can think of K either as a matrix of 2-forms or as a 2-form of matrices. Part of the challenge in learning this material is learning all the different ways of talking about the same thing!

To close this first introduction to curvature, we give a computation where the expression $d\Omega - \Omega \wedge \Omega$ ends up appearing naturally. Let $\mathfrak{X}$ and $\mathcal{Y}$ be vector fields on B. Start with $D_{\mathcal{Y}} s_i = \sum_j \omega_{ij}(\mathcal{Y}) s_j$. Next calculate that

$$\begin{aligned} D_{\mathfrak{X}} D_{\mathcal{Y}} s_i = \sum_j D_{\mathfrak{X}}(\omega_{ij}(\mathcal{Y}) s_j) &= \sum_j \Big[\big(\partial_{\mathfrak{X}} \omega_{ij}(\mathcal{Y})\big) s_j + \omega_{ij}(\mathcal{Y}) D_{\mathfrak{X}} s_j \Big] \\ &= \sum_j \Big[\big(\partial_{\mathfrak{X}} \omega_{ij}(\mathcal{Y})\big) s_j + \sum_k \omega_{ij}(\mathcal{Y}) \omega_{jk}(\mathfrak{X}) s_k \Big] \\ &= \sum_k \Big[\partial_{\mathfrak{X}} \omega_{ik}(\mathcal{Y}) + \sum_j \omega_{ij}(\mathcal{Y}) \omega_{jk}(\mathfrak{X}) \Big] s_k. \end{aligned}$$

By symmetry one gets a formula for $D_{\mathcal{Y}} D_{\mathfrak{X}} s_i$, and hence

$$\begin{aligned} D_{\mathfrak{X}} D_{\mathcal{Y}} s_i - D_{\mathcal{Y}} D_{\mathfrak{X}} s_i = \sum_k \Big[&(\partial_{\mathfrak{X}} \omega_{ik}(\mathcal{Y}) - \partial_{\mathcal{Y}} \omega_{ik}(\mathfrak{X})) + \\ &\sum_j (\omega_{ij}(\mathcal{Y}) \omega_{jk}(\mathfrak{X}) - \omega_{ij}(\mathfrak{X}) \omega_{jk}(\mathcal{Y})) \Big] s_k. \end{aligned}$$

Now recall that $(\alpha \wedge \beta)(u, v) = \alpha(u)\beta(v) - \alpha(v)\beta(u)$. So the term inside the second sum is $-(\omega_{ij} \wedge \omega_{jk})(\mathfrak{X}, \mathcal{Y})$. Provided that $\mathfrak{X}$ and $\mathcal{Y}$ are vector fields associated to a coordinate system $x_1, \ldots, x_r$ (that is to say, $\mathfrak{X} = \partial_{x_u}$ and $\mathfrak{X} = \partial_{x_v}$ for some u and v), then $\partial_{\mathfrak{X}} \omega_{ik}(\mathcal{Y}) - \partial_{\mathcal{Y}} \omega_{ik}(\mathfrak{X})$ is precisely $(d\omega_{ik})(\mathfrak{X}, \mathcal{Y})$ (this is a nice exercise, or else see Exercise A.2.6(b) below for some hints). So we have

$$\begin{aligned} D_{\mathfrak{X}} D_{\mathcal{Y}} s_i - D_{\mathcal{Y}} D_{\mathfrak{X}} s_i &= \sum_k \Big[(d\omega_{ik})(\mathfrak{X}, \mathcal{Y}) - (\Omega \wedge \Omega)_{ik}(\mathfrak{X}, \mathcal{Y}) \Big] s_k \\ &= \sum_k (d\Omega - \Omega \wedge \Omega)(\mathfrak{X}, \mathcal{Y})_{ik} s_k. \end{aligned}$$

This computation illustrates two important things. First, it shows us that the matrix of 2-forms $d\Omega - \Omega \wedge \Omega$ has something to do with differentiating in two different directions, and how that fails to commute. Secondly, it is telling us that the matrix $(d\Omega - \Omega \wedge \Omega)(\mathfrak{X}, \mathcal{Y})$ should really be interpreted as an *endormorphism*: it is the matrix for the transformation $D_{\mathfrak{X}} D_{\mathcal{Y}} - D_{\mathcal{Y}} D_{\mathfrak{X}}$. The next section takes up this discussion in much more detail.

A.2.5. Constructing curvature in terms of iterated D operators. Let B be a smooth manifold. A **derivation** on $C^\infty(B)$ is an $\mathbb{R}$-linear map $D \colon C^\infty(B) \to \mathbb{R}$ such that $D(fg) = (Df)g + f(Dg)$. A simple computation shows that if D and D' are derivations then the commutator $[D, D'] = DD' - D'D$ is another derivation.

There is a map

$$\Gamma(TB) \to (\text{Derivations on } C^\infty(B)), \qquad \mathfrak{X} \to \partial_{\mathfrak{X}},$$

and this turns out to be an isomorphism. So if $\mathfrak{X}$ and $\mathcal{Y}$ are vector fields on B, there is a unique vector field which is the preimage of $[\partial_{\mathfrak{X}}, \partial_{\mathfrak{X}}]$. We denote this vector field by $[\mathfrak{X}, \mathcal{Y}]$. Note that if $f \in C^\infty(B)$ then

$$[f\mathfrak{X}, \mathcal{Y}] = f[\mathfrak{X}, \mathcal{Y}] - (\partial_{\mathcal{Y}} f)\mathfrak{X}.$$

This is a simple computation with differential operators:

$$\partial_{f\mathfrak{X}}\partial_{\mathcal{Y}}g - \partial_{\mathcal{Y}}\partial_{f\mathfrak{X}}g = f\cdot\partial_{\mathfrak{X}}\partial_{\mathcal{Y}}g - \partial_{\mathcal{Y}}(f\cdot\partial_{\mathfrak{X}}g) = f\cdot(\partial_{\mathfrak{X}}\partial_{\mathcal{Y}} - \partial_{\mathcal{Y}}\partial_{\mathfrak{X}})g - (\partial_{\mathcal{Y}}f)\partial_{\mathfrak{X}}g.$$

One important instance of this bracket construction is as follows. Suppose that $x^1,\dots,x^r$ are local coordinates on B, and let $\partial_1,\dots,\partial_r$ be the associated vector fields. Then one readily checks that $[\partial_i,\partial_j]=0$ for all i,j (because partial derivatives commute in $\mathbb{R}^n$).

Exercise A.2.6. Let $x^1,\dots,x^r$ and $\partial_1,\dots,\partial_r$ be as in the preceding paragraph. If $\mathfrak{U}$ is a vector field on B, write $\mathfrak{U}=\sum u^i\partial_i$.

(a) Check that $[\mathfrak{U},\mathcal{V}]^i = \partial_{\mathfrak{U}}v^i - \partial_{\mathcal{V}}u^i$. In other words, $[\mathfrak{U},\mathcal{V}] = \sum_i(\partial_{\mathfrak{U}}v^i - \partial_{\mathcal{V}}u^i)\partial_i$.

(b) Let ω be a 1-form on B. Prove that

$$(d\omega)(\mathfrak{U},\mathcal{V}) = \partial_{\mathfrak{U}}\omega(\mathcal{V}) - \partial_{\mathcal{V}}\omega(\mathfrak{U}) - \omega([\mathfrak{U},\mathcal{V}]).$$

(Hint: Write $\omega=\sum f_i dx^i$, so that $\omega(\mathcal{V})=\sum f_i v^i$. Now just compute everything by brute force.)

Let $E\to B$ be a smooth vector bundle with an affine connection D. We consider the map

$$R\colon \Gamma(TB)\otimes\Gamma(TB)\otimes\Gamma(E)\to\Gamma(E)$$

defined via the formula

$$\mathfrak{X}\otimes\mathcal{Y}\otimes s\mapsto D_{\mathfrak{X}}D_{\mathcal{Y}}s - D_{\mathcal{Y}}D_{\mathfrak{X}}s - D_{[\mathfrak{X},\mathcal{Y}]}s = R_{\mathfrak{X},\mathcal{Y}}(s).$$

Once again this is a formula that I am pulling out of my hat, without much motivation. I don't know any way of thinking about this formula that immediately calls out "Yes, this is the right thing to be looking at!" But there are a few things I can say. First, we saw at the end of the last section that curvature should have something to do with the commutator $D_{\mathfrak{X}}D_{\mathcal{Y}} - D_{\mathcal{Y}}D_{\mathfrak{X}}$. However, if our formula only used this commutator then it would not be $C^\infty(B)$-linear in the variables $\mathfrak{X}$ and $\mathcal{Y}$ (this is an easy check)—in other words, it would not describe a tensor. Upon realizing this, one might start looking around for terms to add to the formula that guarantee the $C^\infty(B)$-linearity—and eventually one would stumble upon the $D_{[\mathfrak{X},\mathcal{Y}]}$-term. But secondly, one could almost guess this: looking at the computation from the end of the last section, at a crucial stage we used a formula for $(d\omega)(\mathfrak{X},\mathcal{Y})$ which we explained did not always hold. The version of that formula that *does* always hold is Exercise A.2.6 above, and the extra term needed precisely comes from $[\mathfrak{X},\mathcal{Y}]$. Finally, it is worth observing that when $[\mathfrak{X},\mathcal{Y}]=0$ (e.g., when $\mathfrak{X}$ and $\mathcal{Y}$ are the coordinate vector fields from a Euclidean neighborhood) then the formula for R really *is* just the commutator that came up in the last section.

Setting aside the question of motivation, we claim the amazing thing about the formula defining R is that it is $C^\infty(B)$-linear in all three variables. We will mostly leave the verifcation to the reader, but here is the first part:

$$\begin{aligned} R_{f\mathfrak{X},\mathcal{Y}}(s) &= D_{f\mathfrak{X}}D_{\mathcal{Y}}s - D_{\mathcal{Y}}D_{f\mathfrak{X}}s - D_{[f\mathfrak{X},\mathcal{Y}]}s \\ &= fD_{\mathfrak{X}}D_{\mathcal{Y}}s - (\partial_{\mathcal{Y}}f)D_{\mathfrak{X}}s - fD_{\mathcal{Y}}D_{\mathfrak{X}}s - fD_{[\mathfrak{X},\mathcal{Y}]}s + (\partial_{\mathcal{Y}}f)D_{\mathfrak{X}}s \\ &= fD_{\mathfrak{X}}D_{\mathcal{Y}}s - fD_{\mathcal{Y}}D_{\mathfrak{X}}s - fD_{[\mathfrak{X},\mathcal{Y}]}s \\ &= fR_{\mathfrak{X},\mathcal{Y}}s. \end{aligned}$$

Since R is $C^\infty(B)$-linear in each variable, it comes from an associated map of bundles $R\colon TB\otimes TB\otimes E\to E$. By adjointness this is a map $TB\otimes TB\to\underline{\mathrm{End}}(E)$, and one readily checks that it factors through $\bigwedge^2(TB)$. So we may interpret R as a bundle map $\bigwedge^2(TB) \to \underline{\mathrm{End}}(E)$, or equivalently

$$\begin{aligned} R \in \Gamma(\underline{\mathrm{Hom}}(\textstyle\bigwedge^2(TB), \underline{\mathrm{End}}(E))) &\cong \Gamma(\underline{\mathrm{Hom}}(\textstyle\bigwedge^2(TB), \mathbb{R}) \otimes \underline{\mathrm{End}}(E)) \\ &= \Omega^2(B; \underline{\mathrm{End}}(E)). \end{aligned}$$

So we regard R as a 2-form on B with coefficients in $\underline{\mathrm{End}}(E)$.

As always with defining differential forms, it can be a little obtuse trying to figure out what this is really saying. Keep in mind that for every $b \in B$ and tangent vectors $v, w \in T_bB$, then $R_{v,w}(b)$ is a linear map $E_b \to E_b$. And this is alternating in the tangent vectors, so that $R_{v,w}(b) = -R_{w,v}(b)$.

Let $s_1, \dots, s_n$ be a local basis of sections for E. Since $R_{v,w}$ is a linear transformation, we can represent it as a matrix

$$R_{v,w}(s_i) = \sum_k [R_{v,w}]_{k,i} s_k. \tag{A.2.7}$$

(The choice of indexing comes from the convention of having matrices act on the left). We will move symbols around from time to time, e.g. writing $R_{k,i}(v, w)$ for $[R_{v,w}]_{k,i}$.

Given any independent tangent vectors v and w at b, we can find a coordinate system $x^1, \dots, x^r$ where $\partial_1(b) = v$ and $\partial_2(b) = w$. Taking $\mathfrak{X} = \partial_1$ and $\mathcal{Y} = \partial_2$, the computation at the end of the last section shows that we have the equality $R_{ki}(\mathfrak{X}, \mathcal{Y}) = (d\Omega - \Omega \wedge \Omega)_{ik}(\mathfrak{X}, \mathcal{Y})$. That is, $R = (d\Omega - \Omega \wedge \Omega)^T$. (Note that we could have avoided the transpose by adopting the opposite indexing convention in (A.2.7)).

A.2.8. Expressing curvature in terms of a connection potential. Suppose given a local basis of sections $e_1, \dots, e_n$ for E. Let $x^1, \dots, x^r$ be local coordinates on B, and $\partial_1, \dots, \partial_r$ the corresponding vector fields. In the following computation we will write i, j, k for elements of $\{1, \dots, r\}$ and α, β, γ for elements of $\{1, \dots, n\}$. Given a local section s of E, we will write $s(x) = \sum_\alpha s^\alpha(x) \cdot e_\alpha$.

Let A denote a connection potential, and let $D = D^0 + A$ where D^0 is the standard flat connecion with respect to the basis $e_1, \dots, e_n$ (so really D is just a connection on $E|_U$ where U is the open set on which our basis is defined). So

$$D_v s = \sum_\alpha (\partial_v s^\alpha) \cdot e_\alpha + A_v(s). \tag{A.2.9}$$

It will be convenient to write

$$A_v(e_\alpha) = \sum_\beta A^\beta_{v,\alpha} e_\beta, \tag{A.2.10}$$

and so $A_v(s) = \sum_{\alpha,\beta}(A^\beta_{v,\alpha} s^\alpha) e_\beta$. Note that with this notation the matrix representing A_v, with respect to the basis $e_1, \dots, e_n$, is the matrix $(A_v)_{\beta,\alpha} = A^\beta_{v,\alpha}$ (where matrices act on the left). In other words, β corresponds to the row index and α to the column. Below we will need a consequence of this, that the matrix for the composition $A_w \circ A_v$ has (β, α)-entry equal to $A^\beta_{w,\gamma} A^\gamma_{v,\alpha}$ (using Einstein summation convention, as always).

Write $D_i s$ as an abbreviation for $D_{\partial i} s$. From (A.2.9) and (A.2.10) it follows that

$$D_i s = \sum_\alpha (\partial_i s^\alpha) \cdot e_\alpha + \sum_{\alpha,\beta} A^\beta_{i,\alpha} s^\alpha e_\beta$$

and so in particular

$$D_i e_\gamma = \sum_\beta A^\beta_{i,\gamma} e_\beta.$$

Based on these formulas we can now calculate

$$\begin{aligned}
R_{ij}(e_\beta) &= D_i D_j e_\beta - D_j D_i e_\beta - D_{[\partial_i,\partial_j]} e_\beta \\
&= D_i(A^\gamma_{j\beta} e_\gamma) - D_j(A^\gamma_{i\beta} e_\gamma) \\
&= \partial_i(A^\gamma_{j\beta})e_\gamma + A^\gamma_{j\beta}A^\delta_{i\gamma}e_\delta - (\partial_j A^\gamma_{i\beta})e_\gamma - A^\gamma_{i\beta}A^\delta_{j\gamma}e_\delta \\
&= \Big(\partial_i A^\gamma_{j\beta} - \partial_j A^\gamma_{i\beta} + A^\delta_{j\beta}A^\gamma_{i\delta} - A^\delta_{i\beta}A^\gamma_{j\delta}\Big)e_\gamma \\
&= \Big(\partial_i A^\gamma_{j\beta} - \partial_j A^\gamma_{i\beta} + A^\gamma_{i\delta}A^\delta_{j\beta} - A^\gamma_{j\delta}A^\delta_{i\beta}\Big)e_\gamma \\
&= \Big(\partial_i A^\gamma_{j\beta} - \partial_j A^\gamma_{i\beta} + (A_iA_j)^\gamma_\beta - (A_jA_i)^\gamma_\beta\Big)e_\gamma.
\end{aligned}$$

It follows at once that

$$R_{ij} = \partial_i A_j - \partial_j A_i + [A_i, A_j]. \tag{A.2.11}$$

Recall what everything means here. Each A_i is an $n \times n$ matrix of smooth functions, and so all the terms on the right are also $n \times n$ matrices of smooth functions. So equation (A.2.11) describes a 2-form on our open set with values in $n \times n$ matrices (such matrices are the local form of $\underline{\mathrm{End}}(E)$). We may also write

$$R = \sum_{i<j} (\partial_i A_j - \partial_j A_i + [A_i, A_j])\, dx^i dx^j.$$

Remark A.2.12. The reader might notice that we have seen this calculation before, back in Remark A.2.4. However, the end result there was a similar-looking formula involving Ω's and a *sign change* on the commutator term. As in Remark A.2.3, this is due to a difference in indexing conventions. Whereas we had $De_i = \sum \omega_{ij} e_j$, we also wrote $De_i = \sum A^\beta_\gamma e_\gamma$. At first this change seems very innocuous, but when we formed the product $A_i A_j$ it was key that we were regarding these matrices as having the β's represent the row index and the γ's represent the column index. This shows that the ω-summation and the A-summation are using the opposite indexing conventions, and we are in the domain of Remark A.2.3.

Example A.2.13. Let $E = \mathbb{R}^4 \times \mathbb{C} \to \mathbb{R}^4$ be the trivial complex line bundle over $\mathbb{R}^4$, and let D be the $\mathfrak{u}(1)$-connection on E determined by the vector potential $A = iA_1 + iA_2 + iA_3 + iA_4$???? Note that $[A_i, A_j] = 0$ for all i, j (the Lie algebra $\mathfrak{u}(1)$ is actually commutative), and so $R_{jk} = \partial_j A_k - \partial_k A_j$. This precisely says that the curvature tensor is

$$\mathcal{R} = d(A_1 dx + A_2 dy + A_3 dz + A_4 dt).$$

But the entity on the right is what we usually think of as the electromagnetic 2-form $\mathcal{F}$ corresponding to the magnetic potential A. We will come back to this example in some detail in Section 6.3.

A.2.14. A Sample Curvature Computation. In this section, we will use the theory we've developed to actually compute the curvature in a familiar concrete example. We will consider the two-dimensional sphere with radius R. Let's use the cylindrical coordinates r and θ, with translation back to Euclidian coordinates in the ambient space given by:

$$\begin{aligned} x &= r\cos\theta \\ y &= r\sin\theta \\ z &= \sqrt{R^2 - r^2}. \end{aligned}$$

At any point (in the coordinate patch) the tangent space is spanned by the tangent vectors ∂_r and ∂_θ. In the coordinates of the ambient space, we can compute

$$\begin{aligned} \partial_r &= \left[\cos\theta, \sin\theta, -\tfrac{r}{z}\right] \\ \partial_\theta &= [-r\sin\theta, r\cos\theta, 0]. \end{aligned}$$

Note that we are mixing cylindrical coordinates and Euclidean coordinates whenever convenient.

Finally, we record the inner products of these tangent vectors, which encodes the metric on the sphere induced by its embedding in Euclidian space:

$$\langle \partial_r, \partial_r \rangle = \frac{R^2}{z^2} \qquad \langle \partial_r, \partial_\theta \rangle = 0 \qquad \langle \partial_\theta, \partial_r \rangle = 0 \qquad \langle \partial_\theta, \partial_\theta \rangle = r^2.$$

So now we have a manifold S^2 with a metric. The next piece of data we need to talk about curvature is a connection on the tangent bundle. Using our basis, we can specify this by writing

$$\begin{aligned} \mathcal{D}_v(\partial_r) &= A_{11}(v)\partial_r + A_{21}(v)\partial_\theta \\ \mathcal{D}_v(\partial_\theta) &= A_{12}(v)\partial_r + A_{22}(v)\partial_\theta, \end{aligned}$$

where the A_{ij} are 1-forms, and together we can think of A as a 1-form with values in the endomorphism bundle of TS^2. The defining properties of an affine connection say that, for an arbitrary vector field $s_r\partial_r + s_\theta\partial_\theta$,

$$\begin{aligned} \mathcal{D}_v(s_r\partial_r + s_\theta\partial_\theta) &= \partial_v(s_r)\partial_r + s_r\mathcal{D}_v(\partial_r) + \partial_v(s_\theta)\partial_\theta + s_\theta\mathcal{D}_v(\partial_\theta) \\ &= \big(\partial_v(s_r)\partial_r + \partial_v(s_\theta)\partial_\theta\big) + \big(s_r\mathcal{D}_v(\partial_r) + s_\theta\mathcal{D}_v(\partial_\theta)\big) \\ &= \mathcal{D}^0_v(s_r\partial_r + s_\theta\partial_\theta) + A(v)\cdot\begin{bmatrix} s_r \\ s_\theta \end{bmatrix} \end{aligned}$$

where $\mathcal{D}^0$ is the standard flat connection in these coordinates, and the $\underline{\mathrm{End}}(TS^2)$-valued 1-form A is what we're calling the "connection potential".

