Lecture notes for Mathematical Physics

Joseph A. Minahan[1]

Department of Theoretical Physics
Box 803, SE-751 08 Uppsala, Sweden

[1]E-mail: joseph.minahan@teorfys.uu.se

1 Introduction

This is a course in Mathematical methods in physics. I should stress at the very beginning that I am a physicist and not a mathematician, so many of the proofs and exercises offered, will not be as rigourous as would be given by a proper mathematician. In stead, our goal will be to develop a set of tools that will be useful for a would be physicist. Much of what I hope to cover in this course is commonly used by a practicing theoretical physicist.

This course is still a work in progress, but I hope to cover the following topics:

- Group Theory and Lie Algebras
- Path Integrals
- Topology
- Differential Geometry
- Yang-Mills

Each one of these topics is a course in itself, so much of the presentation here will be somewhat sketchy.

2 Group Theory

A physics problem can be simplified if there is symmetry in the problem. For example, you probably remember that in ordinary one dimensional quantum mechanics, if the potential is invariant under the symmetry $x \to -x$, that is,

$$V(-x) = V(x), \tag{2.0.1}$$

then we immediately know that the eigenfunctions of the Hamiltonian are either even or odd under this transformation. This transformation is known as parity and is an element in one of the simplest examples of a *group*. We say that the wavefunctions transform under a *representation* of the group.

A less trivial example that you should know from your quantum mechanics courses is the example of a central potential, where the potential only depends on a radius and not on the angles. This potential is invariant under rotations and the wave functions can be classified by different eigenvalues of the angular momentum operator. These three dimensional rotations form a group, the *rotational group*, and the wave functions are in representations of this group, the different representations classified by ℓ, where

$$\vec{L}^2 \psi_\ell(r, \theta, \phi) = \hbar^2 \ell(\ell + 1) \psi_\ell(r, \theta, \phi). \tag{2.0.2}$$

These two different groups have some significant differences. In the first case, the group is finite – there is only one nontrivial transformation ($x \to -x$). In the second case, the group is continuous. The group of rotations is described by 3 continuous angles, the Euler angles.

There is another significant difference between these groups. In the first case, under a series of transformations, it does not matter in which order you perform the transformations. The end result is the same. However, in the second case, the order does matter. A rotation in the $x-y$ plane followed by a rotation in the $x-z$ plane leads to a different result than a rotation in the $x-z$ plane followed by a rotation in the $x-y$ plane. The first case is an example of an *Abelian* group, and the second case is the example of a *Non-Abelian* group.

Now that we have given these examples, let us try and give a more concrete definition of a group.

2.1 Groups (the definitions)

A group is a set $\mathcal{G}$ with a collection of objects, g_i in $\mathcal{G}$. The group has associated with it an operation which we write as "$\cdot$". The elements of the group have to satisfy the following properties

- If $g_1, g_2 \in \mathcal{G}$, then $g_1 \cdot g_2 \in \mathcal{G}$.
- There is a unique identity, $1 \in \mathcal{G}$, such that $g_i \cdot 1 = g_i$, $\forall g_i \in \mathcal{G}$.
- For each $g_i \in \mathcal{G}$, there is a unique inverse, $g_i^{-1} \in \mathcal{G}$, such that $g_i \cdot g_i^{-1} = 1$.
- Associativity: $(g_1 \cdot g_2) \cdot g_3 = g_1 \cdot (g_2 \cdot g_3)$.

That's it! Any collection of objects with some operation that satisfies these 4 properties is a group. Note that $g_1 \cdot g_2 = g_2 \cdot g_1$ is not a requirement for this to be a group. However, if this is true for all elements of the group, then we say that the group is *Abelian*. Another term that we will use is the *order* of a group, which is the number of elements.

2.2 Examples

Here are some examples.

1) The integers under addition. In this case "$\cdot$"= "+". $n_1 + n_2 = n_3$. If n_1 and n_2 are integers then clearly n_3 is an integer, so the first property is satisfied. The identity element is 0, and it is obviously unique. The unique inverse of n_1 is $-n_1$. Finally, addition is associative. Therefore this is a group. Moreover, addition is commutative, that is

$n_1 + n_2 = n_2 + n_1$, so the group is Abelian. Notice that the integers under multiplication is *not* a group, since in general the inverse is not an integer.