Any choice of connection potential A gives us a connection, but we want one that is compatible with the metric if the curvature of the connection is going to tell us anything about the geometric curvature. Recall that this means

$$\partial_v\langle X, Y\rangle = \langle \mathcal{D}_v X, Y\rangle + \langle X, \mathcal{D}_v Y\rangle,$$

for all vector fields X and Y. By replacing X and Y with the vector fields ∂_r and ∂_θ, and replacing v with the tangent vectors ∂_r and ∂_θ, one gets six equations involving the A_{jk}'s (it seems like it should be eight equations, but there are duplicates because

$\langle \partial_r, \partial_\theta \rangle = \langle \partial_\theta, \partial_r \rangle$). Let's compute these:

$$\begin{aligned} \partial_v \langle \partial_r, \partial_r \rangle &= \langle \mathcal{D}_v \partial_r, \partial_r \rangle + \langle \partial_r, \mathcal{D}_v \partial_r \rangle \\ &= 2\langle \mathcal{D}_v \partial_r, \partial r \rangle \\ &= 2A_{11}(v) \langle \partial_r, \partial_r \rangle \\ &= 2A_{11}(v) \frac{R^2}{z^2}. \end{aligned}$$

Taking $v = \partial_r$, this says

$$\frac{\partial}{\partial r}\left(\frac{R^2}{R^2 - r^2}\right) = 2A_{11}(\partial_r) \frac{R^2}{R^2 - r^2},$$

and solving for $A_{11}(\partial_r)$ we find

$$A_{11}(\partial_r) = \frac{r}{z^2}.$$

Taking $v = \partial_\theta$ we easily find $A_{11}(\partial_\theta) = 0$. Putting these together, we have found the differential form

$$A_{11} = \frac{r}{z^2}\, dr.$$

If we do the same thing for $\partial_v \langle \partial_\theta, \partial_\theta \rangle$, we end up finding

$$A_{22} = \frac{1}{r}\, dr.$$

And finally, computing $\partial_v \langle \partial_r, \partial_\theta \rangle$ we get the condition

$$A_{12} = -\frac{r^2 z^2}{R^2} A_{21}.$$

So the connection is not yet completely determined; *many* connections are compatible with the metric.

Differential geometers like to impose an extra condition on the connection beyond metric compatibility, and together these two conditions uniquely determine the connection: this is the so-called "Levi-Civita connection". The extra condition is that the connection be **torsion-free**, meaning that

$$\mathcal{D}_X Y - \mathcal{D}_Y X = [X, Y]$$

for any vector fields X and Y, where $[X, Y]$ is the Lie bracket.

So to proceed with our analysis, let us use the torsion-free condition. This gives us the equation

$$0 = \mathcal{D}_{\partial_r}(\partial_\theta) - \mathcal{D}_{\partial_\theta}(\partial_r),$$

which forces

$$A_{21} = \frac{1}{r}\, d\theta, \qquad A_{12} = -\frac{rz^2}{R^2}\, d\theta.$$

We have now completely determined the Levi-Civita connection, and we can succinctly write the connection potential as

$$A = \begin{bmatrix} \frac{r}{z^2} & 0 \\ 0 & \frac{1}{r} \end{bmatrix} dr + \begin{bmatrix} 0 & -\frac{rz^2}{R^2} \\ \frac{1}{r} & 0 \end{bmatrix} d\theta.$$

We would now like to compute the curvature of this connection. The curvature $R \in \Omega^2(B, \underline{\mathrm{End}}(TS^2))$ is an $\underline{\mathrm{End}}(TS^2)$-valued 2-form. As a 2-form, R is alternating, so all we really need to compute is $R_{\partial_r,\partial_\theta}$. Earlier we found an equation for the curvature of a connection in terms of the connection potential:

$$R_{i,j} = \partial_i A_j - \partial_j A_i + [A_i, A_j].$$

Using this we get

$$\begin{aligned} R_{\partial_r,\partial_\theta} &= \frac{\partial}{\partial r}\begin{bmatrix} 0 & -\frac{rz^2}{R^2} \\ \frac{1}{r} & 0 \end{bmatrix} - \frac{\partial}{\partial \theta}\begin{bmatrix} \frac{r}{z^2} & 0 \\ 0 & \frac{1}{r} \end{bmatrix} + \left[\begin{bmatrix} \frac{r}{z^2} & 0 \\ 0 & \frac{1}{r} \end{bmatrix}, \begin{bmatrix} 0 & -\frac{rz^2}{R^2} \\ \frac{1}{r} & 0 \end{bmatrix}\right] \\ &= \begin{bmatrix} 0 & \frac{r^2}{R^2} \\ -\frac{1}{z^2} & 0 \end{bmatrix}. \end{aligned}$$

In other words,

$$R = \begin{bmatrix} 0 & \frac{r^2}{R^2} \\ -\frac{1}{z^2} & 0 \end{bmatrix} dr \wedge d\theta \in \Omega^2(S^2, \underline{\mathrm{End}}(TS^2)).$$

We have now computed the curvature 2-form for the Levi-Civita connection arising from the metric induced on the sphere by its embedding into 3-dimensional Euclidian space. It's worth emphasizing that this curvature depends only upon the metric, since the Levi-Civita connection is uniquely determined by the metric. So what is R telling us about the geometry of the sphere? To see this, we will need one more theorem from differential geometry, which says that the quantity

$$K(v,w) = \frac{\langle R_{v,w}(w), v\rangle}{|v \wedge w|^2}$$

is the sectional curvature in the plane $\mathrm{Span}\{v, w\}$. Applying this to our example, and recalling that

$$\langle \partial_r \wedge \partial_\theta, \partial_r \wedge \partial_\theta\rangle = \det \begin{bmatrix} \langle \partial_r, \partial_r\rangle & \langle \partial_r, \partial_\theta\rangle \\ \langle \partial_\theta, \partial_r\rangle & \langle \partial_\theta, \partial_\theta\rangle \end{bmatrix},$$

we find

$$K(\partial_r, \partial_\theta) = \frac{\langle R_{\partial_r,\partial_\theta}(\partial_\theta), \partial_r\rangle}{|\partial_r \wedge \partial_\theta|^2} = \frac{\frac{r^2}{R^2}\langle \partial_r, \partial_r\rangle}{\langle \partial_r, \partial_r\rangle\langle \partial_\theta, \partial_\theta\rangle} = \frac{\frac{r^2}{R^2}}{r^2} = \frac{1}{R^2}.$$

So the sectional curvature is constant, and depends only on the radius of the sphere—as one would expect.

Example A.2.15. We will re-do the above example in two other ways. Set $e_1 = \partial_r$ and $e_2 = \partial_\theta$. Recall that the matrix Ω for the connection is defined by

$$D_v(e_i) = \sum_j \omega_{ij} e_j.$$

Comparing to the above example, we find that

$$\Omega = \begin{bmatrix} \frac{r}{z^2}\, dr & \frac{1}{r}\, d\theta \\ -\frac{rz^2}{R^2}\, d\theta & \frac{1}{r}\, dr \end{bmatrix}.$$

Then we compute that

$$d\Omega = \begin{bmatrix} 0 & -\frac{1}{r^2}\, dr d\theta \\ (-1 + \frac{3r^2}{R^2})\, dr d\theta & 0 \end{bmatrix}$$

and

$$\Omega \wedge \Omega = \begin{bmatrix} 0 & (\frac{1}{z^2} - \frac{1}{r^2})\, dr d\theta \\ (\frac{2r^2 - R^2}{R^2})\, dr d\theta & 0 \end{bmatrix}.$$

Next we form

$$K = d\Omega - \Omega \wedge \Omega = \begin{bmatrix} 0 & -\frac{1}{z^2}\, dr d\theta \\ \frac{r^2}{R^2}\, dr d\theta & 0 \end{bmatrix}.$$

Finally, we recall from the discussion after (A.2.7) that $d\Omega - \Omega \wedge \Omega$ is the transpose of the matrix representing R. This recovers the R-matrix from the previous example, and then the sectional curvature computation proceeds as before.

Now let us re-do everything using a different coordinate system on S^2. We could use the coordinates x and y, but here ∂_x and ∂_y turn out not to be orthogonal—this doesn't cause any problem with our methods, but it does make all the calculations more cumbersome. Instead let us use spherical coordinates θ and ϕ. One finds that

$$\partial_\theta \cdot \partial_\theta = R^2 \sin^2 \phi, \quad \partial_\theta \cdot \partial_\phi = 0, \quad \partial_\phi \cdot \partial_\phi = R^2.$$

The connection is given by

$$D_v(\partial_\theta) = A_{11}(v)\partial_\theta + A_{21}(v)\partial_\phi$$
$$D_v(\partial_\phi) = A_{12}(v)\partial_\theta + A_{22}(v)\partial_\phi$$

and the Ω-matrix is

$$\Omega = A^T = \begin{bmatrix} A_{11} & A_{21} \\ A_{12} & A_{22} \end{bmatrix}.$$

We start with the metric compatibility equation

$$\partial_v \langle \partial_\theta, \partial_\theta \rangle = 2 \langle D_v \partial_\theta, \partial_\theta \rangle,$$

and upon computing both sides for $v = \partial_\theta$ and $v = \partial_\phi$ we arrive at $A_{11} = (\cot \phi) d\phi$. A similar analysis for $\partial_v \langle \partial_\phi, \partial_\phi \rangle$ yields that $A_{22} = 0$, and the analysis for $\partial_v \langle \partial_\theta, \partial_\phi \rangle$ yields that $A_{21} = -A_{12} \sin^2 \phi$. Finally we write down the equation for the torsion-free condition:

$$0 = D_{\partial_\theta} \partial_\phi - D_{\partial_\phi} \partial_\theta = A_{12}(\partial_\theta)\partial_\theta - [A_{11}(\partial_\phi)\partial_\theta + A_{21}(\partial_\phi)\partial_\phi].$$

It follows that $A_{21}(\partial_\phi) = 0$ (and hence $A_{12}(\partial_\phi) = 0$ as well), and that $A_{12}(\partial_\theta) = A_{11}(\partial_\phi) = \cot \phi$. So $A_{12} = (\cot \phi)\, \partial_\theta$, and we have found that

$$\Omega = \begin{bmatrix} (\cot \phi)\, d\phi & -(\sin \phi \cos \phi)\, d\theta \\ (\cot \phi)\, d\theta & 0 \end{bmatrix}.$$

Next we compute

$$d\Omega = \begin{bmatrix} 0 & \cos(2\phi)\, d\theta d\phi \\ \frac{1}{\sin^2 \phi}\, d\theta\, d\phi & 0 \end{bmatrix}$$

and

$$\Omega \wedge \Omega = \begin{bmatrix} 0 & (\cos^2 \phi)\, d\theta\, d\phi \\ (\cot^2 \phi)\, d\theta\, d\phi & 0 \end{bmatrix}.$$

Then

$$d\Omega - \Omega \wedge \Omega = \begin{bmatrix} 0 & -(\sin^2 \phi)\, d\theta d\phi \\ d\theta d\phi & 0. \end{bmatrix}$$

Again, recall that this is the transpose of the curvature matrix and so

$$R = \begin{bmatrix} 0 & d\theta d\phi \\ -(\sin^2\phi)\, d\theta d\phi & 0. \end{bmatrix}.$$

The sectional curvature is

$$\frac{\langle R_{\partial_\theta,\partial_\phi}(\partial_\phi), \partial_\theta\rangle}{\langle \partial_\theta, \partial_\theta\rangle\langle \partial_\phi, \partial_\phi\rangle} = \frac{\langle \partial_\theta, \partial_\theta\rangle}{\langle \partial_\theta, \partial_\theta\rangle\langle \partial_\phi, \partial_\phi\rangle} = \frac{1}{\langle \partial_\phi, \partial_\phi\rangle} = \frac{1}{R^2}.$$

A.3. DeRham theory with coefficients in a bundle

Let $E \to B$ be a smooth vector bundle equipped with an affine connection D. Given a section s of E and a tangent vector v, the connection gives us a derivatve $D_v s$. We can phrase this differentiation process as a map

$$d_D \colon \Gamma(E) \longrightarrow \Gamma(T^*B \otimes E).$$

We with to extend this via a type of Leibniz rule to get a sequence of maps

$$\Gamma(E) \xrightarrow{d_D} \Gamma(T^*B \otimes E) \xrightarrow{d_D} \Gamma(\textstyle\bigwedge^2 T^*B \otimes E) \xrightarrow{d_D} \Gamma(\textstyle\bigwedge^3 T^*B \otimes E) \xrightarrow{d_D} \cdots$$

It will not be the case that $d_D^2 = 0$, however; instead we will find that d_D^2 measures the curvature of E.

A.3.1. Wedge product of forms.

Recall that we write $\Omega^p(B; E)$ for $\Gamma(\bigwedge^p T^*B \otimes E)$, and elements of this space are called differential p-forms with coefficients in E. Such differential forms are spanned by elements $\alpha \otimes s$ where $\alpha \in \Omega^p(B)$ and $s \in \Gamma(E)$.

Assuming we have two bundles $\mathcal{E}$ and $\mathcal{F}$, there are natural maps

$$\Omega^p(B; \mathcal{E}) \otimes \Omega^q(B; \mathcal{F}) \longrightarrow \Omega^{p+q}(B; \mathcal{E} \otimes \mathcal{F})$$

which are uniquely determined by saying that

$$(\alpha \otimes s) \otimes (\beta \otimes t) \mapsto (\alpha \wedge \beta) \otimes (s \otimes t).$$

If we have a third vector bundle $\mathcal{G}$ and a pairing $\mu \colon \mathcal{E} \otimes \mathcal{F} \to \mathcal{G}$ then we can compose with this pairing to get

$$\Omega^p(B; \mathcal{E}) \otimes \Omega^q(B; \mathcal{F}) \longrightarrow \Omega^{p+q}(B; \mathcal{E} \otimes \mathcal{F}) \longrightarrow \Omega^{p+q}(B; \mathcal{G}).$$

We will call this composite map the wedge product, usually denoting it $\wedge$ but sometimes by $\wedge_\mu$ when we need to be specific.

Example A.3.2. Some important examples of the above setup are as follows:

(a) $\mathcal{E} = \underline{\mathrm{Hom}}(E, E)$, $\mathcal{F} = \mathcal{G} = E$, and $\mathcal{E} \otimes \mathcal{F} \to \mathcal{G}$ is the natural action of $\underline{\mathrm{Hom}}(E, E)$ on E.

(b) $\mathcal{E} = \mathcal{F} = \mathcal{G} = \underline{\mathrm{Hom}}(E, E)$ and $\mathcal{E} \otimes \mathcal{F} \to \mathcal{G}$ is composition. In this example, suppose that $A \in \Omega^1(B; \underline{\mathrm{End}}(E))$ is a connection potential.. In local coordinates on B we may write $A = \sum A_j dx^j$ where A_j is a local section of $\underline{\mathrm{End}}(E)$. Then

$$A \wedge A = \sum_{j,k} A_j A_k dx^j \wedge dx^k = \sum_{j \le k} (A_j A_k - A_k A_j) dx^j \wedge dx^k \tag{A.3.3}$$

$$= \sum_{j<k} [A_j, A_k] dx^j \wedge dx^k.$$

(c) Suppose that $\mathcal{E}$ has a Lie bracket $[-,-]\colon \mathcal{E}\otimes\mathcal{E}\to\mathcal{E}$. The main examples for us will be when $\mathcal{E}$ is the trivial bundle associated to a Lie algebra $\mathfrak{g}$, or when $\mathcal{E}=\underline{\mathrm{End}}(E)$ and the bracket is the commutator. Using the above construction, the bracket gives rise to pairings $\Omega^p(B;\mathcal{E})\otimes\Omega^q(B;\mathcal{E})\to\Omega^{p+q}(B;\mathcal{E})$. Our custom would be to denote this by $\wedge$ or $\wedge_{[-,-]}$, but here it will be convenient to drop the wedge and just write is as $[-,-]$.

Let $A,C\in\Omega^1(B;\mathcal{E})$ and write $A=\sum_i A_i dx^i$ and $C=\sum_j C_j dx^j$ where $x^1,\dots,x^n$ is some local coordinate system on B. Then

$$[A,C]=\sum_{i,j}[A_i,C_j]dx^i dx^j=\sum_{i<j}\big([A_i,C_j]-[A_j,C_i]\big)dx^i dx^j.$$

As a consequence, note that

$$[A,A]=\sum_{i<j}\big([A_i,A_j]-[A_j,A_i]\big)dx^i dx^j=2\sum_{i<j}[A_i,A_j]dx^i dx^j.$$

When $\mathcal{E}=\underline{\mathrm{End}}(E)$, observe that $\frac{1}{2}[A,A]=A\wedge A$. This is another source of multiple notations in differential geometry. For instance, if D^0 is a flat connection on a bundle $E\to B$ and A is a connection potential, then the curvature of the connection $D=D^0+A$ may be written as

$$R=dA+A\wedge A \qquad\text{or}\qquad R=dA+\tfrac{1}{2}[A,A]$$

and the two formulas say exactly the same thing.

(d) As our final example, let $\mathcal{E}=\mathcal{F}=\mathcal{G}=M_{n\times n}(\mathbb{R})$ be the trivial bundle of $n\times n$ matrices, and let $\mathcal{E}\otimes\mathcal{F}\to\mathcal{G}$ be matrix multiplication. An element of $\Omega^p(B;\mathcal{E})$ may be thought of either as a matrix-valued p-form or as an $n\times n$ matrix of real-valued p-forms. Let us use Greek letters $\alpha,\beta,\dots$ to denote the former, and the corresponding capital Roman letters $A,B,\dots$ to denote the latter. So α and A are in some sense the same object, just thought of in different ways.

We claim that the product $\alpha\wedge\beta$ corresponds to the ordinary matrix product AB. This is really an "obvious" fact, although in my opinion it's the kind of thing that is only obvious after one sees the somewhat-wordy explanation. In any case, here is the explanation. By linearity in all the variables we might as well assume that A has a single nonzero entry—in spot (i,j), say—and that B also has a single nonzero entry—in spot (r,s). Note that the matrix product AB is the zero matrix unless $j=r$, in which case its single nonzero entry is

$$(AB)_{is}=A_{ij}\wedge B_{js}.$$

Changing our perspective and looking at these objects as matrix-valued forms, we have

$$\alpha=A_{ij}\otimes e_{ij}\qquad\text{and}\qquad\beta=B_{rs}\otimes e_{rs}$$

where e_{pq} denotes the matrix with all zeros except a single 1 in spot (p,q). Then

$$\alpha\wedge\beta=(A_{ij}\wedge B_{rs})\otimes(e_{ij}\cdot e_{rs}).$$

The matrix product $e_{ij}\cdot e_{rs}$ is zero unless $j=r$, in which case it equals e_{is}, and so one readily sees that the matrix of forms corresponding to $\alpha\wedge\beta$ is AB.

REMARK A.3.4. The following is often useful. Let $\alpha \in \Omega^1(B;\mathcal{E})$ and $\beta \in \Omega^1(B;\mathcal{F})$. Then $\alpha \wedge \beta \in \Omega^2(B;\mathcal{G})$ is the 2-form whose value at tangent vector $v, w \in T_bB$ is

$$(\alpha \wedge \beta)_b(v,w) = \alpha(v)\beta(w) - \alpha(w)\beta(v).$$

To verify this, choose local coordinates $x^1, \dots, x^n$ on B and write $\alpha = \sum A_i dx^i$ and $\beta = \sum B_j dx^j$. Then $\alpha \wedge \beta = \sum_{i,j} A_i B_j dx^i \wedge dx^j$ and so

$$\begin{aligned}(\alpha \wedge \beta)(v,w) &= \sum_{i,j} A_i B_j [dx^i \wedge dx^j](v,w) \\ &= \sum_{i,j} A_i B_j \big[dx^i(v)dx^j(w) - dx^i(w)dx^j(v)\big] \\ &= \Big(\sum_i A_i dx^i(v)\Big)\Big(\sum_j B_j dx^j(w)\Big) - \Big(\sum_i A_i dx^i(w)\Big)\Big(\sum_j B_j dx^j(v)\Big) \\ &= \alpha(v)\beta(w) - \alpha(w)\beta(v).\end{aligned}$$

A.3.5. The deRham maps.

We have already defined $d_D \colon \Gamma(E) \to \Gamma(T^*B \otimes E)$. Extend this to $d_D \colon \Omega^p(B;E) \to \Omega^{p+1}(B;E)$ by requiring that

$$d_D(\alpha \otimes s) = (d\alpha) \otimes s + (-1)^p \alpha \wedge d_D(s)$$

where $d\alpha$ denotes the ordinary deRham differential on $\Omega^*(B)$. We leave it to the reader to check that this gives a well-defined map. So we have a sequence of composable maps

$$\text{(A.3.6)} \qquad \Gamma(E) = \Omega^0(B;E) \xrightarrow{d_D} \Omega^1(B;E) \xrightarrow{d_D} \Omega^2(B;E) \xrightarrow{d_D} \cdots$$

Unlike the ordinary deRham situation, this is not actually a cochain complex. The following proposition in some sense says that the extent to which it fails to be a cochain complex is measured by the curvature of the connection D:

PROPOSITION A.3.7. *Let $R \in \Omega^2(B; \underline{\mathrm{End}}(E))$ be the curvature 2-form for the affine connection D. Then for any $\eta \in \Omega^*(B;E)$, $d_D^2(\eta) = R \wedge \eta$. (The wedge product here is the one from Example A.3.2(a).)*

PROOF. The heart of the proof is checking this when $\eta \in \Omega^0(B;E)$. To see that it reduces to this case, let $s_1, \dots, s_n$ be a local basis of sections for E. Then any $\eta \in \Omega^p(B;E)$ may be written (locally) as $\eta = \sum_i \alpha_i \otimes s_i$ for some $\alpha_i \in \Omega^p(B)$. We compute that

$$d_D(\eta) = \sum_i \big[(d\alpha_i) \otimes s_i + (-1)^p \alpha_i \wedge d_D\big] s_i$$

and so

$$\begin{aligned} d_D^2(\eta) &= \sum_i (-1)^{p+1}(d\alpha_i) \otimes d_D s_i + (-1)^p d\alpha_i \wedge d_D s_i + (-1)^p(-1)^p \alpha_i \wedge d_D^2 s_i \\ &= \sum_i \alpha_i \wedge d_D^2 s_i. \end{aligned}$$

If we know that $d_D^2 s_i = R \wedge s_i$ then we have

$$d_D^2(\eta) = \sum_i \alpha_i \wedge R \wedge s_i = \sum_i R \wedge \alpha_i \wedge s_i = R \wedge \sum_i \alpha_i \wedge s_i = R \wedge \eta.$$

Note that in stating $\alpha_i \wedge R = R \wedge \alpha_i$ we are using two things, first that R is even-dimensional (it is a 2-form) and secondly the commutativity of scalar multiplication on the bundle $\underline{\mathrm{End}}(E)$ (the values of α_i being scalars).

It remains to verify that $d_D^2(s_i) = R \wedge s_i$. If $d_D s_i = D_{(-)}s_i = \sum \omega_{ij}s_j$, where $\Omega = (\omega_{ij})$ is the local matrix for the connection D. Then

$$\begin{aligned} d_D^2(s_i) = d_D\Bigl(\sum_j \omega_{ij} \wedge s_j\Bigr) &= \sum_j (d\omega_{ij}) \wedge s_j - \omega_{ij} \wedge d_D s_j \\ &= \sum_j \Bigl((d\omega_{ij}) \otimes s_j - \sum_k \omega_{ij} \wedge \omega_{jk} \wedge s_k\Bigr) \\ &= \sum_k \Bigl(d\omega_{ik} - \sum_j \omega_{ij} \wedge \omega_{jk}\Bigr) \wedge s_k \\ &= \sum_k (d\Omega - \Omega \wedge \Omega)_{ik} \wedge s_k \\ &= \sum_k R_{ki} \wedge s_k \\ &= R \wedge s_i. \end{aligned}$$

In the last two lines, R_{ki} is the local 2-form giving the matrix element for R with respect to the basis $s_1, \dots, s_n$, and the equation $R \wedge s_i = \sum_k R_{ki} \wedge s_k$ is an exercise in chasing through definitions. □

A.3.8. The Bianchi identity. Let R be the curvature form of a vector bundle $E \to B$ with connection D. As we have remarked before, for any $b \in B$ and any two tangent vectors $v, w \in T_bB$, the endomorphism $R_{v,w} \colon E_b \to E_b$ says something about what happens when you parallel transport vectors in E_b along small "parallelograms" oriented along v and w. Suppose now that one has a third tangent vector $u \in T_bB$. We can consider small "parallelpipeds" oriented along u, v, and w, and it turns out that there is a relation between the different ways of parallel-transporting around this parallelpiped. This results in a special property satisfied by all curvature forms, called the *Bianchi identity*.

The description in the above paragraph is vague, but it gives a rough version of the basic idea. The very short version is: curvature forms satisfy a special equation. As always in differential geometry, there are multiple ways of writing down this equation—different contexts, different notations, etc. We will explain a few of these.

The first version of the Bianchi identity we present is, in my opinion, the most elegant. But it also requires the most machinery, and it will use a version of the d_D map constructed in the last section. More precisely, recall from ???? that if one has connections on bundles $\mathcal{E}$ and $\mathcal{F}$ then there is an induced connection on the bundle $\underline{\mathrm{Hom}}(\mathcal{E}, \mathcal{F})$. So the connection D on E induces a connection on $\underline{\mathrm{End}}(E)$—we will also call this new connection D, by abuse of notation. As a result, we have maps $d_D \colon \Omega^p(B; \underline{\mathrm{End}}(E)) \to \Omega^{p+1}(B; \underline{\mathrm{End}}(E))$.

Proposition A.3.9. *Let $E \to B$ be a smooth vector bundle with connection D, and let $R \in \Omega^2(B; \underline{\mathrm{End}}(E))$ the associated curvature form. Then $d_D(R) = 0$.*

Proof. For $\eta \in \Omega^*(B; E)$ we compute $d_D^3(\eta)$ in two different ways:

$$d_D^3(\eta) = d_D(d_D^2\eta) = d_D(R \wedge \eta) = (d_D R) \wedge \eta + R \wedge (d_D \eta)$$

and

$$d_D^3(\eta) = d_D^2(d_D(\eta)) = R \wedge (d_D\eta).$$

Comparing these two, we find that $(d_DR) \wedge \eta = 0$ for all forms $\eta \in \Omega^*(B;E)$. The only way this can happen is if $d_DR = 0$ (if there were some $p \in B$ where $(d_DR)_p \neq 0$ then one could choose η to be a section of E that is nonzero in a neighborhood of p, and this would yield a contradiction). □

Although the equation $d_DR = 0$ is very simple to state, it is not so clear what actual geometry is encoded in it! This is often the curse of modern mathematics: machinery is developed that leads to extremely efficient and natural statements, but at the same time makes the meaning more obtuse. So let us look at some other versions of the Bianchi identity.

Recall that there is an induced connection on the bundle End(E), and we also denote this connection by D. If $\mathfrak{X}$ and $\mathcal{Y}$ are vector fields on B then $R_{\mathfrak{X},\mathcal{Y}}$ is a section of End(E), and hence we can apply D to it and obtain another section of End(E). The Bianchi identity says the following:

$$D_{\mathfrak{U}}(R_{\mathfrak{X},\mathcal{Y}}) + D_{\mathfrak{X}}(R_{\mathcal{Y},\mathfrak{U}}) + D_{\mathcal{Y}}(R_{\mathfrak{U},\mathfrak{X}}) = 0 \tag{A.3.10}$$

for all vector fields $\mathfrak{U}$, $\mathfrak{X}$, and $\mathcal{Y}$ on B.

Another way to write the Bianchi identity is in terms of commutators:

$$[D_{\mathfrak{U}}, R_{\mathfrak{X},\mathcal{Y}}] + [D_{\mathfrak{X}}, R_{\mathcal{Y},\mathfrak{U}}] + [D_{\mathcal{Y}}, R_{\mathfrak{U},\mathfrak{X}}] = 0. \tag{A.3.11}$$

At first this might seem very different from (A.3.10), so it could be surprising to learn that they are equivalent. The key is to realize that the symbol D means something different in the two equations. In (A.3.10) it refers to the connection on $\underline{\text{End}}(E)$, whereas in (A.3.11) it refers to the connection on E. Note that both $R_{\mathfrak{X},\mathcal{Y}}$ and $D_{\mathfrak{U}}$ (where $D = D^E$) are operators on the space of sections of E, and we may compose them (in either order) to obtain other operators. The commutators in (A.3.11) are to be interpreted in this context.

To explain the equivalence of (A.3.10) and (A.3.11), let us for just a moment write $\tilde{D}$ for the connection on $\underline{\text{End}}(E)$. Recall that this is defined by

$$(\tilde{D}\alpha)(s) = D(\alpha s) - \alpha(Ds),$$

where α is any section of $\underline{\text{End}}(E)$. Writing $\tilde{D}(\alpha) = [D, \alpha]$ is a reasonable encoding of the same statement. Thus, the term $D_{\mathfrak{U}}(R_{\mathfrak{X},\mathcal{Y}})$ in (A.3.10) would be written $\tilde{D}_{\mathfrak{U}}(R_{\mathfrak{X},\mathcal{Y}})$ in our present discussion, and is exactly equal to the term $[D_{\mathfrak{U}}, R_{\mathfrak{X},\mathcal{Y}}]$ in (A.3.11). The same of course applies to the other two terms.

Note that if $\mathfrak{U}$, $\mathfrak{X}$, and $\mathcal{Y}$ are commuting vector fields—e.g., coordinate vectors fields for a certain system of coordinates on B—then $R_{\mathfrak{U},\mathfrak{X}} = [D_{\mathfrak{U}}, D_{\mathfrak{X}}]$ and similarly for the other curvature operators. So (A.3.11) is equivalent to

$$[D_{\mathfrak{U}}, [D_{\mathfrak{X}}, D_{\mathcal{Y}}]] + [D_{\mathfrak{X}}, [D_{\mathcal{Y}}, D_{\mathfrak{U}}]] + [D_{\mathcal{Y}}, [D_{\mathfrak{U}}, D_{\mathfrak{X}}]] = 0,$$

which is a triviality—it is just the Jacobi identity for the operators $D_{\mathfrak{U}}$, $D_{\mathfrak{X}}$, and $D_{\mathcal{Y}}$. This does not exactly serve to *prove* the Bianchi identity, as of course it is not true that all vectors fields commute, but it at least helps provide some intuition about why such a formula might hold.

REMARK A.3.12. The formulas we have called the "Bianchi identity" are in some texts called the *second* Bianchi identity. In those texts the "first" Bianchi identity is something that holds only for the curvature form of the Levi-Civita connection on the tangent bundle of a Riemannian manifold:

$$R_{u,v}(w) + R_{v,w}(u) + R_{w,u}(v) = 0$$

for all $p \in B$ and all $u, v, w \in T_pB$. (Note that such a formula wouldn't even make sense for the curvature form of an arbitrary bundle, because u, v, and w would be two tangent vectors and a section—it would not make sense to permute them)

The first and second Bianchi identities are of a very different nature: note that the former contains no derivatives. There is a certain similarity between them in that they both can be written in a form involving cyclic permutations of three variables—in fact, they both connected to the Jacobi identity—but other than that they are not particularly related.

A.4. Principal connections

Let G be a Lie group and let $P \to B$ be a principal G-bundle. Note that the right action of G on P gives rise to a right action of G on TP. Precisely, if $v \in T_pP$ and $g \in G$ then define

$$v \cdot g = \frac{d}{dt}\bigg|_{t=0} (\gamma(t) \cdot g)$$

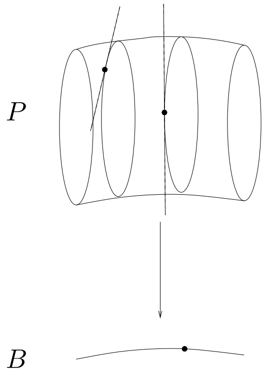
where γ is any path in P having tangent vector v when $t = 0$. As another perspective, if $R_g \colon P \to P$ is right multiplication by g then its derivative at $p \in P$ is $(DR_g)_p \colon T_pP \to T_{pg}P$. Our definition just says that

$$v \cdot g = (DR_g)_p(v).$$

It is clear from this description that we indeed have a well-defined action.

Given $p \in P$ there is a smooth embedding $j_p \colon G \hookrightarrow P$ given by $g \mapsto pg$. The derivative of j_p at the identity is a map $(Dj_p)_e \colon \mathfrak{g} \to T_pP$. Let $V_p \subseteq T_pP$ be the image of this map, which we will call the *vertical subspace* of T_pP. It is precisely the tangent space of the fiber of $P \to B$ that passes through p. Note that the collection of all V_p forms a subbundle of TP, called the *vertical subbundle.*

The picture below depicts a principal bundle $P \to B$—with the fibers drawn as circles—and two fibers of the vertical subbundle of TP.



While at every point in P there is a well-defined vertical subspace, notice that there is no canonical choice of a complementary "horizontal" subspace. We will discuss this more in just a moment, as it is intimately tied up with the notion of a connection on a principal bundle.

There are certain canonical vector fields that exist on any principal bundle. In the above picture, for instance, there is the vector field that circulates "counterclockwise" around the fibers. In general, any element of T_eG determines a vector field on P. if $X \in \mathfrak{g}$ we define this precisely by

$$X_p = (Dj_p)_e(X).$$

Then $p \mapsto X_p$ gives a vector field on P which we will denote by ϕX but sometimes abbreviate to just X. It is important to note that $X_{pg} \neq X_p \cdot g$. Instead there is the more complicated relation

$$X_{pg} = (gXg^{-1})_p \cdot g = [\mathrm{Ad}(g)(X)]_p \cdot g. \tag{A.4.1}$$

Here we have written gXg^{-1} as a useful abbreviation for the adjoint action of G on $\mathfrak{g}$, since in the case of matrix groups the adjoint action is exactly conjugation.

To understand (A.4.1) we argue as follows. Pick a path $\gamma\colon [-1,1] \to G$ such that $g(0) = e$ and $g'(0) = X$. Then

$$\begin{aligned} X_{pg} = \frac{d}{dt}\bigg|_{t=0} (pg \cdot \gamma(t)) = \frac{d}{dt}\bigg|_{t=0} (p \cdot g\gamma(t)) &= \frac{d}{dt}\bigg|_{t=0} (p \cdot g\gamma(t)g^{-1} \cdot g) \\ &= \frac{d}{dt}\bigg|_{t=0} (p \cdot g\gamma(t)g^{-1}) \cdot g \\ &= (gXg^{-1})_p \cdot g. \end{aligned}$$

The second-to-last equality is the definition of the right G-action on TP, and the last equality is because $t \mapsto g\gamma(t)g^{-1}$ is a path in G whose tangent vector at $t = 0$ is gXg^{-1}.

Example A.4.2.

(a) Extending the above ideas, produce a map $\mu\colon P \times TG \to TP$ such that
   (1) If $x \in G$, $p \in P$, and $v \in T_xG$ then $\mu(p,x)$ lies in $T_{px}P$;
   (2) The restriction of μ to $P \times T_eG$ sends (p, X) to $X_p \in T_pP$ as defined above.

(b) Let Z be any smooth manifold and consider smooth maps $\sigma\colon Z \to P$ and $g\colon Z \to G$. Let $\sigma g\colon Z \to P$ denote $z \mapsto \sigma(z)g(z)$. For $z \in Z$ and $v \in T_zZ$, verify that

$$D(\sigma g)_z(v) = (D\sigma_z)(v) \cdot g(z) + \sigma(v) \cdot (Dg)_z(v).$$

Here the product in the first term refers to the right action of G on TP, whereas the product in the second term refers to the map μ constructed in the first part of the exercise.

(c) If $g \in G$ and $X \in T_xG$ then $Xg \in T_{xg}G$. This is the right action of G on its tangent bundle. Prove that for any $p \in P$ and $X \in T_xG$ one has

$$[p \cdot X] \cdot g = p \cdot (Xg).$$

Here the first and the third multiplication symbols refer to the map μ constructed in (a), whereas the second such symbol is the right action of G on TP. Note also that $p \cdot X \in T_{px}P$, and so $[p \cdot X] \cdot g \in T_{pxg}P$. Likewise, $Xg \in T_{xg}G$ and so $p \cdot (Xg) \in T_{pxg}P$.

Here is a useful lemma to help reinforce your understanding of some of these concepts:

Lemma A.4.3. *Let $X \in \mathfrak{g}$ and let ϕX denote the associated vertical vector field. Let $\mathcal{Y}$ be any vector field on P that is G-equivariant, in the sense that $\mathcal{Y}_{p\cdot g} = \mathcal{Y}_p \cdot g$ for all $p \in P$, $g \in G$. Then $[\phi X, \mathcal{Y}] = 0$.*

Proof. Let $\varphi_t(p) = p \cdot e^{tX}$, for every $p \in P$; equivalently, φ_t is the right multiplication map $R_{e^{tX}}$. Then φ is a local flow for the vector field ϕX. Differential geometry (???) tells us that

$$[\phi X, \mathcal{Y}]_p = \lim_{t\to 0} \frac{1}{t}\Big[\mathcal{Y}_{\varphi_t(p)} - (D\varphi_t)_p(\mathcal{Y}_p)\Big] = \lim_{t\to 0} \frac{1}{t}\Big[\mathcal{Y}_{p\cdot e^{tX}} - (\mathcal{Y}_p \cdot e^{tX})\Big].$$

The G-equivariance of $\mathcal{Y}$ tells us that the final expression inside the brackets is zero. □

A.4.4. Connections. We are now ready to define the notion of a connection on a principle bundle. As often is the case in differential geometry, there are several ways of looking at this same concept. We concentrate on two main approaches.

DEFINITION A.4.5. *Let $P \to B$ be a principal G-bundle.*

*(1) A **horizontal connection** on P is a sub-vector-bundle $H \subseteq TP$ such that for every $p \in P$ one has $H_p \oplus V_p = T_pP$ and such that $H_{pg} = H_p \cdot g$ for every $p \in P$, $g \in G$. The subspaces $H_p \subseteq T_pP$ are called horizontal subspaces.*

*(2) A **principal connection** on P is a form $\omega \in \Omega^1(P; \mathfrak{g})$ such that*
   (a) $\omega_p(X_p) = X$ for every $X \in \mathfrak{g}$, and
   (b) $\ker \omega_{pg} \supseteq (\ker \omega_p) \cdot g$ for every $p \in P$ and $g \in G$.

Condition (b) looks different from how most texts define the notion principal connection, but we will shortly see that it is equivalent to the commonly-used version.

To understand the relationship between the definitions in (1) and (2), note that if ω is a principal connection then ω_p is a linear map $T_pP \to \mathfrak{g}$ which is surjective and becomes an isomorphism when restricted to the vertical subbundle (because of property (a)). The kernel of ω_p is therefore a complement to V_p inside of T_pP, and is our candidate for the horizontal subspace. Note that $\dim(\ker \omega_p)$ is therefore $\dim P - \dim G$, and this holds for all $p \in P$. It follows by equality of dimensions that the subset in property (b) is actually an equality, and therefore these kernels satisfy the condition for being a horizontal connection.

Conversely, given a horizontal connection $H \subseteq TP$ then one defines $\omega_p \colon T_pP \to \mathfrak{g}$ to be the unique map which is zero on H_p and satisfies property (a) on V_p (in other words, it sends X_p to X for every $X \in \mathfrak{g}$). The kernel of this form is exactly H_p, and so property (b) is satisfied and we have a principal connection. Thus, we see that the two notions of connection are completely equivalent.

Principal connections enjoy a certain equivariance property, giving a relationship between $\omega_{pg}(v \cdot g)$ and $\omega_p(v)$:

PROPOSITION A.4.6. *If $\omega \in \Omega^1(P; \mathfrak{g})$ is a principal connection then*

$$\omega_{pg}(v \cdot g) = g^{-1}[\omega_p(v)]g = \mathrm{Ad}(g^{-1})[\omega_p(v)]$$

for every $p \in P$, $v \in T_pP$, and $g \in G$.

PROOF. Since V_p and $H_p = \ker \omega_p$ are complementary subspaces of T_pP, it will suffice to check the equation for $v \in V_p$ and $v \in H_p$. In the latter case, note that $v \cdot g \in \ker \omega_{pg}$ by property (b) from the definition of principal connection, and so both sides of the equation are zero.

It remains to check the equation for $v \in V_p$. We claim that this is actually forced by condition (a) in the definition of principal connection. Explicitly, condition (a) in conjunction with (A.4.1) gives that

$$X = \omega_{pg}(X_{pg}) = \omega_{pg}((gXg^{-1})_p \cdot g)$$

for every $X \in \mathfrak{g}$. Since any vector $Y \in \mathfrak{g}$ may be written as gXg^{-1} for some X (namely $X = g^{-1}Yg$), this is equivalent to

$$\omega_{pg}(Y_p \cdot g) = g^{-1}Yg = g^{-1}[\omega_p(Y_p)]g$$

for every $Y \in \mathfrak{g}$. Since every vector $v \in V_p$ is of the form Y_p for some $Y \in \mathfrak{g}$, this completes the proof. □

Remark A.4.7. Many texts replace condition (b) in the definition of principal connection with the equation from Proposition A.4.6. We have shown that the former implies the latter, and it is easy to see that the converse is also true. So our definition is equivalent to the more common one.