2) Parity: $x \to -x$. This group has two elements, 1 and Π, where $\Pi^2 = 1$,[2] hence Π is its own inverse. This group is clearly Abelian, and has order 2.

3) Permutation groups on N elements. The elements of this group are *generated* by, g_{ij}, the operations that exchange element i with j. The *generators* of a group are a subset of $\mathcal{G}$ out of which all group elements can be constructed. As an example, consider the permutation group on 3 elements. The generators of this group are g_{12}, g_{23} and g_{13}. Actually, we don't even need g_{13}, since we can generate it with the other two elements:

$$g_{13} = g_{12} g_{23} g_{12}. \tag{2.2.1}$$

To see this, act with $g_{12}g_{23}g_{12}$ on 3 elements (a, b, c).

$$g_{12}g_{23}g_{12}(a, b, c) = g_{12}g_{23}(b, a, c) = g_{12}(b, c, a) = (c, b, a) = g_{13}(a, b, c). \tag{2.2.2}$$

We also see that this group is nonabelian, since $g_{12}g_{23} \neq g_{23}g_{12}$. Since there are $N!$ ways to order N elements, the order of the permutation group is $N!$.

4) Rotations in the plane. There are many ways to represent this. In matrix notation, we can express the transformation corresponding to a rotation by angle θ as

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \tag{2.2.3}$$

The transformation of a two dimensional vector is

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x\cos\theta - y\sin\theta \\ y\cos\theta + x\sin\theta \end{pmatrix} \tag{2.2.4}$$

The product of two such rotations is

$$\begin{pmatrix} \cos\theta_1 & -\sin\theta_1 \\ \sin\theta_1 & \cos\theta_1 \end{pmatrix} \begin{pmatrix} \cos\theta_2 & -\sin\theta_2 \\ \sin\theta_2 & \cos\theta_2 \end{pmatrix} = \begin{pmatrix} \cos(\theta_1+\theta_2) & -\sin(\theta_1+\theta_2) \\ \sin(\theta_1+\theta_2) & \cos(\theta_1+\theta_2) \end{pmatrix} \tag{2.2.5}$$

Clearly, this is an abelian group. There is also another way to represent this group. Note that if we let $z = x + iy$, then under the transformation in (2.2.4), z transforms as

$$z \to e^{i\theta} z \tag{2.2.6}$$

[2]Note, that where the notation is clear, we will often drop the "."

In other words, we can represent the group elements as $e^{i\theta}$. Since these can be represented by 1 dimensional complex numbers, we call this group $U(1)$. The "U" stands for unitary, since for every element g of the group, we have that

$$g^\dagger = g^{-1} \tag{2.2.7}$$

$U(1)$ is an example of what is called a *compact* group. Roughly speaking, this is because every group element is described by a θ over a compact space, $0 \leq \theta < 2\pi$. This will be made more precise later on.

5) $SL(2,R)$. Another interesting continuous group that we will come across is $SL(2,R)$, which stands for "Special Linear 2 dimensional, Real". The group elements are made up of 2 by 2 matrices with real entries whose determinant is 1. In other words, we have the elements

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \qquad\qquad ad - bc = 1. \tag{2.2.8}$$

We can show that this is a group, because if we multiply two such matrices, say A and B, we are left with a matrix whose entries are real. Furthermore, since $\det(AB) = \det A \det B$, we see that the determinant of the product is also real, hence the product is an element of the group. There is clearly a unique identity element $I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, and since the determinant of the matrices are nonzero, every element has an inverse. Therefore, $SL(2,R)$ is a group. However, the group is nonabelian, since usually, 2 by 2 matrices don't commute with one another.