Finally, we give a way of looking at connections in terms of vector bundles. Let π denote the map $P \to B$. Then there is a short exact sequence of vector bundles on P

$$0 \to V \to TP \to \pi^*(TB) \to 0.$$

We claim that specifying a horizontal connection on $P \to B$ is equivalent to specifying a splitting of this sequence. Clearly a splitting χ *gives* a horizontal connection, simply by taking the horizontal subspaces to be the image of the splitting map. Conversely, suppose give a horizontal connection. Then for every $p \in P$ and $v \in T_{\pi(p)}B$ there is a *unique* vector in H_p that is a preimage for v under $(D\pi)_p$, and this readily translates into a definition for the splitting χ.

For any $p \in B$ and $v \in T_{\pi(p)}B$, let $v_p \in H_p$ denote the vector given by the above splitting. This gives a vector field $p \mapsto v_p$ defined on the fiber $\pi^{-1}(p)$, and it is readily seen to be G-equivariant: the tangent vector $v_p \cdot g$ lies in H_p and is a lifting for v, so by uniqueness we must have $v_p \cdot g = v_{pg}$. Generalizing this somewhat, we find that any vector field $\mathfrak{X}$ on B lifts uniquely to a horizontal vector field on P (obtained by applying the splitting χ), and that such vector fields are G-equivariant. Vector fields on P that are both horizontal and G-equivariant are called **basic** vector fields; they are in bijective correspondence with vector fields on B.

A.4.8. Local form of a principal connection. Suppose that $U \subseteq B$ is an open set over which P is trivial. This means that there is a section $\sigma\colon U \to P$. Let $\Omega_U^\sigma = \sigma^*\omega \in \Omega^1(U;\mathfrak{g})$. If $\chi\colon U \to P$ is another section, then we can write $\chi(u) = \sigma(u)g(u)$ for a unique map $g\colon U \to G$. We claim that

$$\Omega_U^\chi = g\Omega_U^\sigma g^{-1} + g^{-1}\,dg. \tag{A.4.9}$$

As usual, let us first process what the terms in this equation even mean. If v is a tangent vector to U at the point u, then $\Omega_U^\sigma(v)$ is an element of $\mathfrak{g}$. Therefore $g\Omega_U^\sigma(v)g^{-1}$ refers to the action of g on $\Omega_U^\sigma(v)$ under the adjoint representation (usually written $\mathrm{Ad}(g)[\Omega_U^\sigma(v)]$)—this is also an element of $\mathfrak{g}$. The object $(dg)_u(v)$ is another name for $(Dg)_u(v)$ and lies in $T_{g(u)}G$. Left-multiplying by $g(u)^{-1}$ moves this tangent vector into T_eG, and so $g(u)^{-1} \cdot dg_u(v)$ is again an element of $\mathfrak{g}$.

Let us prove equation (A.4.9). Let $b \in U$ and $v \in T_bU$. Then

$$\begin{aligned}
(\Omega_U^\chi)_b(v) &= \omega_{\chi(b)}\Big((D\chi)_b(v)\Big)\\
&= \omega_{\chi(b)}\Big((D\sigma)_b(v)\cdot g(b) + \sigma(b)\cdot(Dg)_b(v)\Big) \quad \text{using Exercise A.4.2(b)}\\
&= \omega_{\chi(b)}\Big((D\sigma)_b(v)\cdot g(b) + \big[\sigma(b)\cdot(Dg)_b(v)g(b)^{-1}\big]\cdot g(b)\Big) \quad \text{using (A.4.2)(c)}
\end{aligned}$$

To ease notation let us write $p = \sigma(b)$, $g = g(b)$, and $X = (Dg)_b(v) \in T_{g(b)}G$. Continuing, we have

$$\begin{aligned}(\Omega_U^\chi)_b(v) &= \omega_{pg}\Big(\big[(D\sigma)_b(v) + p\cdot Xg^{-1}\big]\cdot g\Big)\\ &= g^{-1}\Big[\omega_p\big((D\sigma)_b(v) + p\cdot Xg^{-1}\big)\Big]g \quad \text{by Proposition A.4.6}\\ &= g^{-1}\Big[\omega_p\big((D\sigma)_b(v)\big) + \omega_p\big(p\cdot Xg^{-1}\big)\Big]g\\ &= g^{-1}\Big[(\Omega_U^\sigma)_b(v) + Xg^{-1}\Big]g \quad \text{by (a) from defn. of a principal connection}\\ &= g^{-1}\Big[(\Omega_U^\sigma)_b(v)\Big]g + g^{-1}X\\ &= g(b)^{-1}\Big[(\Omega_U^\sigma)_b(v)\Big]g(b) + g(b)^{-1}(Dg)_b(v).\end{aligned}$$

This completes our explanation of (A.4.9).

????

Remark A.4.10. The reader should note the similarity between the present discussion and the discussion of affine connections for vector bundles. If $E \to B$ is a real vector bundle then $\mathrm{Fr}(E) \to B$ is a principal $GL_n(\mathbb{R})$-bundle. A local trivialization of E is the same as a local section of $\mathrm{Fr}(E)$. The Lie algebra of $\mathrm{GL}_n(\mathbb{R})$ is simply $M_{n\times n}(\mathbb{R})$ (the general linear group is an open subset of $M_{n\times n}(\mathbb{R})$ and so they share the same tangent spaces). ?????

A.4.11. Horizontal and vertical forms, and the covariant derivative. Let $P \to B$ be a principal G-bundle and let ω be a principal connection. Recall that ω determines horizontal subspaces $H_p \subseteq T_pP$ via $H_p = \ker \omega_p$. Let us say that an r-form $\alpha \in \Omega^r(P;\mathfrak{g})$ is **horizontal** if $\alpha_p(u_1,\ldots,u_r) = 0$ whenever at least one u_i belongs to V_p. In other words, horizontal forms vanish if any of their inputs is a vertical tangent vector. Let $\Omega^*_H(P;\mathfrak{g}) \subseteq \Omega^*(P;\mathfrak{g})$ denote the subspace of all horizontal forms. There is a similar definition for vertical forms. Note that ω itself is vertical.

The connection ω gives us a retraction $\rho\colon \Omega^*(P;\mathfrak{g}) \to \Omega^*_H(P;\mathfrak{g})$. To see this, for any $p \in P$ let $\pi_p\colon T_pP \to H_p$ be the projection that annihilates V_p. For $\alpha \in \Omega^r(P;\mathfrak{g})$ define $\rho(\alpha)$ to be the form given by

$$[\rho(\alpha)]_p(u_1,\ldots,u_r) = \alpha(\pi_p(u_1),\ldots,\pi_p(u_r)).$$

We leave it as an exercise to check that if α is horizontal then $\rho(\alpha) = \alpha$.

The **covariant derivative** associated to the principal connection is the map $D\colon \Omega^r(P;\mathfrak{g}) \to \Omega^{r+1}_H(P;\mathfrak{g})$ given by $D(\alpha) = \rho(d\alpha)$. In other words, D is the composite

$$\Omega^r(P;\mathfrak{g}) \xrightarrow{d} \Omega^{r+1}(P;\mathfrak{g}) \xrightarrow{\rho} \Omega^{r+1}_H(P;\mathfrak{g}).$$

A.4.12. Curvature of a principal connection. Again suppose that $P \to B$ is a principal G-bundle with a connection $\omega \in \Omega^1(P;\mathfrak{g})$. The curvature form associated to this connection is defined to be

$$R = D\omega \in \Omega^2_H(P;\mathfrak{g}).$$

This 2-form can also be described as follows:

Proposition A.4.13. $R = d\omega + \frac{1}{2}[\omega,\omega]$.

The reader should compare the above description to ????, keeping in mind Example ????.

Proof. We must prove, for every $p \in P$ and $u, v \in T_pP$, that

$$(d\omega)(\pi_p u, \pi_p v) = (d\omega)(u,v) + \frac{1}{2}[\omega,\omega](u,v). \tag{A.4.14}$$

By bilinearity and antisymmetry, it sufficese to check this in three cases: (i) $u, v \in H_p$, (ii) $u \in H_p$ and $v \in V_p$, and (iii) $u, v \in V_p$. Note that $[\omega,\omega](u,v) = 2[\omega(u), \omega(v)]$ by Remark A.3.4. So if either u or v lies in H_p then this term vanishes.

Case (i): $\boldsymbol{u, v \in H_p}$. Since $\pi_p(u) = u$ and $\pi_p(v) = v$ and the third term of (A.4.14) vanishes, this case is obvious.

Case (ii): $\boldsymbol{u \in H_p,\ v \in V_p}$. Here $\pi_p(v) = 0$, so the first and third terms in (A.4.14) vanish. To analyze the second term we first extend u locally to a basic vector field $\mathfrak{U}$ (recall that "basic" means horizontal and G-equivariant). Since v is vertical, we can write $v = Y_p$ for some $Y \in \mathfrak{g}$. Let us write $\mathcal{Y}$ for the vertical vector field on P determined by Y. Next, use the formula

$$(d\omega)(\mathfrak{U}, \mathcal{Y}) = \partial_{\mathfrak{U}}(\omega(\mathcal{Y})) - \partial_{\mathcal{Y}}(\omega(\mathfrak{U})) - \omega([\mathfrak{U}, \mathcal{Y}])$$

from Exercise A.2.6. The vector field $\omega(\mathfrak{U})$ is identically zero because $\mathfrak{U}$ is horizontal, and $\omega(\mathcal{Y})$ is the constant function Y by property (a) in the definition of principal connection—therefore $\partial_{\mathfrak{U}}(\omega(\mathcal{Y}))$ also vanishes. Finally, the bracket $[\mathfrak{U}, \mathcal{Y}]$ vanishes by Lemma A.4.3.

Case (iii): $\boldsymbol{u, v \in V_p}$. Here $\pi_p(u) = \pi_p(v) = 0$, so the left-hand-side of (A.4.14) vanishes. Since both u and v are vertical we can write $u = X_p$ and $v = Y_p$ for some $X, Y \in \mathfrak{g}$. Write $\mathfrak{X}$ and $\mathcal{Y}$ for the vector fields on P determined by X and Y. As in case (ii), we can write

$$(d\omega)(\mathfrak{X}, \mathcal{Y}) = \partial_{\mathfrak{X}}(\omega(\mathcal{Y})) - \partial_{\mathcal{Y}}(\omega(\mathfrak{X})) - \omega([\mathfrak{X}, \mathcal{Y}]).$$

The functions $\omega(\mathfrak{X})$ and $\omega(\mathcal{Y})$ are constant, and so the first two terms vanish. The Lie bracket $[\mathfrak{X}, \mathcal{Y}]$ is just the vector field associated to $[X, Y]$ by ???, so $(d\omega)(\mathfrak{X}, \mathcal{Y}) = -[X, Y]$. Finally, the third term of (A.4.14) is equal to $[\omega(\mathfrak{X}), \omega(\mathcal{Y})] = [X, Y]$. At this point one readily verifies that (A.4.14) holds. □

APPENDIX B

Background on compact Lie groups and their representations

When examining the progression of physics over the last century, one thing that stands out is the ever-increasing role of *symmetry*. These symmetries appear most often in terms of an action of some Lie group G. The rotation group $SO(3)$ is perhaps the most intuitive example here, but as physics has progressed it has found ample use for less familiar groups like Spin(3) (in the Dirac theory of electron spin), $SU(3)$ (in the most basic theory of quarks), and even the exceptional Lie groups E_6, E_7, and E_8. It is therefore important that one has a working knowledge of basic Lie theory: namely, what are all the compact Lie groups and what do we know about their representations?

There are ample texts that discuss these issues, but not always in a brief, "this-is-what-you-need-to-know" kind of way. We try to provide such a survey in the present section. Our presentation is heavily influenced by [**Ad**] and [**BtD**].

B.1. Root systems

Root systems are purely geometric objects. While it seems unlikely (to me) that they would have arisen naturally, without their role in Lie theory, it is useful to introduce them as if that had happened. Doing so makes a clear separation between what parts of the classification are "about Lie groups" and what parts are really about "configurations of vectors in Euclidean space".

Definition B.1.1. *Let V be a finite-dimensional real vector space with a positive-definite inner product. A **root system** in V is a finite collection Φ of nonzero vectors such that*

(1) The elements of Φ span V.
(2) If $\alpha \in \Phi$ then $-\alpha \in \Phi$, and no other scalar multiples of α belong to Φ.
(3) For every $\alpha \in \Phi$, the set Φ is closed under reflection in the hyperplane orthogonal to α.
(4) For every $\alpha, \beta \in \Phi$, the projection of β onto the line spanned by α lies in $\mathbb{Z}\langle\frac{1}{2}\rangle.\alpha$. That is,

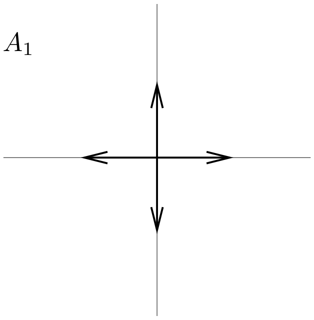
$$\frac{\langle \alpha, \beta\rangle}{\langle \alpha, \alpha\rangle} \in \tfrac{1}{2}\mathbb{Z}.$$

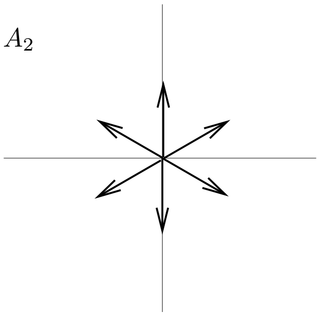
*The elements of Φ are called **roots**. The root system will usually be denoted as a pair (V, Φ). An isomorphism of root systems is an isomorphism of vector spaces $V \to V'$ (not necessarily an isometry!) which sends Φ bijectively to Φ'. The **rank** of a root system is the dimension of V.*

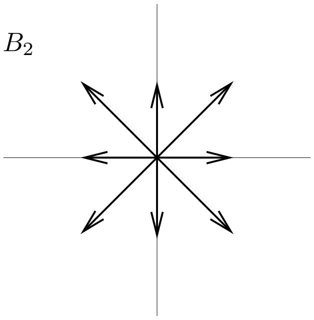
Remark B.1.2. Note that if (V, Φ) and (V', Φ') are two root systems, then $(V \oplus V', \Phi \cup \Phi')$ is also a root system (where $V \oplus V'$ is given the evident inner product in which V and V' are orthogonal). Root systems that do not arise in this way are called **irreducible**.

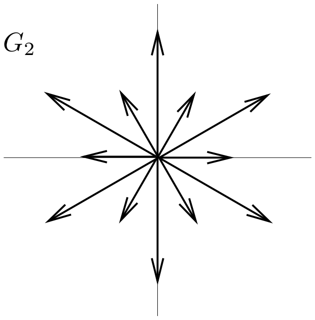
Remark B.1.3. If (V, Φ) is a root system and $U \subseteq V$ is a subspace, then $(U, U \cap \Phi)$ is also a root system.

Up to isomorphism there is only one root system of rank 1, namely $(\mathbb{R}, \{1, -1\})$. With a little work one finds that there are exactly four root systems of rank 2. They are called A_1, A_2, B_2, and G_2 (the reasons for these names will become clear a little later):









If α and β are independent roots in a root sytem, then restricting to the subspace $\mathbb{R}\langle \alpha, \beta \rangle$ gives a 2-dimensional root system (see Remark B.1.3). Since there are only four such 2-dimensional systems, we deduce the following:

Lemma B.1.4. *Let (V, Φ) be a root system. For any $\alpha, \beta \in \Phi$, the angle between them (in degrees) lies in the set $\{0, 30, 45, 60, 90, 120, 135, 150, 180\}$. Also, the following hold:*

(1) If the angle is $0°$ or $180°$, then $|\alpha| = |\beta|$.
(2) If the angle is $30°$ or $150°$, then $|\alpha|/|\beta|$ is $\sqrt{3}$ or $\frac{1}{\sqrt{3}}$.
(3) If the angle is $60°$ or $120°$ then $|\alpha| = |\beta|$.
(4) If the angle is $45°$ or $135°$ then $|\alpha|/|\beta|$ is $\sqrt{2}$ or $\frac{1}{\sqrt{2}}$.

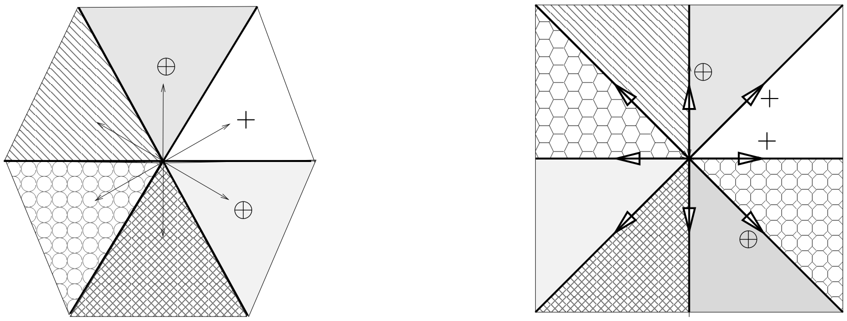
If $\alpha \in \Phi$ then let $R_\alpha \colon V \to V$ be reflection in the hyperplane orthogonal to α. The **Weyl group** of Φ is the subgroup of isometries of B generated by the reflections R_α. This is always a finite group. For the rank 2 root systems, the Weyl groups turn out to be $\mathbb{Z}/2 \times \mathbb{Z}/2$ for A_1, the symmetric group Σ_3 for A_2, the dihedral group D_4 (of order eight) for B_2, and the dihedral group D_6 (of order 12) for G_2.

For each i let H_i be the hyperplane orthogonal to the root θ_i. Removing these hyperplanes from V leaves a collection of connected components called **Weyl chambers**. The chambers are all conjugate under the action of the Weyl group. We choose one chamber C (it doesn't matter which one) to be called the **fundamental Weyl chamber**. (See Example B.1.5 below for two pictures).

Choose a hyperplane H that doesn't contain any of the roots, and such that the fundamental Weyl chamber is all contained on one side of the hyperplane. Exactly half the roots lie on one side of H, and the remaining half lie on the other. The roots that lie on the same side of H as C are called the **positive roots**, and the set of such roots is denoted Φ^+. Another (and faster) way to say all of this is that Φ^+ is the set of roots θ with the property that $\langle \theta, v \rangle \geq 0$ for all $v \in C$.

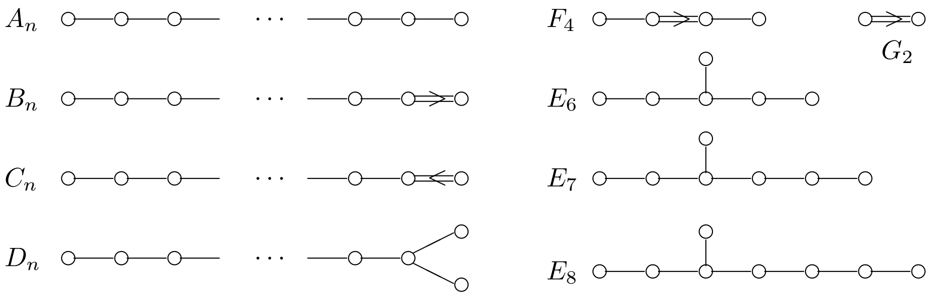
A root in Φ^+ is called **simple** if it cannot be written as the sum of two positive roots, and the set of simple roots is denoted Δ. One can prove that Δ forms a basis for V, and that the angle between any two simple roots is at least 90°.

Example B.1.5. The following diagram shows the hyperplanes H_i and the Weyl chambers for the systems A_2 and B_2; our choice of fundamental Weyl chamber is the non-shaded chamber. The positive roots are marked with $+$ or $\oplus$, with the *simple* roots all marked with $\oplus$.

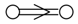


A root system gives a so-called **Dynkin diagram** in the following way. The diagram contains one node for every simple root. If two roots are orthogonal, they are not connected by an edge. If two roots make an angle of 120°, the nodes are connected by a single undirected edge. If the roots make an angle of 135°, they are connected by a double edge that is directed toward the shorter root. If the roots make an angle of 150°, they are connected by a triple edge that is again directed toward the shorter root.

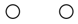
One proves that root systems are in bijective correspondence with Dynkin diagrams, with irreducible root systems corresponding to connected Dynkin diagrams. Without much trouble one can write down all the possible root systems: they break up into four infinite families and five sporadic ones, with the following Dynkin diagrams:



Note that in the families one starts counting basis elements from the right. So B_2 and C_2 are both the root system



which is also known as G_2, whereas D_2 is the root system



which is isomorphic to $A_1 \times A_1$. Note also that $D_3 \cong A_3$, but $D_4 \ncong A_4$.

B.1.6. Weyl groups. It is a theorem that the Weyl group acts simply-transitively on the set of Weyl chambers. In particular, the order of the Weyl group is the same as the number of Weyl chambers. Using this, it is not hard to determine the structure of the Weyl group for each of the root systems. To describe this, let $\Omega_n = \{e_1, \dots, e_n, -e_1, \dots, -e_n\} \subseteq \mathbb{R}^n$. Let $\Sigma_n \subseteq \text{Aut}(\Omega_n)$ be the subgroup of permutations of the indices, and let $U_n \subseteq \text{Aut}(\Omega_n)$ be the subgroup of all sign changes (so $U_n \cong (\mathbb{Z}/2)^n$). Finally, let $U_n^{ev} \subseteq \text{Aut}(\Omega_n)$ be the subgroup of even numbers of sign changes. The Weyl groups for types A_n–D_n are as follows:

A_n	Σ_n
B_n	the subgroup of $\text{Aut}(\Omega_n)$ generated by Σ_n and U_n
C_n	the subgroup of $\text{Aut}(\Omega_n)$ generated by Σ_n and U_n^{ev}
D_n	the subgroup of $\text{Aut}(\Omega_n)$ generated by Σ_n and U_n.