It turns out that $SL(2,R)$ is an example of a group that is *noncompact*. To see this, note that for any given a, b, and c, the expression for d is given by $d = (1+bc)/a$. Hence as long as a is nonzero, there is a solution for d. If $a = 0$, then we must have $bc = -1$. In any event, a, b and c can range over all real values. Now this in itself does not make the group noncompact. It turns out that we will have to define something called a *group measure*, such that integrating over this measure gives an infinite result, in the sense that integrating dx over all x gives an infinite result.

6) $SU(2)$. The group $U(N)$ has elements which are unitary $N \times N$ complex matrices. If we restrict ourselves to those matrices whose determinant is 1, then these matrices form a group called $SU(N)$, where the "SU" stands for special unitary. Let us concentrate on the case of $SU(2)$. We note that a unitary matrix U can be written as $U = exp(iH)$, where H is hermitian ($H^\dagger = H$). In 2 dimensions, any Hermitian matrix can be expressed as

$$H = a_0 I + a_1\sigma_1 + a_2\sigma_2 + a_3\sigma_3, \tag{2.2.9}$$

where $I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, σ_i are the Pauli matrices

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad \sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad , \tag{2.2.10}$$

and a_0 and a_i are real numbers. The determinant is given by $\det U = \exp(i\mathrm{Tr}H)$ (prove this!), therefore, we see that $\mathrm{Tr}H = 0$ and so $a_0 = 0$. Hence, we see that our unitary matrix is described by three numbers a_1, a_2 and a_3.

We can now say more about $SU(2)$. If we use the fact that $\sigma_i\sigma_j + \sigma_j\sigma_i = 2I\delta_{ij}$, then it must be true that any element U can be written as

$$U = b_0 I + i(b_1\sigma_1 + b_2\sigma_2 + b_3\sigma_3) \tag{2.2.11}$$

where all coefficients are real and where

$$b_0^2 + b_1^2 + b_2^2 + b_3^2 = 1 \tag{2.2.12}$$

This last condition insures that the determinant is 1. But the condition in (2.2.12) tells us something else. Namely that the elements of $SU(2)$ map to the 3 sphere! In other words, for every element of $SU(2)$, there is a corresponding point on a 3 dimensional sphere. This will be of importance later on in this course. This 3-sphere is called the *group manifold* of $SU(2)$.

2.3 Subgroups and Cosets

A *subgroup* $\mathcal{H}$ of a group $\mathcal{G}$ is a subset of the elements of $\mathcal{G}$, such that the elements of the subset form a group. In other words, we have that $h_i \in \mathcal{H}$ so that the h_i satisfy all of the properties for a group mentioned above. Every group has at least one subgroup, namely the trivial group made up of one element, the identity element 1. Note that the identity element for $\mathcal{G}$ is the identity element for $\mathcal{H}$.

Given a subgroup we can define a *coset*. A coset of the group $\mathcal{G}$ by the subgroup $\mathcal{H}$ is written as $\mathcal{G}/\mathcal{H}$ (or sometimes as $\mathcal{H}\backslash\mathcal{G}$) and is defined as an identification. In other words, two elements g_i and g_j are considered identical elements in the coset if

$$g_i = g_j \cdot h \qquad \text{for some } h \in \mathcal{H}. \tag{2.3.1}$$

Actually, this is what is known as a *right coset*. A *left coset* has the identification

$$g_i = h \cdot g_j \qquad \text{for some } h \in \mathcal{H}. \tag{2.3.2}$$

If the group is abelian, then there is no difference between right and left. However, if the group is nonabelian, then there could be a difference in the cosets, although there will be a one to one map of the elements in the left coset to the elements of the right coset. *n.b.* The coset is not always a group!

One other thing to note is that the order of $\mathcal{G}$ divided by the order of the subgroup is the order of the coset, in other words, the number of elements in the coset. Of course, this must mean that the order of any subgroup $\mathcal{H}$ divides the order of the original group $\mathcal{G}$ (prove this!)

2.4 Examples of Subgroups and Cosets

1)The integers. A subgroup of the integers is all integers that are a multiple of m, where m is some positive integer greater than 1. Clearly the "product" of any two such elements gives an integer which is a multiple of m. The coset of such a subgroup is the integers mod m. Note that these subgroups have an infinite number of elements but the coset only has a finite number. Note further that the cosets also form a group, the group of modular addition.