The Weyl group of G_2 is D_6 (order 12). The Weyl groups of $F4$, $E6$, $E7$, and $E8$ have respective orders of 1152; 51,840; 2,903,040; and 696,729,600.

B.2. Classification of simply-connected, compact Lie groups

Let G be a compact, simply-connected Lie group, and let $T \hookrightarrow G$ be a maximal torus of rank n. We have the exponential map $\exp\colon T_eG \to G$, which maps T_eT into T. As the notation T_eT is horrible, let us just denote this tangent space as M. The **integral lattice** inside of M is the set of points that map to e under the exponential map; let us denote this lattice as $M_{\mathbb{Z}} \subseteq M$.

The irreducible complex representations of a torus are all 1-dimensional, and therefore correspond bijectively with group homomorphisms $T \to S^1$. In turn, group homomorphisms $T \to S^1$ are in bijective correspondence with maps $M \to \mathbb{R}$ that send the integral lattice into $\mathbb{Z}$—in other words, with elements $\theta \in M^*$ having the property that $\theta(M_{\mathbb{Z}}) \subseteq \mathbb{Z}$. Such θ are called **weights**, and the subgroup of M^* consisting of all weights is called the **weight lattice** $M^*_{\mathbb{Z}} \subseteq M^*$. It is an easy exercise to check that $M^*_{\mathbb{Z}} \cong \mathrm{Hom}_{\mathbb{Z}}(M_{\mathbb{Z}}, \mathbb{Z})$; the two lattices $M^*_{\mathbb{Z}}$ and $M_{\mathbb{Z}}$ are naturally dual.

Next regard T_eG as a G-representation via the adjoint action. Restricting via $T \hookrightarrow G$ gives an induced T-action on T_eG, and therefore also on $\mathbb{C} \otimes_{\mathbb{R}} T_eG$. This complex representation of T will split up into irreducible representations (necessarily 1-dimensional). Some of these are trivial, and for the nontrivial ones let $\theta_1, \dots, \theta_r \in M^*_{\mathbb{Z}}$ be the corresponding weights. Finally, let $\Phi = \{\theta_1, \dots, \theta_r, -\theta_1, \dots, -\theta_r\}$. With a little work one can prove that Φ is a root system for M^*.

Here is a list of basic Lie groups and their associated root systems:

$SU(n)$	A_n
$SO(2n+1)$ or $\mathrm{Spin}(2n+1)$	B_n
$Sp(n)$	C_n
$SO(2n)$ or $\mathrm{Spin}(2n)$	D_n

Recall that $Sp(n)$ is the subgroup of $\mathrm{GL}_n(\mathbb{H})$ consisting of those matrices that preserve the standard Hermitian form on $\mathbb{H}^n$.

We explain some portions of the above table through the following examples:

Example B.2.1. Let $G = SU(n)$. For our maximal torus T we choose the set of diagonal matrices in $SU(n)$, namely the matrices of the form

$$\Omega = \begin{bmatrix} \omega_1 & & & \\ & \omega_2 & & \\ & & \ddots & \\ & & & \omega_n \end{bmatrix}$$

where $\omega_1, \dots, \omega_{n-1} \in S^1$ and $\omega_n = (\omega_1 \cdots \omega_{n-1})^{-1}$.

The tangent space T_eG consists of trace zero matrices X such that $X + \overline{X}^T = 0$, and it has the inner product given by $\langle X, Y\rangle = \sum_{j,k} x_{jk}\overline{y_{jk}}$. The space $M = T_eT$ is the subspace of diagonal matrices in T_eG. Thus M consists of matrices

$$\begin{bmatrix} ia_1 & & \\ & \ddots & \\ & & ia_n \end{bmatrix}$$

where $a_1, \dots, a_{n-1} \in \mathbb{R}$ and $a_n = -(a_1 + \cdots + a_{n-1})$. It will be convenient to identify elements of M with the corresponding vectors $\mathbf{a} = (a_1, \dots, a_n) \in \mathbb{R}^n$; under this identification M corresponds to the subspace of $\mathbb{R}^n$ consisting of vectors $\mathbf{a}$ with $\sum_k a_k = 0$.

Let e_i denote the diagonal matrix with (i,i)-entry equal to 1 and all other entries zero, which of course corresponds to the standard basis for $\mathbb{R}^n$ under our identification. Then a basis for M is $e_1 - e_n, e_2 - e_n, \dots, e_{n-1} - e_n$. Write $b_i = e_i - e_n$, so that $b_1, \dots, b_{n-1}$ is this basis. Note that $\langle b_j, b_k\rangle$ is 2 if $j = k$ and 1 if $j \neq k$.

Note that there are two natural candidates for a "dual basis" for M^*. One is the standard algebraic dual, defined by $\tilde{b}_i(b_j) = \delta_{ij}$. The other is defined using the inner product on M, and is given by $\hat{b}_i(b_j) = \langle b_i, b_j\rangle$. To understand the relationship between them, look at the vector of values $\tilde{b}_i(\mathbf{b}) = (\tilde{b}(b_1), \tilde{b}(b_2), \dots, \tilde{b}(b_{n-1}))$ and $\hat{b}_i(\mathbf{b}) = (\hat{b}(b_1), \hat{b}(b_2), \dots, \hat{b}(b_{n-1}))$. One has

$$\tilde{b}_i(\mathbf{b}) = (0, 0, \dots, 0, 1, 0, \dots, 0)$$

with the 1 in the ith spot, whereas

$$\hat{b}_i(\mathbf{b}) = (1, 1, \dots, 1, 2, 1, \dots, 1)$$

with the 2 in the ith spot. Note that under the identification of M with a subspace of $\mathbb{R}^n$, $\tilde{b}_i$ is the (restriction of the) functional that takes $\mathbf{a}$ to a_i. The basis $\tilde{b}$ generates the integral lattice for M^*, whereas the basis $\hat{b}$ does not.

There is one last bit of preparation which will be handy before we look into the Lie theory of this example. The inclusion $M \subseteq \mathbb{R}^n$ dualizes to give a surjection $(\mathbb{R}^n)^* \twoheadrightarrow M^*$. If $e_1, \dots, e_n$ is the standard basis for $\mathbb{R}^n$, let $\hat{e}_1, \dots, \hat{e}_n$ be the dual basis (note that the dual basis with respect to the algebraic pairing and the Euclidean inner product coincide). Then the map $(\mathbb{R}^n)^* \to M^*$ sends $\hat{e}_i$ to $\tilde{b}_i$ for $i < n$, whereas it sends $\hat{e}_n$ to $-(\tilde{b}_1 + \cdots + \tilde{b}_{n-1})$. It will be convenient to use this notation for elements of M^*, so that $\hat{e}_i$ and $\tilde{b}_i$ are synonymous for $i < n$ and $\hat{e}_n$ is just another name for $-(\tilde{b}_1 + \cdots + \tilde{b}_{n-1}) = -\frac{1}{n}(\hat{b}_1 + \cdots + \hat{b}_{n-1})$. In other words, we can identify M^* with the quotient $(\mathbb{R}^n)^*/\langle \hat{e}_1 + \cdots + \hat{e}_n \rangle$.

This use of $\mathbb{R}^n$ and $(\mathbb{R}^n)^*$, which are in some sense "external" to the situation, is not necessary at all: we could do everything that follows just using the $\tilde{b}$'s and $\hat{b}$'s. But the use of the standard basis for $(\mathbb{R}^n)^*$ makes certain formulas a little easier to remember. However, we need a WARNING: with this notation one must be careful about inner products. The vector $\hat{e}_n$ has norm 1 in $(\mathbb{R}^n)^*$, but $-(\tilde{b}_1 + \cdots + \tilde{b}_{n-1})$ does not have norm 1 in M^*. It is possible to get oneself confused via this. There is a way to make things work out, however. The map $(\mathbb{R}^n)^* \to M^*$ has a splitting which sends $\hat{b}_i$ to $\hat{e}_i - \hat{e}_n$ for all i, and this splitting map preserves all inner products. The image is the set of all sums $\sum_i c_i \hat{e}_i$ such that $\sum c_i = 0$. So as long as we represent cosets in the quotient $(\mathbb{R}^n)^*/\langle \hat{e}_1 + \cdots + \hat{e}_n \rangle$ by representatives with $\sum c_i = 0$, we can compute inner products and get the same answers as in M^*.

After this long preamble we can start looking at Lie theory. Our first task will be to analyze the adjoint action of T on T_eG and to determine the roots. For $r \neq s$ and $z \in \mathbb{C}$ write $X^{rs}(z)$ for the matrix with z in the (r,s)-spot, $-\bar{z}$ in the (s,r)-spot, and zeros elsewhere. Abbreviate $X^{rs}(1)$ as just X^{rs}. If $\Omega \in T$ then one readily checks that

$$\Omega \cdot X^{rs}(z) \cdot \Omega^{-1} = \omega_r \omega_s^{-1} \cdot X^{rs}(z).$$

It follows that $\{X^{rs}(z) \,|\, z \in \mathbb{C}\}$ is a two-dimensional irreducible real representation of T. When one complexifies, this breaks up into two 1-dimensional representations with characters $\Omega \mapsto \omega_r \omega_s^{-1}$ and $\Omega \mapsto \omega_r^{-1} \omega_s$. The corresponding weights are given by $\theta_{rs}(\mathbf{a}) = a_r - a_s$ and its negative. Note that $\theta_{rs} = \hat{e}_r - \hat{e}_s$, or else we could write

$$\theta_{rs} = \begin{cases} \hat{b}_r - \hat{b}_s & \text{if } r < n \text{ and } s < n; \\ \hat{b}_r & \text{if } s = n. \end{cases}$$

We choose our positive cone to be determined by the roots θ_{rs} for $r < s$. The simple roots are then

$$\theta_{12} = \hat{e}_1 - \hat{e}_2, \qquad \theta_{23} = \hat{e}_2 - \hat{e}_3, \qquad \dots \qquad \theta_{n-1,n} = \hat{e}_{n-1} - \hat{e}_n.$$

Write $\Theta_i = \theta_{i,i+1}$ for brevity. Note that in the $\hat{b}$-basis we would write

$$\Theta_1 = \hat{b}_1 - \hat{b}_2, \quad \Theta_2 = \hat{b}_2 - \hat{b}_3, \quad \ldots \quad \Theta_{n-2} = \hat{b}_{n-2} - \hat{b}_{n-1}, \quad \Theta_{n-1} = \hat{b}_{n-1}.$$

One readily computes that $\langle \Theta_i, \Theta_j \rangle = 0$ unless $j \in \{i-1, i\}$, and $\langle \Theta_i, \Theta_{i+1} \rangle = -1$ (use that the $\hat{e}$ basis for $\mathbb{R}^n$ is orthonormal). It follows from this that we are looking at the root system A_{n-1}.

Next let us consider the Weyl group. Let R_{rs} be the reflection in the hyperplane orthogonal to $\theta_{rs} = \hat{b}_r - \hat{b}_s$. For $i \notin \{r, s\}$, $\hat{b}_i$ is orthogonal to θ_{rs} and hence $R_{rs}(\hat{b}_i) = \hat{b}_i$. Likewise, since $\hat{b}_r = \frac{1}{2}(\hat{b}_r + \hat{b}_s) + \frac{1}{2}(\hat{b}_r - \hat{b}_s)$ and the first term is orthogonal to θ_{rs}, it follows that

$$R_{rs}(\hat{b}_r) = \tfrac{1}{2}(\hat{b}_r + \hat{b}_s) - \tfrac{1}{2}(\hat{b}_r - \hat{b}_s) = \hat{b}_s.$$

A similar computation shows that $R_{rs}(\hat{b}_s) = \hat{b}_r$ (or else use that $R_{rs}(\hat{b}_r - \hat{b}_s) = \hat{b}_s - \hat{b}_r$).

???

One gets a map of groups $W \to \Sigma_n$, where Σ_n is the symmetric group on n letters. It turns out that this is an isomorphism.

B.3. Representation theory

We start with some basic observations. Let $R_{\mathbb{C}}(G)$ denote the complex representation ring of G. Additively, this is a free abelian group with basis consisting of the irreducible representations; the multiplication is given by the tensor product.

A representation V of G has an induced **character** $\chi_V \colon G \to \mathbb{C}$ given by $\chi_V(g) = \mathrm{tr}(g_V)$ where $g_V \colon V \to V$ is left multiplication by g. These characters are **class functions**, meaning that they are continuous maps satisfying $\chi_V(hgh^{-1}) = \chi_V(g)$ for all $g, h \in G$. The set of all class functions $\mathrm{Cl}(G)$ forms a ring under pointwise addition and multiplication, and we obtain a ring map

$$\chi \colon R_{\mathbb{C}}(G) \to \mathrm{Cl}(G).$$

This turns out to be a monomorphism.

For a torus T the representation ring is easy to understand. The irreducible representations of T are 1-dimensional, and for such representations the character is a group homomorphism and must take its values in $S^1 \subseteq \mathbb{C}$. The 1-dimensional representations of T are in bijective correspondence with group homorphisms $T \to S^1$, which are in turn in bijective correspondence with elements of the weight lattice $M^*_{\mathbb{Z}} \subseteq M^*$ (recall $M = T_eT$ is the tangent space of T at the identity). If V and W are 1-dimensional representations with corresponding weights v and w, then $V \otimes_{\mathbb{C}} W$ is a 1-dimensional representation with weight $v + w$.

Recall that if H is a group then the group ring $\mathbb{Z}[H]$ consists of finite, formal linear combinations $\sum n_i[h_i]$ with $n_i \in \mathbb{Z}$, where the multiplication is determined by the group structure in H via the formula $[h_i] \cdot [h_j] = [h_ih_j]$. When H is the free abelian group $\mathbb{Z}^k$ note that $\mathbb{Z}[H]$ is an algebra of Laurent polynomials $\mathbb{Z}[x_1^{\pm 1}, \ldots, x_k^{\pm 1}]$. Consider $\mathbb{Z}[M^*_{\mathbb{Z}}]$, the group ring on the abelian group $M^*_{\mathbb{Z}}$; by the previous remarks this is an algebra of Laurent polynomials in k variables where $k = \mathrm{rank}\, T$. The considerations on representations of T from the above paragraph give us an isomorphism

$$\mathbb{Z}[M^*_{\mathbb{Z}}] \xrightarrow{\cong} \mathbb{R}_{\mathbb{C}}(T),$$

with weights $w \in M^*_{\mathbb{Z}}$ mapping to their associated 1-dimensional representation.

The notation for elements of $\mathbb{Z}[M^*_{\mathbb{Z}}]$ can be a little confusing in the present situation, as for $v, w \in M^*_{\mathbb{Z}}$ we have distinct elements $[v] + [w]$ and $[v + w]$ and in fact $[v] \cdot [w] = [v + w]$. To help with this, it will be convenient to write e^v for $[v]$. Note that this is purely notational, the exponential doesn't really *mean* anything! With this notation, however, elements of $\mathbb{Z}[M^*_{\mathbb{Z}}]$ look like $\sum n_i e^{v_i}$ with $v_i \in M^*_{\mathbb{Z}}$, and the multiplication is determined by $e^v \cdot e^w = e^{v+w}$.

Let G be a compact, connected Lie group with maximal torus $T \hookrightarrow G$. Then any G-representation becomes a T-representation by restriction, and this gives us a map $\mathrm{res}\colon R_{\mathbb{C}}(G) \to R_{\mathbb{C}}(T)$. Comparing with the rings of class functions, we have

$$\begin{tikzcd}
R_{\mathbb{C}}(G) \arrow[r] \arrow[d, tail] & R_{\mathbb{C}}(T) \arrow[d, tail] \\
\mathrm{Cl}(G) \arrow[r] & \mathrm{Cl}(T).
\end{tikzcd}$$

We have already remarked that the vertical maps are injective. It is a fact from Lie theory that G is covered by the conjugates of T, and from this it follows at once that $\mathrm{Cl}(G) \to \mathrm{Cl}(T)$ is injective. We deduce that $R_{\mathbb{C}}(G) \to R_{\mathbb{C}}(T)$ is injective as well. In other words, representations of G are uniquely determined by their restrictions to T.

Let $W = N_G(T)/T$ be the Weyl group (the quotient of the normalizer of T by T). Each element $n \in W$ gives a group homomorphism $C_n\colon G \to G$ via $g \mapsto ngn^{-1}$, which restricts to a homomorphism $T \to T$. These homomorphisms induce horizontal maps

$$\begin{tikzcd}
R_{\mathbb{C}}(G) \arrow[r, "C_n"] \arrow[d, "\mathrm{res}"'] & R_{\mathbb{C}}(G) \arrow[d, "\mathrm{res}"'] \\
R_{\mathbb{C}}(T) \arrow[r, "C_n"] & R_{\mathbb{C}}(T).
\end{tikzcd}$$

The top map is readily checked to be the identity: restricting a representation along an inner automorphism always gives an isomorphic representation. The maps C_n give an action of W on $R_{\mathbb{C}}(T)$, and the above square shows that the restriction $R_{\mathbb{C}}(G) \to R_{\mathbb{C}}(T)$ has image inside of the ring of invariants:

$$\mathrm{res}\colon R_{\mathbb{C}}(G) \hookrightarrow R_{\mathbb{C}}(T)^W.$$

The following is a key result:

Theorem B.3.1. *For any compact Lie group G the map* $\mathrm{res}\colon R_{\mathbb{C}}(G) \hookrightarrow R_{\mathbb{C}}(T)^W$ *is an isomorphism.*

At this point we find ourselves wanting to understand $\mathbb{Z}[M^*_{\mathbb{Z}}]^W$, where the Weyl group W acts on T and therefore has an induced action on the the weight lattice $M^*_{\mathbb{Z}}$. The orbits $\mathcal{O}$ of W in $M^*_{\mathbb{Z}}$ are all finite, so let $S(\mathcal{O}) = \sum_{v \in \mathcal{O}} e^v$. We call $S(\mathcal{O})$ the "symmetric sum" corresponding to the orbit $\mathcal{O}$; it clearly lies in $\mathbb{Z}[M^*_{\mathbb{Z}}]^W$. It is easy to see that additively $\mathbb{Z}[M^*_{\mathbb{Z}}]^W$ is precisely the free abelian group generated by such symmetric sums, and we can write this fact as

$$\mathbb{Z}[M^*_{\mathbb{Z}}]^W \cong \mathbb{Z}\langle M^*_{\mathbb{Z}}/W\rangle$$

(one basis element for every element of the orbit space $M^*_{\mathbb{Z}}/W$). One final point about terminology: if $u \in M^*_{\mathbb{Z}}$ we will abuse notation somewhat by writing $S(u)$ for $S(W.u)$.

Let V be any representation of G. Then $\chi(V|_T) \in \mathbb{Z}[M_{\mathbb{Z}}^*]^W$ may be written uniquely as

$$\chi(V|_T) = \sum n_i S(\mathcal{O}_i)$$

for integers $n_i \in \mathbb{Z}_{>0}$ and orbits $\mathcal{O}_i \in M_{\mathbb{Z}}^*/W$. Each orbit $\mathcal{O}_i$ has a unique element ω_i lying in the positive cone C_+. In this way, every representation gives us a collection of positive weights and multiplicities.

One can define a partial ordering on the weights as follows. Say $\omega_1 \leq \omega_2$ if ω_1 lies inside the convex hull of the points in the orbit $W\omega_2$. This is clearly reflexive and transitive. It is not quite antisymmetric, but if $\omega_1 \leq \omega_2$ and $\omega_2 \leq \omega_1$ then $\omega_2 = \phi(\omega_1)$ for some $\phi \in W$. Note that a given pair of weights ω_1, ω_2 may or may not be comparable: it is not necessarily true that $\omega_1 \leq \omega_2$ or $\omega_2 \leq \omega_1$.

Theorem B.3.2. *Suppose that V is an irreducible representation of G.*

*(a) Among the associated weights of V there is one that is larger than all others with respect to $\leq$; this weight occurs with multiplicity one. It is called the **highest weight** (or **dominant weight**) of V.*

Said differently, there is a unique weight $\omega \in C_+$ such that

$$\chi(V|_T) = S(\omega) + (\textit{lower terms}).$$

(b) Every weight in C_+ occurs as the dominant weight for a unique irreducible representation. So there is a bijection between irreducible representations and weights lying in the positive cone.