2) Parity. This group only has the identity element as a subgroup.

3) Permutations. Examples of subgroups of the permutation group on N elements are the permutation groups on M elements where $M < N$. So for example, the permutation group on 3 elements has as a subgroup, the permutation group on two elements. This subgroup has the elements 1 and g_{12}. A left coset for this subgroup has 3 elements up to identification: 1, g_{13} and g_{23}. Note that $1 \equiv g_{12}$, $g_{13} \equiv g_{12}g_{23}$ and $g_{23} \equiv g_{12}g_{13}$. The coset is *not* a group, since for example $g_{13}^2 = 1$, but $(g_{12}g_{23})^2 = g_{12}g_{13}$, so the identification is not preserved under group multiplication.

4) $U(1)$. The subgroups of this group are given by Z_N, whose elements are given by the solutions of the equation $z^N = 1$. The solutions are given by $e^{2\pi i n/N}$. Clearly these form a group under multiplication. The cosets of these groups are made up of the elements $e^{i\phi}$, where we identify elements if $\phi_1 = \phi_2 + 2\pi n/N$.

5) $SL(2,R)$. One interesting subgroup is the group $SL(2,Z)$, where the entries of the matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \qquad ad - bc = 1. \tag{2.4.1}$$

are all integers. Clearly any product gives a matrix whose entries are integers. The identity also has integer entries. Less trivial to check is that the inverse of any element only has integer entries. However, this is guaranteed by the fact that the determinant of

the matrix is 1. Therefore, the inverse of (2.4.1) is

$$\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \tag{2.4.2}$$

and so the inverse also has integer entries.

Another subgroup of $SL(2,R)$ is $U(1)$. This is clear since our original formulation of $U(1)$ was given in terms of real 2×2 matrices with determinant 1. We will defer the discussion on the coset.

6) $SU(2)$. $SU(2)$ also has a $U(1)$ subgroup, actually, it has many $U(1)$ subgroups. For example, we can choose the subgroup to be those $U = \exp i\phi\sigma_3$. The coset is very interesting. Let a general $SU(2)$ element be given by (2.2.11). Then if we multiply it by an element of the subgroup, we get

$$U = b_0 \cos\phi - b_3 \sin\phi + i(b_1 \cos\phi + b_2 \sin\phi)\sigma_1 + i(b_2 \cos\phi - b_1 \sin\phi)\sigma_2 + i(b_3 \cos\phi + b_0 \sin\phi)\sigma_3. \tag{2.4.3}$$

Let us define

$$w = b_0 - ib_3 \qquad\qquad z = b_1 + ib_2. \tag{2.4.4}$$

Therefore under the transformation of U in (2.4.3), we see that

$$w \to e^{-i\phi} w \qquad\qquad z \to e^{-i\phi} z. \tag{2.4.5}$$

The constraints on the b's results in the constraint

$$|z|^2 + |w|^2 = 1 \tag{2.4.6}$$

Identification under the subgroup means that equal rotations of z and w in the complex plane correspond to the same coset element. Now let us define $\tilde{z} = \rho z$ and $\tilde{w} = \rho w$, where ρ is any positive real. Then it is clear from z and w, that we have the same point in the coset if we identify $\{\tilde{z}, \tilde{w}\} \equiv \{\rho\tilde{z}, \rho\tilde{w}\}$, since both expressions can come from the same z and w. If we also include the identification under the phase rotation, then we see that we can describe the coset using the identification $\{\tilde{z}, \tilde{w}\} \equiv \{\lambda\tilde{z}, \lambda\tilde{w}\}$, where λ is any complex number other than zero. This space is known as $CP(1)$, for 1 dimensional complex projective plane. The dimension of this space is one complex dimension, which is the same as two real dimensions. This is because $\tilde{z}$ and $\tilde{w}$ each have one complex dimension, but the identification removes one dimension. Even though $\tilde{z}$ and $\tilde{w}$ can be any complex number (except, both can't be zero), the space is compact. We will show this when we discuss topology.