Theorem B.3.3. *Assume that G is simply-connected, with simple roots $\theta_1, \dots, \theta_k$. Then there exists a unique collection of weights $\omega_1, \dots, \omega_k$ such that $\frac{\langle \theta_r, \omega_t \rangle}{\langle \theta_r, \theta_r \rangle} = \frac{1}{2}\delta_{r,t}$ for all r and t. We have that*

(1) $\omega_1, \dots, \omega_k$ is a $\mathbb{Z}$-basis for the weight lattice $M_{\mathbb{Z}}^*$, and

(2) The set of weights in the positive cone coincides with the free abelian semi-group generated by the ω_i's. That is, every weight in the positive cone can be written uniquely as $\sum n_i \omega_i$ with $n_i \in \mathbb{Z}_{\geq 0}$.

We call the $\omega_1, \dots, \omega_k$ in the above theorem the **simple weights** of G. Note that every simple root θ_r has the corresponding simple weight ω_r.

Corollary B.3.4. *Assume that G is simply-connected. Let $\rho_1, \dots, \rho_k$ be the irreducible representations whose highest weights are $\omega_1, \dots, \omega_k$. Then $\mathbb{Z}[G]$ is the polynomial ring $\mathbb{Z}[\rho_1, \dots, \rho_k]$.*

Remark B.3.5. If G is a simply-connected, compact Lie group then the following sets of objects are in bijective correspondence:

(1) The nodes of the Dynkin diagram of G;
(2) The simple roots $\theta_1, \dots, \theta_k$;
(3) The simple weights $\omega_1, \dots, \omega_k$;
(4) The irreducible representations $\rho_1, \dots, \rho_k$ that generate $\mathbb{Z}[G]$.

B.3.6. Examples in complex representation theory.

Example B.3.7. $G = SU(n)$. We will use all the notation and information from Example B.2.1. Our first task is to determine the simple weights $\omega_1, \dots, \omega_{n-1}$. If we write $\omega_1 = c_1\hat{e}_1 + \ldots + c_n\hat{e}_n$ with $\sum c_i = 0$ then $\langle \omega_1, \Theta_i \rangle = c_i - c_{i+1}$ for all i. So we need $c_1 - c_2 = 1$, $c_i - c_{i+1} = 0$ for $i > 1$, and $\sum c_i = 0$. It follows that

$$\omega_1 = \left(\tfrac{n-1}{n}\right)\hat{e}_1 - \tfrac{1}{n}\hat{e}_2 - \tfrac{1}{n}\hat{e}_3 - \cdots - \tfrac{1}{n}\hat{e}_n.$$

Using that $\hat{e}_n = -\big(\hat{e}_1 + \cdots + \hat{e}_{n-1}\big)$ we can also just write $\omega_1 = \hat{e}_1$.

Similar computations shows that

$$\omega_1 = \hat{e}_1, \quad \omega_2 = \hat{e}_1 + \hat{e}_2, \quad \omega_3 = \hat{e}_1 + \hat{e}_2 + \hat{e}_3, \quad \dots \quad \omega_{n-1} = \hat{e}_1 + \cdots + \hat{e}_{n-1}.$$

Recall that these weights are supposed to lie in the positive cone. As an exercise, check that

$$\omega_1 = \big(\tfrac{n-1}{n}\big)\Theta_1 + \big(\tfrac{n-2}{n}\big)\Theta_2 + \cdots \big(\tfrac{1}{n}\big)\Theta_{n-1}.$$

The fact that the coefficients are positive confirms that ω_1 is in the positive cone. Find similar formulas for the other simple weights.

We next compute the elements $S(\omega_i)$ in $\mathbb{Z}[M_{\mathbb{Z}}^*]^W$. It will be visually somewhat more pleasing if we introduce yet another notation and write $\xi_i = \hat{e}_i$. Recall that the weight lattice $M_{\mathbb{Z}}^*$ has basis $\xi_1, \dots, \xi_{n-1}$ and that ξ_n is shorthand for $-(\xi_1 + \cdots + \xi_{n-1})$. Also recall that the Weyl group $W = \Sigma_n$ acts on $M_{\mathbb{Z}}^*$ by permuation of the ξ's. Then

$$\begin{aligned} S(\omega_1) = S(\xi_1) &= e^{\xi_1} + e^{\xi_2} + \cdots + e^{\xi_n} \\ &= e^{\xi_1} + \cdots + e^{\xi_{n-1}} + e^{-(\xi_1 + \cdots + \xi_{n-1})} \end{aligned}$$

and

$$\begin{aligned} S(\omega_2) = S(\xi_1 + \xi_2) = \sum_{i<j} e^{\xi_i + \xi_j} &= \sum_{i<j} e^{\xi_i} e^{\xi_j} \\ &= \sum_{1 \le i < j < n} e^{\xi_i} e^{\xi_j} + \sum_i e^{-(\xi_1 + \cdots + \hat{\xi}_i + \cdots + \xi_{n-1})} \end{aligned}$$

(where here the hat on ξ_i indicates that that term is omitted).

In general, $S(\omega_k)$ is the kth elementary symmetric function in the terms $e^{\xi_1}, \dots, e^{\xi_n}$, with of course the usual proviso that ξ_n is shorthand for $-(\xi_1 + \cdots + \xi_{n-1})$.

Our final goal is to determine the fundamental representations $\rho_1, \dots, \rho_{n-1}$ corresponding to the simple weights $\omega_1, \dots, \omega_{n-1}$. Let V denote the standard representation of $SU(n)$ on $\mathbb{C}^n$. The character of this representation is clearly $e^{\xi_1} + \cdots + e^{\xi_n}$ which is $S(\omega_1)$, which confirms that this representation is irreducible with highest weight ω_1. One readily checks that for $k < n$ the character of $\Lambda^k V$ is the kth elementary symmetric function on the e^{ξ_i}'s, and so $\Lambda^k V$ is irreducible with highest weight ω_k. Our fundamental representations are therefore

$$\rho_1 = V, \quad \rho_2 = \Lambda^2 V, \quad \dots \quad \rho_{n-1} = \Lambda^{n-1} V$$

and we have

$$R_{\mathbb{C}}(SU(n)) = \mathbb{Z}[V, \Lambda^2 V, \dots, \Lambda^{n-1} V].$$

It is worth being very specific about what the representation theory of $SU(n)$ is like for low values of n, so we do this next.

Example B.3.8 (A closer look at $SU(2)$). For $SU(2)$ the rank is 1. The weight lattice is generated by an element ξ_1, and the roots of $SU(2)$ are $\pm 2\xi_1$. We have $R_{\mathbb{C}}(SU(2)) = \mathbb{Z}[V]$ where V is the standard representation of $SU(2)$ on $\mathbb{C}^2$. Unfortunately, this description of $R_{\mathbb{C}}(SU(2))$ does *not* tell us the irreducible representations of $SU(2)$. It tells us that they can all be found inside the tensor powers $V^{\otimes k}$, but it does not tell us exactly how to find them. So we have to work harder.

Let z_1 and z_2 be two formal variables, and let H_n denote the complex vector space of degree n homogeneous polynomials in z_1 and z_2. Elements of H_n look like

$$f(z_1, z_2) = a_0 z_1^n + a_1 z_1^{n-1} z_2 + a_2 z_1^{n-2} z_2^2 + \cdots + a_n z_2^n.$$

Note that $\dim H_n = n + 1$. Elements of H_n are certain functions $\mathbb{C}^2 \to \mathbb{C}$, and we make $SU(2)$ act on them via the natural action of $SU(2)$ on $\mathbb{C}^2$. That is, if $X \in SU(2)$ then

$$(X.f)(z) = f(X^{-1}z).$$

Note that

$$[X.(Yf)](z) = (Yf)(X^{-1}z) = f(Y^{-1}(X^{-1}z)) = f((XY)^{-1}z) = [(XY).f](z)$$

and so this really is an action.

Let us compute the character of H_n. The monomial functions $z_1^{n-k} z_2^k$ generate 1-dimensional representations of the torus, with $\begin{bmatrix} \omega & 0 \\ 0 & \omega^{-1} \end{bmatrix}$ acting via ω^{n-2k}. It follows that

$$\chi(H_n) = e^{n\xi_1} + e^{(n-2)\xi_1} + e^{(n-2)\xi_1} + \cdots e^{-n\xi_1}.$$

So H_n is the irreducible representation with highest weight $n\xi_1$, and therefore

$$\mathbb{C} = H_0,\ H_1,\ H_2,\ H_3,\ \ldots$$

is a complete list of the irreducible representations of $SU(2)$.

Note that we now have two bases for $R_{\mathbb{C}}(SU(2))$: the basis consisting of the irreducible representations H_n, and the basis consisting of the powers of V. It is perhaps useful to think a little about how these two bases relate to each other. Clearly $V = H_1$ (as both are irreducible representations of dimension 2, and only one such representation exists). The character of $V^{\otimes 2}$ is

$$\begin{aligned} \chi(V^{\otimes 2}) = \chi(V)^2 = (e^{\xi_1} + e^{-\xi_1})^2 = e^{2\xi_1} + 2 + e^{-2\xi_1} &= (e^{2\xi_1} + 1 + e^{-2\xi_1}) + 1 \\ &= \chi(H_2) + \chi(\mathbb{C}). \end{aligned}$$

This immediately tells us that $V^{\otimes 2} \cong H_2 \oplus \mathbb{C}$. Similar computations work for the higher tensor powers; for instance,

$$\chi(V^{\otimes 3}) = e^{3\xi_1} + 3e^{\xi_1} + 3e^{-\xi_1} + e^{-3\xi_1} = \chi(H_3) + 2\chi(H_1)$$

and hence $V^{\otimes 3} \cong H_3 \oplus H_1 \oplus H_1$.

A natural question is whether one can see these decompositions in a more concrete way, without recourse to character calculations. This gets into some interesting representation theory, and mostly we won't pursue it here, but let us at least remark on the first instance of this. We saw above that $V^{\otimes 2}$ contains a trivial representation, but how can we write down the explicit one-dimensional subspace that is fixed by the action?

Consider the antisymmetric form on V given by $\langle v, w \rangle = \det[v|w]$ (where elements $v \in V = \mathbb{C}^2$ are identified with column vectors). Then $SU(2)$ preserves this form, since $\langle Xv, Xw \rangle = \det[Xv|Xv] = \det(X \cdot [v|w]) = \det(X)\det[v|w]$ and $\det(X) = 1$ for $X \in SU(2)$. Regarding our form as a map $V \otimes V \to \mathbb{C}$, we see that this is a surjective map of representations: so the trivial representations is a quotient, and therefore a summand, of $V^{\otimes 2}$. This approach doesn't give the splitting for free, but one can easily guess it: the element $1 \in \mathbb{C}$ should be sent to $e_1 \otimes e_2 - e_2 \otimes e_1$. One readily checks that this antisymmetric tensor is fixed by the $SU(2)$ action, and therefore generates our copy of $\mathbb{C}$ inside $V \otimes V$. Now having guessed this, though, we can engineer an even better explanation. Because the space of antisymmetric 2-tensor is precisely $\Lambda^2 V$, and the quotient of $V \otimes V$ by these antisymmetric tensors is $\mathrm{Sym}^2 V$. That is, we have an exact sequence

$$0 \to \Lambda^2 V \to V \otimes V \to \mathrm{Sym}^2 V \to 0$$

and this is clearly an exact sequence of $SU(2)$-represenations. Since V is 2-dimensional $\Lambda^2 V$ is the determinant representation of $SU(2)$, which is trivial. So $V^{\otimes 2} \cong \mathbb{C} \oplus \mathrm{Sym}^2(V)$. The representation $\mathrm{Sym}^2(V)$ is essentially H_2 (although technically $H_2 = \mathrm{Sym}^2(V^*)$, but one readily checks that $V \cong V^*$ as representations by using the adjoint of the determinant form $V \otimes V \to \mathbb{C}$).

As one final example concerning $SU(2)$, let us consider the adjoint action of $SU(2)$ on its complexfied tangent space $\mathbb{C} \otimes T_I SU(2)$. As we have seen before (????), the character of this representation is

$$\chi(\mathbb{C} \otimes T_I SU(2)) = e^{2\xi_1} + 1 + e^{-2\xi_1}.$$

It follows that $\mathbb{C} \otimes T_I SU(2) \cong H_2$, the irreducible representation of dimension three.

Example B.3.9 (A closer look at $SU(3)$). Here our group is rank 2. The weight lattice is generated by ξ_1 and ξ_2, and the roots are $\pm(\xi_1 - \xi_2)$, $\pm(\xi_1 - \xi_3)$, and $\pm(\xi_2 - \xi_3)$ where ξ_3 is shorthand for $-(\xi_1 + \xi_2)$. Thus we can also say that the roots are $\pm(\xi_1 - \xi_2)$, $\pm(2\xi_1 + \xi_2)$, and $\pm(\xi_1 + 2\xi_2)$. The simple roots are

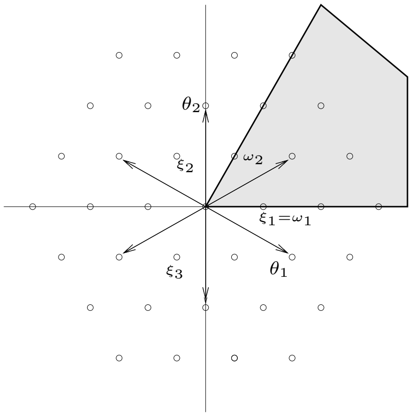
$$\theta_1 = \xi_1 - \xi_2 \qquad \text{and} \qquad \theta_2 = \xi_2 - \xi_3 = \xi_1 + 2\xi_2,$$

and the simple weights are

$$\omega_1 = \xi_1, \qquad \omega_2 = \xi_1 + \xi_2.$$

We know that $R_{\mathbb{C}}(SU(3)) = \mathbb{Z}[V, \Lambda^2 V]$. Again, this description by itself does not tell us much about the irreducible representations. We do know that there will be one irreducible representation with highest weight vector $k\omega_1 + n\omega_2$ for any $k, n \in \mathbb{Z}_{\geq 0}$, and that these will be *all* the irreducible representations.

The following picture shows the weight lattice of $SU(3)$ (the hexagonal lattice in the background). It also shows the root system, and the shaded region is the fundamental dual Weyl chamber.

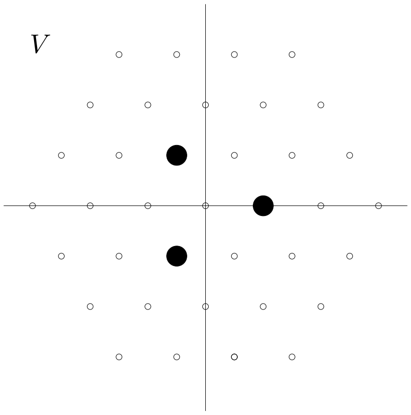


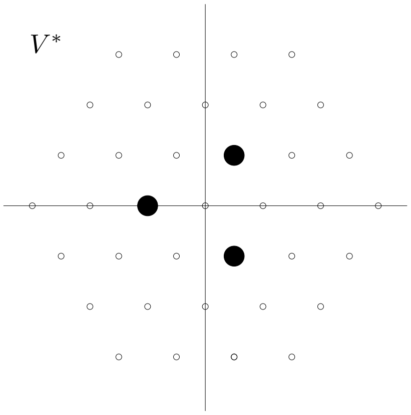
The Weyl group is Σ_3 and is generated by the reflections in the lines $\langle \xi_1 + \xi_2 \rangle$, $\langle \xi_1 + \xi_3 \rangle$, and $\langle \xi_2 + \xi_3 \rangle$.

The characters of V and V^* are

$$\chi(V) = e^{\xi_1} + e^{\xi_2} + e^{\xi_3} \qquad \text{and} \qquad \chi(V^*) = e^{-\xi_1} + e^{-\xi_2} + e^{-\xi_3}.$$

We can also depict these by drawing weight diagrams, as follows:





Note that the weights of $\Lambda^2(V)$ coincide with those of V^*, showing that these two representations are isomorphic.

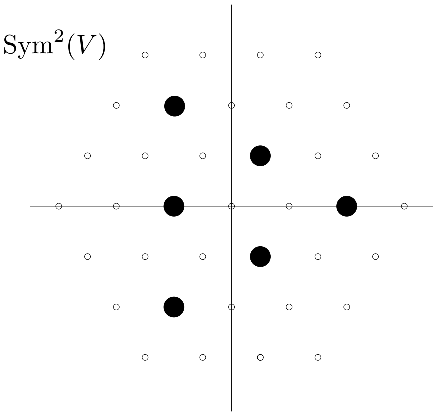
We know that there will be an irreducible representation $\Gamma_{a,b}$ with highest weight $a\omega_1 + b\omega_2$, for each $a, b \in \mathbb{Z}_{\geq 0}$. So far we have constructed $\Gamma_{1,0} = V$ and $\Gamma_{0,1} = V^*$. To construct others, note that if W and W' are representations then the weights of $W \otimes W'$ are the set of weights $w \otimes w'$ where w is a weight of W and w' is a weight of w'. This shows immediately that $V \otimes V$ will have highest weight $2\omega_1$; but just as we saw in the $SU(2)$ example, $V \otimes V$ will not be irreducible. There is an exact sequence

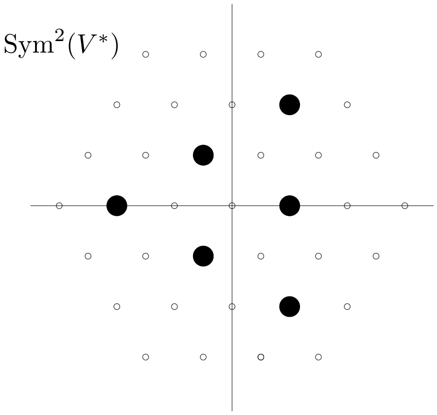
$$0 \to \Lambda^2 V \to V \otimes V \to \operatorname{Sym}^2(V) \to 0.$$

So let us instead focus on $\operatorname{Sym}^2(V)$. The weights are readily computed, and we find that

$$\chi(\operatorname{Sym}^2 V) = e^{2\xi_1} + e^{2\xi_2} + e^{2\xi_3} + e^{\xi_1+\xi_2} + e^{\xi_1+\xi_3} + e^{\xi_2+\xi_3}.$$

A similar computation holds for $\operatorname{Sym}^2(V^*)$, where the only change is the sign on all the exponents. The results are again best depicted in terms of diagrams:

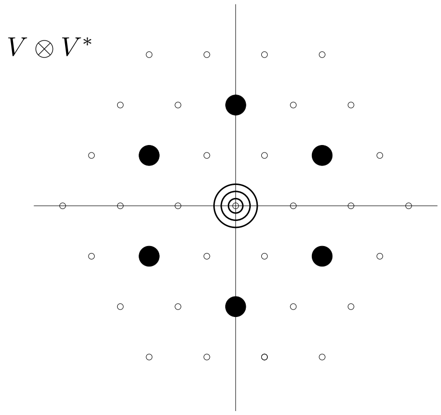




One observes, of course, that these weight diagrams all have a high degree of symmetry—and we knew this already, as they must be symmetric under the action of the Weyl group.

At this point we have constructed irreducible representations whose highest weights are ω_1, ω_2, $2\omega_2$, and $2\omega_1$. They have dimensions 3, 3, 6, and 6 (respectively), and they come in dual pairs of W and W^*.

To get a representation whose highest weight is $\omega_1 + \omega_2$ one naturally is led to look at $V \otimes V^*$, which has the following weight diagram:



Here the three circles at the origin indicate that this weight occurs with multiplicity three. The representaton $V \otimes V^*$ is *not* irreducible, as one readily sees by considering the evaluation map $V \otimes V^* \to \mathbb{C}$. This is a surjective map of representations, which shows that $\mathbb{C}$ splits off $V \otimes V^*$. The remaining eight-dimensional representation *is* irreducible, although it takes some work to see this. It coincides with the adjoint representation of $SU(3)$ on its Lie algebra $T_ISU(3)$, as one readily sees from the root computation (the two weights that are zero correspond to the maximal torus of $SU(3)$ acting trivially on the torus's Lie algebra).

In general, to construct an irreducible representation with highest weight $a\omega_1 + b\omega_2$ one can start with the representation $\mathrm{Sym}^a(V) \otimes \mathrm{Sym}^b(V^*)$. This representation has the correct highest weight, but is not necessarily irreducible. There is a map

$$\mathrm{Sym}^a(V) \otimes \mathrm{Sym}^b(V^*) \to \mathrm{Sym}^{a-1}(V) \otimes \mathrm{Sym}^{b-1}(V^*)$$

given by

$$(v_1 \otimes \cdots \otimes v_a) \otimes (\beta_1 \otimes \cdots \beta_b) \mapsto \sum_{i,j} \beta_i(v_j)(v_1 \otimes \cdots \otimes \hat{v}_i \otimes \cdots v_a) \otimes (\beta_1 \otimes \cdots \hat{\beta}_j \otimes \cdots \beta_b).$$

This map turns out to be surjective, and its kernel $\Gamma_{a,b}$ clearly has highest weight $a\omega_1 + b\omega_2$ (because this weight is too small to appear in $\mathrm{Sym}^{a-1}(V) \otimes \mathrm{Sym}^{b-1}(V^*)$). It is far from obvious, but $\Gamma_{a,b}$ turns out to be irreducible. See [**FuH**, Section 13.2] for more details about this. The complete details are not crucial for us, as these $\Gamma_{a,b}$'s will not be used in the main text.

B.3.10. Real representations. Let $R_{\mathbb{R}}(G)$ represent the real representation ring of G. As a group, it is the free abelian group on the irreducible representations of G on real vector spaces. We would like to understand these real representations, just as we have understood the complex ones. It turns out that one cannot understand the real representations without also studying the *quaternionic* representations—the representations of G on vector spaces over $\mathbb{H}$. Let $R_{\mathbb{H}}(G)$ denote the Grothendieck group of such representations, which is again the free abelian group generated by the irreducible quaternionic representations. (Note that $\mathbb{R}_{\mathbb{H}}(G)$ is not a ring due to the lack of a tensor product for quaternionic vector spaces). Our goal is to understand both $R_{\mathbb{R}}(G)$ and $R_{\mathbb{H}}(G)$ by comparing them to $R_{\mathbb{C}}(G)$.

We will need the following functors on representations:

$$\begin{tikzcd}
\mathrm{Rep}_{\mathbb{R}}(G) \arrow[r, shift left, "c"] & \mathrm{Rep}_{\mathbb{C}}(G) \arrow[l, shift left, "r"] \arrow[r, shift left, "q"] \arrow[loop above, "t"] & \mathrm{Rep}_{\mathbb{H}}(G). \arrow[l, shift left, "r_{\mathbb{H}}"]
\end{tikzcd}$$

Here $c(V) = V \otimes_{\mathbb{R}} \mathbb{C}$ and $q(V) = V \otimes_{\mathbb{C}} \mathbb{H}$. The functors r and $r_{\mathbb{H}}$ are restriction functors, or forgetful functors: if V is a complex representation then rV is the same set, with the same G-action, but regarded as a real vector space. Similarly, $r_{\mathbb{H}}$ takes a quaternionic representation to its underlying complex representation. Finally, if V is a complex representation then tV is its conjugate: the same set, and same G-action, but given the conjugate complex structure.

The following relations between these functors are readily checked (we write the relations in terms of isomorphism classes of representations):

$$\begin{gathered}
rc = 2, \quad tc = c, \quad cr = 1 + t, \quad rt = r; \qquad t^2 = 1; \\
r_{\mathbb{H}}q = 1 + t, \quad qr_{\mathbb{H}} = 2, \quad tr_{\mathbb{H}} = r_{\mathbb{H}}, \quad qt = q.
\end{gathered}$$

These relations mean, for instance, that if V is a real representation the $rc(V) \cong V \oplus V$, whereas if W is a complex representation then $cr(W) \cong W \oplus tW$.

The above functors induce maps between the Grothendieck groups:

$$\begin{tikzcd}
R_{\mathbb{R}}(G) \arrow[r, shift left, "c"] & R_{\mathbb{C}}(G) \arrow[l, shift left, "r"] \arrow[r, shift left, "q"] \arrow[loop above, "t"] & R_{\mathbb{H}}(G). \arrow[l, shift left, "r_{\mathbb{H}}"]
\end{tikzcd}$$

Here both c and t are ring homomorphisms, but r is not (and recall that $R_{\mathbb{H}}(G)$ doesn't have a ring structure at all). The relations $rc = 2$ and $qr_{\mathbb{H}} = 2$ show that c and $r_{\mathbb{H}}$ are injective.

Note that of the functors listed above, only t is guaranteed to send irreducibles to irreducibles (which follows because $t^2 = 1$).

Complex representations that are in the image of r or $r_{\mathbb{H}}$ are called **real-type** and **quaternionic-type**, respectively. If a complex representation V satisfies $tV \cong V$ we say it is **self-conjugate**. Note that both real-type and quaternionic-type representations are self-conjugate. It is a theorem that every irreducible, self-conjugate representation is either real-type or quaternionic-type, but not both.

Remark B.3.11. It is commonplace to shorten "real-type" to just "real", and "quaternionic-type" to just "quaternionic". This creates some confusion about what one means by the phrase "real representation": does it mean a representation on a real vector space, or does it mean a complex represenation of real type? While the answer is usually clear from context, it does create some confusion when one wants to talk about all these things at once. We will therefore stick to the longer, but less ambiguous, adjectives.

Here is a useful result:

Lemma B.3.12. *Let V be a complex representation of G.*

(a) V is real-type (resp. quaternionic-type) if and only if V admits a symmetric (resp. skew-symmetric) bilinear form $V \otimes V \to \mathbb{C}$ that is invariant under the action of G.

(b) V is real-type (resp. quaternionic-type) if and only if V admits a G-equivariant, conjugate-linear, self-map $j\colon V \to V$ such that $j^2 = 1$ (resp. $j^2 = -1$). Such a j is called a ***structure map*** *of real- or quaternionic- type.*

Example B.3.13. Consider the standard representation of $SU(2)$ on $V = \mathbb{C}^2$. The bilinear form on V given by $\langle v, w\rangle = \det[v|w]$ (with v and w regarded as column vectors) is skew-symmetric and invariant under the $SU(2)$-action. Therefore V is a quaternionic representation. Structure map ????

Write $\operatorname{Irr}(G;\mathbb{C})$ for the isomorphism classes of irreducible complex representations. Write $\operatorname{Irr}(G;\mathbb{C})_{\mathbb{R}}$ and $\operatorname{Irr}(G;\mathbb{C})_{\mathbb{H}}$ for the isomorphism classes of irreducible real-type and quaternionic-type representations. Finally, let $\operatorname{Irr}(G;\mathbb{C})_{\mathbb{C}}$ denote the isomorphism classes of irreducible representations that are not self-conjugate. Such isomorphism classes come in conjugate pairs, so write $\frac{1}{2}\operatorname{Irr}(G;\mathbb{C})_{\mathbb{C}}$ for any subset which contains exactly one isomorphism class from each conjugate pair. One has a disjoint union

$$\operatorname{Irr}(G;\mathbb{C}) = \operatorname{Irr}(G;\mathbb{C})_{\mathbb{R}} \amalg \operatorname{Irr}(G;\mathbb{C})_{\mathbb{C}} \amalg \operatorname{Irr}(G;\mathbb{C})_{\mathbb{H}}.$$

The first and third components together equal the self-conjugate representations.

Let $V \in \operatorname{Irr}(G;\mathbb{R})$. Then cV is self-conjugate, and so it can be written as

$$cV = (U_1 \oplus tU_1) \oplus \cdots \oplus (U_k \oplus tU_k) \oplus (W_1 \oplus \cdots \oplus W_s)$$

where each U_i is in $\operatorname{Irr}(G;\mathbb{C})_{\mathbb{C}}$ and W_i is in $\operatorname{Irr}(G;\mathbb{C})_{\mathbb{R}} \cup \operatorname{Irr}(G;\mathbb{C})_{\mathbb{H}}$. Applying r gives that

$$2V = rcV = 2(rU_1) \oplus \cdots \oplus 2(rU_k) \oplus rW_1 \oplus \cdots \oplus rW_s.$$

But V is irreducible, and so the number of summands on the right hand side can be at most two: (k, s) is either $(1, 0)$, $(0, 1)$, or $(0, 2)$. In the first case we conclude that $V = rU_1$; in the second, cV is irreducible and therefore lies in $\operatorname{Irr}(G;\mathbb{C})_{\mathbb{R}}$. In the final case we have $cV = W_1 \oplus W_2$ where W_i is self-conjugate, and $V = rW_1 = rW_2$. Note that if Z is a real-type representation then rZ is necessarily reducible (because $rc = 2$), hence neither W_1 nor W_2 are real-type; that is, they are quaternionic type.

The above paragraph gives three possibilities for what an object in $\operatorname{Irr}(G;\mathbb{R})$ can look like. More of these kinds of arguments—playing around with the functors r, c, t, $r_{\mathbb{H}}$, and q—easily yield the following result:

Theorem B.3.14. *The set $\operatorname{Irr}(G;\mathbb{R})$ is the disjoint union of the three pieces*

(1) $\{V \in \operatorname{Irr}(G;\mathbb{R}) \,|\, cV \in \operatorname{Irr}(G;\mathbb{C})_{\mathbb{R}}\}$

(2) $\{rW \,|\, W \in \operatorname{Irr}(G;\mathbb{C})_{\mathbb{C}}\}$

(3) $\{rW \,|\, W \in \operatorname{Irr}(G;\mathbb{C})_{\mathbb{H}}\}$.

(Note that it is not immediately obvious that sets (2) and (3) are subsets of $\operatorname{Irr}(G;\mathbb{R})$ *in the first place). Likewise, the set* $\operatorname{Irr}(G;\mathbb{H})$ *is the disjoint union of the three pieces*

(1) $\{V \in \operatorname{Irr}(G;\mathbb{H}) \,|\, r_{\mathbb{H}}V \in \operatorname{Irr}(G;\mathbb{C})_{\mathbb{H}}\}$

(2) $\{qW \,|\, W \in \operatorname{Irr}(G;\mathbb{C})_{\mathbb{C}}\}$

(3) $\{qW \,|\, W \in \operatorname{Irr}(G;\mathbb{C})_{\mathbb{R}}\}$.

Our next goal is to understand real-type and quaternionic-type representations in terms of their weights. Let K denote the set of weights in the fundamental Weyl chamber. We know that there is a bijection between K and the set of irreducible complex representations; if $\mu \in K$ then let V_μ denote the irreducible complex representaton with higest weight μ. We also know that K is the free semi-group on the fundamental weights $\omega_1, \dots, \omega_k$. Our goal is to understand which weights in K correspond to elements of $\operatorname{Irr}(G;\mathbb{C})_\mathbb{R}$ and which weights correspond to elements of $\operatorname{Irr}(G;\mathbb{C})_\mathbb{H}$.

We can define a $\mathbb{Z}/2$-action on K as follows. If $\mu \in K$, let $\overline{\mu}$ denote the highest weight in the irreducible representation tV_μ. Let us also say that a weight $\mu \in K$ is self-conjugate, real-type, complex-type, or quaternionic-type if the corresponding representation V_μ is so. Finally, for a self-conjugate weight μ define the **index** of μ to be 1 if μ is real-type, and -1 is μ if quaternionic-type.

THEOREM B.3.15. *Let $\mu \in K$.*

(a) μ is self-conjugate if and only if $-\mu \in W.\mu$ (where $W.\mu$ is the orbit of μ under the Weyl group).
(b) Let $\mu, \nu \in K$ be self-conjugate, with indices ϵ and ϵ'. Then $\mu + \nu$ is self-conjugate and has index equal to $\epsilon\epsilon'$.
(c) For any $\mu \in K$, the weight $\mu + \overline{\mu}$ is self-conjugate and real-type.

EXAMPLE B.3.16 ($SU(2)$ and $SU(3)$). For $SU(2)$ the weight lattice is $\mathbb{Z}\langle\xi_1\rangle$, so that $K = \{0, \xi_1, 2\xi_1, 3\xi_1, \dots\}$. The representation V_{ξ_1} has character $e^{\xi_1} + e^{-\xi_1}$ and so is clearly self-conjugate. The $\mathbb{Z}/2$-action on K therefore fixes ξ_1, and hence fixes all of K: all weights are self-conjugate.

We also know that V_{ξ_1} equals the standard representation of $SU(2)$ on $\mathbb{C}^2$, and we saw in Example B.3.13 that this is quaternionic-type. It follows from Theorem B.3.15(b) that $2\xi_1$ is real-type, $3\xi_1$ is quaternionic-type, and so forth. So we have

$$\begin{aligned}&\operatorname{Irr}(SU(2);\mathbb{C})_\mathbb{R} = \{\mathbb{C}, H_2, H_4, H_6, \dots\}, \qquad \operatorname{Irr}(SU(2);\mathbb{C})_\mathbb{C} = \emptyset,\\ &\operatorname{Irr}(SU(2);\mathbb{C})_\mathbb{H} = \{H_1, H_3, H_5, \dots\}.\end{aligned}$$

For $SU(3)$ the weight lattice is $\mathbb{Z}\langle\xi_1, \xi_2\rangle$ and K is generated by $\omega_1 = \xi_1$ and $\omega_2 = \xi_1 + \xi_2$. Recall that V_{ω_1} is the standard representation on $\mathbb{C}^3$ and V_{ω_2} is its dual. This shows that $\overline{\omega}_1 = \omega_2$, and by linearity this determines the $\mathbb{Z}/2$-action on all of K. In particular, the self-conjugate weights are those of the form $n(\omega_1 + \omega_2)$ for $n \geq 0$; all of the other weights are complex-type.

But $\omega_1 + \omega_2 = \omega_1 + \overline{\omega}_1$, and so by Theorem B.3.15(c) this is real-type. By part (b) of the same theorem, each multiple $n(\omega_1 + \omega_2)$ is also real-type.

To summarize:

$$\begin{aligned}&\operatorname{Irr}(SU(3);\mathbb{C})_\mathbb{R} = \{V_{n(\omega_1+\omega_2)} \,|\, n \geq 0\}\\ &\operatorname{Irr}(SU(3);\mathbb{C})_\mathbb{C} = \{V_{p\omega_1+q\omega_2} \,|\, p, q \geq 0,\ p \neq q\}\\ &\operatorname{Irr}(SU(3);\mathbb{C})_\mathbb{H} = \emptyset.\end{aligned}$$

Returning to our discussion of general compact Lie groups G, let $\omega_1, \dots, \omega_k$ be a fundamental system of weights for G. It is clear that the $\mathbb{Z}/2$-action on K must permute these ω_i's. Write

$$\{\omega_1, \dots, \omega_k\} = \{\nu_1, \dots, \nu_r\} \amalg \{\sigma_1, \overline{\sigma}_1, \dots, \sigma_s, \overline{\sigma}_s\}.$$

where $\nu_i = \overline{\nu}_i$ for all i. Then the self-conjugate weights are precisely the ones of the form

$$\mu = \sum_i n_i(\sigma_i + \overline{\sigma}_i) + \sum_j m_j \nu_j.$$

If ϵ_j denotes the index of ν_j, then Theorem B.3.15(b,c) tell us that the index of μ is $\prod_j \epsilon_j^{m_j}$. So we can readily identify the type of every weight in K. Here is a brief summary of this:

Corollary B.3.17. *Let $Q \subseteq \{1, \dots, k\}$ be the set of indices i such that ω_i is self-conjugate and quaternionic-type. If $\mu = \sum_i n_i\omega_i$ with $n_i \in \mathbb{Z}_{\geq 0}$ and μ is self-conjugate, then the index of μ is 1 or -1 depending on whether $\sum_{i\in Q} n_i$ is even or odd.*

We close this section with a discussion of how to determine $R_{\mathbb{R}}(G)$ and $R_{\mathbb{H}}(G)$. These are embedded into $R_{\mathbb{C}}(G)$ via the maps

$$R_{\mathbb{R}}(G) \xrightarrow{c} R_{\mathbb{C}}(G) \xleftarrow{r_{\mathbb{H}}} R_{\mathbb{H}}(G).$$

The image of c is generated, as an abelian group, by the set

$$\mathrm{Irr}(G;\mathbb{C})_{\mathbb{R}} \cup \{2W \mid W \in \mathrm{Irr}(G;\mathbb{C})_{\mathbb{C}}\} \cup \{2W \mid W \in \mathrm{Irr}(G;\mathbb{C})_{\mathbb{H}}\}.$$

??????

B.4. The groups $SO(n)$ and $\mathrm{Spin}(n)$.

Recall that for $n > 2$ one has $\pi_1 SO(n) \cong \mathbb{Z}/2$, and that $\mathrm{Spin}(n)$ is the universal cover of $SO(n)$. These two Lie groups therefore share the same Lie algebra, and in particular the same root system. However: the representation theory of the two groups, although intricately linked, is somewhat different. Recall that many of our general results on representation theory assumed that the group G was simply-connected, so those results will only apply to $\mathrm{Spin}(n)$.

The material in this section is organized as follows. We first consider only the groups $SO(n)$, ignoring their double covers. The analysis of these groups is different when n is even or odd. We discuss the root systems in each case, then discuss what we can about the representation theory. Only afterwards do we turn our attention to $\mathrm{Spin}(n)$, again considering the odd and even cases separately.

B.4.1. Preliminaries. We begin with some preliminaries. For any $\alpha \in \mathbb{R}$ write

$$R_\theta = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}.$$

The set of matrices R_θ constitutes the group $SO(2)$ (which is isomorphic to S^1, of course). Consider the action of this group on $M = M_{2\times 2}(\mathbb{R})$ given by conjugation: $A.X = AXA^{-1}$ for $A \in SO(2)$ and $X \in M$. If we equip M with the standard inner product $\langle X, Y\rangle = \sum x_{ij}y_{ij}$, then this is invariant under the $SO(2)$-action. The subspace of symmetric matrices is a subrepresentation of M, the subspace of antisymmetric matrices is another subrepresentation, and the subspace of trace zero matrices is yet another subrepresentation. Taking intersections of these readily yields the following decomposition of M into irreducible $SO(2)$-representations:

$$M = \langle I\rangle \oplus \langle J\rangle \oplus \langle K, L\rangle$$

where

$$J = R_{\pi/2} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \qquad K = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \qquad L = JK = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

(and I is the identity matrix). Note that $\langle K, L\rangle$ is the set of trace zero, symmetric matrices, and $\langle J\rangle$ is the set of antisymmetric matrices.

B.4.2. The root systems. First consider the group $SO(n)$ where $n = 2k$. The maximal torus T is the set of block matrices

$$\Omega = \begin{bmatrix} R_{\alpha_1} & 0 & 0 & \cdots & 0 \\ 0 & R_{\alpha_2} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & & 0 \\ 0 & 0 & 0 & \cdots & R_{\alpha_k} \end{bmatrix}$$

where $\alpha_1, \dots, \alpha_k \in \mathbb{R}$. This is clearly a cartesian product of k copies of $SO(2)$, and hence a torus as claimed.

The tangent space $\mathfrak{g} = T_I SO(2k)$ is the set of $(2k) \times (2k)$ antisymmetric matrices, which has dimension $\binom{2k}{2}$. For $B \in M_{2\times 2}(\mathbb{R})$ write $X_{rs}(B)$ for the matrix with B in the (r, s) block and $-B^T$ in the (s, r) block, and zeros elsewhere. For $r < s$ write M_{rs} for the space of matrices $X_{rs}(B)$ with B arbitrary, and write M_{rr} for the space of matrices $X_{rr}(B)$ with B antisymmetric. Then $\mathfrak{g}$ is the direct sum of the M_{rs} subspaces for $r \leq s$.

We need to analyze the adjoint action of T on $\mathfrak{g}$. If Ω is an element of T as above, then

$$\Omega.X_{rs}(B) = \Omega X_{rs}(B)\Omega^{-1} = X_{rs}(R_{\theta_r} B R_{-\theta_s}).$$

So each M_{rs} is a T-subrepresentation of $\mathfrak{g}$. Note that each M_{rr} is one-dimensional and is therefore the trivial representation (a fact that is also readily checked by hand). We claim that each M_{rs} splits into two 2-dimensional irreducible representations, determined by the maps $T \to SO(2)$ given by $\Omega \mapsto R_{\alpha_r - \alpha_s}$ and $\Omega \mapsto R_{\alpha_r + \alpha_s}$. The former is the submodule spanned by $X_{rs}(I)$ and $X_{rs}(J)$, and this is easy to see because both I and J commute with all elements of $SO(2)$. The latter representation is the submodule spanned by $X_{rs}(K)$ and $X_{rs}(L)$; this is seen using that $KR_\alpha = R_{-\alpha}K$ for all angles α, and that $L = JK$.

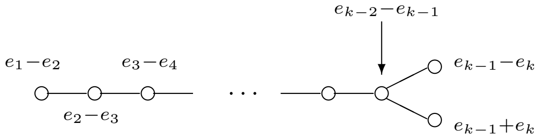
If $e_1, \dots, e_k$ are the ???? then our roots are

$$\theta_{rs} = e_r - e_s \qquad \text{and} \qquad \theta'_{rs} = e_r + e_s$$

(and their negatives), for $r < s$. The positive roots are the ones just listed, and the simple roots are

$$e_1 - e_2, \quad e_2 - e_3, \quad \dots, \quad e_{k-1} - e_k, \quad \text{and} \quad e_{k-1} + e_k.$$

Computing the inner products between them, one readily finds that the root system is of type D_k:



Next we turn to the case of $SO(2k+1)$. Here the maximal torus T is the set of matrices

$$\Omega = \begin{bmatrix} R_{\alpha_1} & O & O & \cdots & O & 0 \\ O & R_{\alpha_2} & O & \cdots & O & 0 \\ \vdots & \vdots & \ddots & & O & 0 \\ O & O & O & \cdots & R_{\alpha_k} & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}$$

where each O represents a 2×2 block matrix of zeros, and each 0 represents either a 2×1 or 1×2 matrix of zeros. The Lie algebra $\mathfrak{g}$ is again the space of $(2k+1) \times (2k+1)$ antisymmetric matrices. Define $X_{rr}(B)$ and $X_{rs}(B)$ as before, for $B \in M_{2\times 2}(\mathbb{R})$. If $b \in \mathbb{R}^2$ define $X_i(b)$ to be the matrix with the column vector b in the $(2i-1, 2k+1)$- and $(2i, 2k+1)$-entries, and the row vector $-b^T$ in the $(2k+1, 2i-1)$- and $(2k+1, 2i)$-entries; here $1 \leq i \leq k$. Write M_{rs} for the space of matrices $X_{rs}(B)$ with B arbitrary, write M_{rr} for the space of matrices $X_{rr}(B)$ with B antisymmetric, and write M_i for the space of matrices $M_i(b)$ with $b \in \mathbb{R}^2$. Then $\mathfrak{g}$ is the direct sum of the M_{rs} spaces for $1 \leq r \leq s \leq k$ and the M_i spaces for $1 \leq i \leq k$.