2.5 Representations of Groups

A *representation*, is a mapping of group elements that preserves the group multiplication law. In fact, up to now, we have been describing examples of groups through their representations. The elements of the group are a somewhat abstract notion. The representations give a concrete description of the groups.

That is not too say that all representations are equivalent. For example, when we write group elements as matrices, this is a representation of the group. Let us call an element of this representation $M(g_i)$, where $M(g_i)$ is the matrix corresponding to the element g_i in $\mathcal{G}$. Therefore, we have that

$$M(g_i)M(g_j) = M(g_i \cdot g_j). \tag{2.5.1}$$

Then an equivalent representation is where

$$M(g_i) \to \widetilde{M}(g_i) = AM(g_i)A^{-1}, \tag{2.5.2}$$

where A is a matrix that is the same for all group elements. If an A exists such that all $M(g_i)$ can be transformed to the form

$$\widetilde{M}(g_i) = \begin{pmatrix} \widetilde{M}_1(g_i) & 0 \\ 0 & \widetilde{M}_2(g_i) \end{pmatrix}, \tag{2.5.3}$$

where the matrices are all in block diagonal form, then we say that the representation is *reducible*. Notice that all of the blocks form representations of the group. So our original representation is just a combination of smaller representations of the group.

As an example, consider the case of the exchange between two elements. We can then write the elements as

$$M(1) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad M(g_{12}) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{2.5.4}$$

This is one representation of the group. However, if we perform the similarity transformation $M(g_i) \to AM(g_i)A^{-1}$, where

$$A = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, \tag{2.5.5}$$

then the new matrices are

$$\widetilde{M}(1) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \widetilde{M}(g_{12}) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{2.5.6}$$

Hence the representation is comprised of two representations. One of these is the trivial represenation where $\widetilde{M}_1(g_i) = 1$. The other representation has $\widetilde{M}_2(1) = 1$ and $\widetilde{M}_2(g_{12}) = -1$. Obviously, these representations cannot be reduced further. Representations that cannot be reduced any further are said to be *irreducible*.

Next consider the group $U(1)$. The matrices in (2.2.3), can be diagonalized to the form

$$\begin{pmatrix} e^{i\theta} & 0 \\ 0 & e^{-i\theta} \end{pmatrix} \tag{2.5.7}$$

Hence the original two dimensional representation is reducible to two one dimensional representations. However, these are not the only representations of $U(1)$. We can easily see that $e^{in\theta}$ where n is any integer is a representation of $U(1)$. In fact, these are all of the irreducible representations.

In the case of $SU(2)$, the representation in (2.2.11) is irreducible. There is no way to block diagonalize all elements simultaneously, since the different Pauli matrices don't commute with each other. This is the smallest nontrivial representation of $SU(2)$. The one dimensional representation is trivial.

2.6 Lie Groups and Lie Algebras

A *Lie group* is a continuous group that is specified by a set of parameters. For a *compact Lie Group*, the parameter space is compact. The groups $U(1)$ and $SU(N)$ are examples of Lie groups. A Lie group has associated with it a set of *generators*, T_a, that can generate the entire group. If a group element can be smoothly continued to the identity, then such a group element can be written as

$$U = \exp\left(i\sum_a \theta_a T_a\right) = \exp(i\theta_a T_a), \tag{2.6.1}$$

where the θ_a are the parameters.

Let us now assume that the parameters are very small, so that U is very close to the identity. I will write the parameters as $\theta_a = \epsilon_a$ hence

$$U \approx 1 + i\epsilon_a T_a - \frac{1}{2}\epsilon_a\epsilon_b T_a T_b \ + \mathrm{O}(\epsilon^3). \tag{2.6.2}$$

Suppose I consider the product $UVU^{-1}V^{-1}$, where V is also close to the identity, with parameters ϵ'_a. Then the resulting product is also a group element. Performing the infinitesimal multiplication, we find that

$$UVU^{-1}V^{-1} \approx 1 - \epsilon_a\epsilon'_b T_a T_b + \epsilon_a\epsilon'_b T_b T_a + ... \tag{2.6.3}$$

Since this is a group element, comparing with (2.6.2), we see that the lowest order correction to the identity in (2.6.3) is a linear combination of the generators. Hence the generators must satisfy an algebra of the form

$$[T_a, T_b] = i f_{abc} T_c. \tag{2.6.4}$$

This algebra is called a *Lie algebra* and the coefficients f_{abc} are called the *structure constants*. The number of independent generators is called the *dimension* of the algebra. We will denote this number by D.