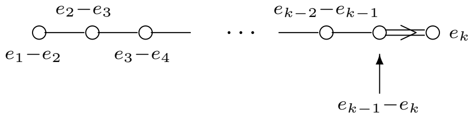
In computing the adjoint action of T on $\mathfrak{g}$, one finds that the action on the M_{rs} spaces is exactly the same as in the case of $SO(2k)$. One also finds that T acts on M_i via

$$\Omega \cdot X_i(b) = \Omega X_i(b) \Omega^{-1} = X_i(R_{\alpha_i} b).$$

So M_i is an irreducible subrepresentation with character $\Omega \mapsto R_{\alpha_i}$. It follows that the roots of $SO(2k+1)$ are $e_i - e_j$, $e_i + e_j$ $(i < j)$, and e_i (and their negatives). We take the positive roots to be $e_i - e_j$ for $i < j$, $e_i + e_j$ for $i < j$, and all the e_i's. The simple roots are then

$$e_1 - e_2, \quad e_2 - e_3, \quad \ldots, \quad e_{k-1} - e_k, \quad \text{and} \quad e_k.$$

Note that these all have length $\sqrt{2}$ except for e_k, which is shorter. One readily computes the inner products and finds that the Dynkin diagram is type B_k, as shown below:



B.4.3. Representation theory of the groups $SO(n)$. If $SO(2k)$ were simply-connected then the first thing we would do is to determine the fundamental weights $\omega_1, \ldots, \omega_k$. It is informative to *attempt* to do this anyway: if $\theta_i = e_i - e_{i+1}$ for $i < k$ and $\theta_k = e_{k-1} + e_k$, then one must solve the system of equations $\langle \theta_r, \omega_s \rangle = \delta_{rs}$. One finds that

$$\omega_1 = e_1, \qquad \omega_2 = e_1 + e_2, \qquad \ldots, \qquad \omega_{k-2} = e_1 + \ldots + e_{k-2}$$

but then something interesting happens because

$$\omega_{k-1} = \tfrac{1}{2}e_1 + \tfrac{1}{2}e_2 + \cdots + \tfrac{1}{2}e_{k-1} - \tfrac{1}{2}e_k \quad \text{and} \quad \omega_k = \tfrac{1}{2}e_1 + \tfrac{1}{2}e_2 + \cdots + \tfrac{1}{2}e_{k-1} + \tfrac{1}{2}e_k.$$

These are not weights, as they do not lie in the integer lattice of $[T_I SO(2n)]^*$. For $SO(2k+1)$ the computation is similar, and one would find that $\omega_i = e_1 + e_2 + \cdots + e_i$ for $i < k$ but that $\omega_k = \frac{1}{2}(e_1 + \cdots + e_k)$. At the moment these computations just serve to remind us that $SO(n)$ is not simply-connected and hence Theorem B.3.3 does not apply; but they will be more helpful when we start to talk about Spin(n).

Despite the fact that we don't have a fundamental system of weights, we still know by Theorem B.3.1 that the map

$$R_{\mathbb{C}}(SO(n)) \to R_{\mathbb{C}}(T)^W$$

is an isomorphism. The representation ring of the torus is the algebra of Laurent series $\mathbb{Z}[e^{\pm\alpha_1}, \dots, e^{\pm\alpha_k}]$ (where either $n = 2k$ or $n = 2k+1$). In the case $n = 2k+1$ the Weyl group W consists of all permuations of the α_i's together with all sign changes. Some examples of invariants are the elementary symmetric functions

$$\sigma_i = \sigma_i(e_1^{\alpha}, e^{-\alpha_1}, e^{\alpha_2}, e^{-\alpha_2}, \dots, e^{\alpha_k}, e^{-\alpha_k})$$

and it is not hard to see that these are algebraically independent for $1 \le i \le k$ and generate the entire ring of invariants. That is, $R(T)^W = \mathbb{Z}[\sigma_1, \dots, \sigma_k]$.

In the case $n = 2k$ the Weyl group consists of all permutations of the α_i's together with *even numbers* of sign changes. Of course all the σ_i's are still invariants, but when one gets up to σ_k one notices that it splits as the sum of two new invariants:

$$\sigma_k = \sigma_k^+ + \sigma_k^-.$$

Here σ_k^+ is the sum of all terms $e^{\pm\alpha_1 \pm \alpha_2 \pm \cdots \pm \alpha_k}$ in which there are an even number of minus signs, whereas σ_k^- is the sum of all such terms in which there are an odd number of minus signs. In fact this suggests a nice approach for understanding $R_{\mathbb{C}}(T)^W$: observe that it has an involution S defined by "changing an odd number of signs". This involution splits the ring into $+1$ and -1 eigenspaces, where the $+1$ eigenspace is exactly $\mathbb{Z}[\sigma_1, \dots, \sigma_k]$ (the ring of invariants considered earlier). The -1-eigenspace is a free module over the $+1$-eigenspace, generated by the element σ_k^-. For a proof of this, see [**Ad**, Proof of 7.9]. This gives us a description of $R_{\mathbb{C}}(T)^W$. Note that this ring is not a polynomial algebra!

For $SO(n)$ let U be the standard representation on $\mathbb{R}^n$ and let $V = U \otimes_{\mathbb{R}} \mathbb{C}$. So V consists of column-vectors over $\mathbb{C}$ of length n, with the action of $SO(n)$ given by left multiplication. Now assume $n = 2k$. Then it is easy to see that

$$\chi(V) = e^{\alpha_1} + e^{-\alpha_1} + \cdots + e^{\alpha_k} + e^{-\alpha_k} = \sigma_1.$$

More generally, $\chi(\Lambda^i V) = \sigma_i$. So $\mathbb{Z}[V, \Lambda^2 V, \dots, \Lambda^k V] \subseteq R_{\mathbb{C}}(G)$ is a polynomial subalgebra which maps to $\mathbb{Z}[\sigma_1, \dots, \sigma_k]$ under the character map $R_{\mathbb{C}}(G) \to R_{\mathbb{C}}(T)^W$. We will return to this in a moment, but let us pause to look at the odd case which is easier. Because for $n = 2k+1$ we have instead that

$$\chi(V) = e^{\alpha_1} + e^{-\alpha_1} + \cdots + e^{\alpha_k} + e^{-\alpha_k} + 1 = \sigma_1 + 1.$$

From this one readily deduces that $\chi(\Lambda^i V) = \sigma_i + \sigma_{i-1}$. Since $R_{\mathbb{C}}(T)^W = \mathbb{Z}[\sigma_1, \dots, \sigma_k] = \mathbb{Z}[\sigma_1 + 1, \sigma_2 + \sigma_1, \dots, \sigma_k + \sigma_{k-1}]$, we deduce that

$$R_{\mathbb{C}}(SO(2k+1)) = \mathbb{Z}[V, \Lambda^2 V, \dots, \Lambda^k V].$$

It follows immediately that the embedding $c\colon R_{\mathbb{R}}(SO(2k+1)) \to R_{\mathbb{C}}(SO(2k+1))$ is actually an isomorphism, since every $\Lambda^i V$ is the image of $\Lambda^i U$. In particular, every irreducible complex representation of $SO(2k+1)$ is real-type.

Remark B.4.4. As we have said before, be aware that knowing the representation ring of $R_{\mathbb{C}}(SO(2k+1))$ still doesn't give us a complete list of the irreducible complex representations. But for the moment it is the end of our story as far as the odd case is concerned.

Now return to the case of $SO(2k)$, with $U = \mathbb{R}^{2k}$ the standard representation and $V = \mathbb{C}^{2k}$ its complexification. Consider the Hodge $*$-operator

$$*\colon \Lambda^k(U) \to \Lambda^k(U),$$

which satisfies $*(v_1 \wedge \cdots \wedge v_k) = v_{k+1} \wedge \cdots \wedge v_{2k}$ for any oriented, orthonormal basis $v_1, \ldots, v_{2k}$. Note that if $A \in O(2k)$ and $\omega \in \Lambda^k(U)$ then

$$*(A\omega) = (\det A) \cdot A\omega$$

(it suffices to check this when $\omega = v_1 \wedge \cdots \wedge v_k$ with v as above, where the formula follows from the fact that $Av_1, \ldots, Av_{2k}$ is still orthonormal). In particular, the map $*$ is a map of $SO(2k)$-representations.

Note that $*^2 = (-1)^k$. If k is even then $\Lambda^k(U)$ decomposes into the $+1$ and -1 eigenspaces for $*$; if k is odd then we need to first complexify, after which $\Lambda^k(V)$ decomposes into the $+i$ and $-i$ eigenspaces. In what follows we will try to treat these two cases more or less simultaneously. If we let W_1 and W_2 be the eigenspaces, we claim that by naming things appropriately one has

$$\chi(W_1) = \sigma_k^+ + \text{ (lower terms)} \qquad \text{and} \qquad \chi(W_2) = \sigma_k^- + \text{ (lower terms)}.$$

To see this, let $L \in O(2k)$ be any element of determinant -1, for example the diagonal matrix with all 1's and a single -1 on the diagonal. Then $\omega \to L\omega$ gives a vector space isomorphism between the two eigenspaces. The map $A \mapsto LAL^{-1}$ is a map of groups $SO(2k) \to SO(2k)$, and we may restrict any representation along this map to get a new representation: let us call this the "L-conjugate" representation. Then $\omega \to L\omega$ is an isomorphism between the $+1$ (or $+i$) eigenspace and the -1 (or $-i$) eigenspace with its L-conjugate representation.

Write $\Lambda^k(V)_+$ and $\Lambda^k(V)_-$ for the subrepresentations of $\Lambda^k(V)$ whose characters have σ_k^+ and σ_k^- in them, respectively.

Theorem B.4.5. *$R_{\mathbb{C}}(SO(2k))$ is a free module over $\mathbb{Z}[V, \Lambda^2 V, \ldots, \Lambda^k V]$ on two generators, 1 and $\Lambda_k(V)_+$. When k is even all of these generators are real, so the map $c\colon R_{\mathbb{R}}(SO(2k)) \to R_{\mathbb{C}}(SO(2k))$ is an isomorphism. When k is odd then $V, \Lambda^2(V), \ldots, \Lambda^k(V)$ are all real, and the representations $\Lambda_k(V)_+$ and $\Lambda_k(V)_-$ are conjugate to each other.*

B.4.6. Spin representations. Let N be the kernel of $\mathrm{Spin}(n) \to SO(n)$. Then N is a subgroup of order 2, let us write it as $\{I, E\}$. Note that representations of $SO(n)$ are in bijective correspondence with representations of $\mathrm{Spin}(n)$ on which E acts as the identity. This correspondence clearly sends irreducibles to irreducibles.

Example B.4.7 (Irreducible representations of $SO(3)$). The group $\mathrm{Spin}(3)$ is just S^3, also known as the group $SU(2)$. The only elements of $SU(2)$ that square to I and I and $-I$, hence $E = -I$. Consider the irreducible complex representation H_n of $SU(2)$, consisting of the homogeneous degree n polynomials in z_1 and z_2. The action of E is to change the sign of z_1 and z_2, and this will correspond to the identity only if n is even. So the complete list of irreducible, complex representations of $SO(3)$ is

$$\mathbb{C} = H_0, \quad H_2, \quad H_4, \quad H_6, \quad \ldots$$

Recall that H_n has dimension $n + 1$, so the dimensions are $1, 3, 5, 7, \ldots$.

We have seen previously that all of the irreducible complex representations of $SO(3)$ are real-type, so the problem arises of giving real representations whose complexifications are the H_{2n}'s. This is usually done in terms of the so-called *spherical harmonics*, which we briefly review here.

Consider functions $F\colon \mathbb{R}^3 \to \mathbb{R}$ satisfying Laplace's equation $\nabla^2 F = 0$. To solve this equation one works in spherical coordinates and first looks for solutions of the form $F(\rho,\theta,\phi) = R(\rho)Y(\theta,\phi)$. This leads to separate differential equations in R and Y, and clearly $SO(3)$ acts on the solution space for Y. The equation for Y can be solved by again assuming a separation of variables $Y(\theta,\phi) = \Theta(\theta)\Phi(\phi)$, and one ends up discovering that there are a discrete set of solutions of the form $Y_\ell^m(\theta,\phi)$ indexed by integers ℓ and m where $-\ell \leq m \leq \ell$. Fixing ℓ, one can check that the space spanned by $Y_\ell^{-\ell}$, $Y_\ell^{-\ell+1}, \dots, Y_\ell^{\ell}$ is a subrepresentation of $SO(3)$. It turns out to be irreducible, and its complexification is $H_{2\ell}$.

Since $SO(n)$-representations are precisely Spin(n)-representations on which E acts as the identity, restricting along Spin$(n) \to SO(n)$ gives a embedding of representation rings $R_{\mathbb{C}}(SO(n)) \hookrightarrow R_{\mathbb{C}}(\mathrm{Spin}(n))$. From the general theory of simply-connected, compact Lie groups we know that $R_{\mathbb{C}}(\mathrm{Spin}(n))$ is a polynomial ring on fundamental representations $\rho_1, \dots, \rho_k$ (where $n = 2k$ or $n = 2k+1$). Let us see what we can determine about these representations.

Let $\tilde{T}$ be the maximal torus in Spin(n), so that we have a double cover $\tilde{T} \to T$. This induces a surjective map of tangent spaces $\tilde{M} \to M$ which sends $\tilde{M}_{\mathbb{Z}}$ onto $M_{\mathbb{Z}}$, and therefore an *injective* map of the duals: $M^* \hookrightarrow (\tilde{M})^*$ and $M_{\mathbb{Z}}^* \hookrightarrow (\tilde{M}_{\mathbb{Z}})^*$. In the latter case it must be that $M_{\mathbb{Z}}^*$ is an index 2 subgroup. If one now goes back to our attempt, at the beginning of Section B.4.3, to calculate the fundamental weights of $SO(n)$, it is clear that $(\tilde{M}_{\mathbb{Z}})^*$ must contain $\frac{1}{2}(\xi_1 + \dots + \xi_k)$. The fact that $\langle \xi_1, \dots, \xi_k \rangle$ is an index 2 subgroup then implies that we can identify

$$\tilde{M}_{\mathbb{Z}}^* = \big\langle \xi_1, \dots, \xi_k, \tfrac{1}{2}(\xi_1 + \dots + \xi_k) \big\rangle.$$

The calculation of the fundamental weights of Spin(n) is exactly what we tried to do in the case of $SO(n)$: for $n = 2k$ one gets

$$\omega_i = \begin{cases} \xi_1 + \dots + \xi_i & \text{if } i \leq k-2, \\ \frac{1}{2}(\xi_1 + \dots + \xi_{k-1} - \xi_k) & \text{if } i = k-1, \\ \frac{1}{2}(\xi_1 + \dots + \xi_{k-1} + \xi_k) & \text{if } i = k. \end{cases}$$

For $n = 2k+1$ one has instead that

$$\omega_i = \begin{cases} \xi_1 + \dots + \xi_i & \text{if } i \leq k-1, \\ \frac{1}{2}(\xi_1 + \dots + \xi_k) & \text{if } i = k. \end{cases}$$

Let $V = \mathbb{C}^n$ be the standard representation of $SO(n)$, regarded as a representation of Spin(n). Assume $n = 2k+1$. Then

$$\chi(V) = 1 + e^{\xi_1} + e^{-\xi_1} + \dots + e^{\xi_k} + e^{-\xi_k}.$$

We see that the highest weight is $\xi_1 = \omega_1$, so $V = \rho_1$. The character computation for $\Lambda^i(V)$ likewise reveals that the highest weight is $\xi_1 + \dots + \xi_i$. So for $i < k$ we have $\rho_i = \Lambda^i(V)$. For $i = k$ something interesting happens, though, because ω_k is only half of $\xi_1 + \dots + \xi_k$. So it is not true that $\Lambda^k(V) = \rho_k$. The irreducible representation with highest weight ω_k is called the **half-spin representation**, and denoted Δ. ????

THEOREM B.4.8. *The representation ring* $R_{\mathbb{C}}(\mathrm{Spin}(2k+1))$ *is the polynomial ring* $\mathbb{Z}[V, \Lambda^2(V), \ldots, \Lambda^{k-1}(V), \Delta]$*. The map* $R_{\mathbb{C}}(SO(2k+1)) \to R_{\mathbb{C}}(\mathrm{Spin}(2k+1))$ *sends* $\Lambda^k(V)$ *to* $\Delta^2 - (1 + V + \Lambda^2(V) + \cdots + \Lambda^{k-1}(V))$*.*

Now consider the case $n = 2k$. The first part of the analysis is the same, in the sense that $\Lambda^i(V)$ is an irreducible representation with highest weight $\xi_1 + \cdots + \xi_i$. Here we conclude that $\rho_i = \Lambda^i(V)$ for $i \leq k-2$, but we don't yet know ρ_{k-1} or ρ_k. These irreducible representations are again called **half-spin representations**, denoted by Δ_- and Δ_+, respectively.

THEOREM B.4.9. *The representation ring* $R_{\mathbb{C}}(\mathrm{Spin}(2k))$ *is the polynomial ring* $\mathbb{Z}[V, \Lambda^2(V), \ldots, \Lambda^{k-2}(V), \Delta_+, \Delta_-]$*. The map* $R_{\mathbb{C}}(SO(2k)) \to R_{\mathbb{C}}(\mathrm{Spin}(2k))$ *sends*

$$\begin{aligned} \Lambda^{k-1}(V) &\longrightarrow \Delta_+\Delta_- - (\Lambda^{k-3}V + \Lambda^{k-5}V + \cdots) \\ \Lambda^k(V)_+ &\longrightarrow \Delta_+^2 - (\Lambda^{k-2}V + \Lambda^{k-4}V + \cdots) \\ \Lambda^k(V)_- &\longrightarrow \Delta_-^2 - (\Lambda^{k-2}V + \Lambda^{k-4}V + \cdots). \end{aligned}$$

The construction of the fundamental half-spin representations is usually done via Clifford algebras and their modules. We will not recall this here, but see [**BtD**].

Bibliography

[Ad] J. F. Adams, *Lectures on Lie Groups*, W. A. Benjamin, Inc., 1969.

[A] V. I. Arnold, *Mathematical methods of classical mechanics, second edition*, Graduate Texts in Mathematics **60**, Springer, 1989,

[At] M. Atiyah, *The geometry and physics of knots*, Cambridge University Press, 1990.

[BM] J. Baez and J. Muniaian, *Gauge fields, knots and gravity*, World Scientific Publishing Company, 1994.

[BT] R. Bott and L. Tu, *Differential forms in algebraic topology*.

[BtD] T. Bröcker and T. tom Dieck, *Representations of compact Lie groups*, Springer-Verlag New York, 1985.

[De] E. A. Desloge, *Classical mechanics, Volume 1 & 2*, John Wiley & Sons, Inc., 1982.

[FLS] R. P. Feynman, R.B. Leighton, and M. Sands, *The Feynman lectures on physics*, Addison-Wesley Publishing Company, 1965.

[FH] R. P. Feynman and A. Hibbs, *Quantum mechanics and path integrals*, McGraw-Hill, 1965.

[FuH] W. Fulton and J. Harris, *Representation theory: a first course*, Springer.

[Go] Goldstein, *Classical mechanics*, ??????.

[G1] D. J. Griffiths, *Introduction to quantum mechanics*, Prentice-Hall, New Jersey, 1995.

[LM] H. B. Lawson and M-L. Michselsohn, *Spin geometry*, Princeton University Press, 1989.

[M] I. H. Madsen and J. Tornehave, *From calculus to cohomology: De Rham cohomology and characteristic classes*, Cambridge University Press, 1997.

[P] A.M. Polyakov, *Fermi-Bose transmutations induced by gauge fields* , Mod. Phys. Lett. A 3 (1988) 325

[R] J. Rabin, *Introduction to quantum field theory for mathematicians*, in Geometry and Quantum Field Theory (eds. D. Freed and K. Uhlenbeck), American Mathematical Society, 1995.

[S1] J. J. Sakurai, *Advanced quantum mechanics*, Addison-Wesley Publishing Company, Inc., 1967.

[S] Shankar, *Principles of quantum mechanics, second edition*, Plenum Press, New York, 1994.

[W] E. Witten, *Quantum field theory and the Jones polynomial*, ????.

[Z] A. Zee, *Quantum field theory in a nutshell, second edition*, Princeton University Press, 2010.