If we write the generators in matrix form, then we say that these are in a *representation* of the algebra. The structure constants obviously satisfy the relation $f_{abc} = -f_{bac}$. But they also have other properties of note. Notice that by the *Jacobi identity*

$$[[T_a, T_b], T_c] + [[T_c, T_a], T_b] + [[T_b, T_c], T_a] = 0 \tag{2.6.5}$$

Therefore, using the algebra in (2.6.4) in (2.6.5) we find

$$-f_{abd} f_{dce} T_e - f_{cad} f_{dbe} T_e - f_{bcd} f_{dae} T_e = 0. \tag{2.6.6}$$

Since the T_e are assumed to be independent, this reduces to just an equation for the structure constants. Using their antisymmetry properties, we find

$$f_{acd} f_{bde} - f_{bcd} f_{ade} = -f_{abd} f_{dce}. \tag{2.6.7}$$

In other words, $i f_{acd}$ is a representation for the generator T_a, where the c and d label in the structure constant refers to the row and column of the matrix. This representation is called the *adjoint* representation and its dimension is equal to the dimension of the algebra, D.

Associated with the Lie algebra (2.6.4) is a subalgebra known as the *Cartan subalgebra*. The Cartan subalgebra is made up of a subset of generators, which we call H_i, that satisfy the algebra

$$[H_i, H_j] = 0. \tag{2.6.8}$$

The H_i can all be chosen to be Hermitian. The number of generators in the Cartan subalgebra is known as the *rank*, n of the Lie algebra. Note that there are many ways to choose a Cartan subalgebra among the elements of the Lie algebra.

Let us now turn to the specific example of $SU(2)$. From eq. (2.2.11) we see that one representation of the group has generators which are the Pauli matrices, so the dimension of the algebra is 3. Let us call the generators $T_a = \frac{1}{2}\sigma_a$. Then the algebra is given by

$$[T_a, T_b] = i\epsilon_{abc} T_c, \qquad \epsilon_{123} = 1. \tag{2.6.9}$$

Hence the algebra of $SU(2)$ is isomorphic to the algebra for angular momentum in quantum mechanics. The Cartan subalgebra has one generator, namely T_3 (although we could have chosen any other generator), hence the rank of $SU(2)$ is 1. We should also expect the irreducible representations of $SU(2)$ to be those representations that we found for the angular momentum, that is, the representations should be related to the allowed spins for particles consistent with the algebra.

2.7 Roots

Having chosen a cartan subalgebra, we can then classify the other generators. To do this, let us first recall something we have learned from quantum mechanics. In quantum mechanics we learned about "bra" and "ket" states. For our purposes, we will define a set of D independent ket states as

$$|T_a\rangle. \tag{2.7.1}$$

In other words, for every independent T_a in the Lie algebra we have a corresponding independent ket state. These states satisfy the necessary linearity properties of quantum mechanics, namely that

$$|\alpha T_a + \beta T_b\rangle = \alpha|T_a\rangle + \beta|T_b\rangle. \tag{2.7.2}$$

For the bra states we have

$$\langle T_a| = (|T_a\rangle)^\dagger . \tag{2.7.3}$$

We also need an inner product that satisfies the requirement that

$$\langle T_a|T_b\rangle^* = \langle T_b|T_a\rangle. \tag{2.7.4}$$

It is easy to see that

$$\langle T_a|T_b\rangle = \mathrm{Tr}(T_a^\dagger T_b) \tag{2.7.5}$$

satisfies the requirement in (2.7.4), since

$$\mathrm{Tr}(T_a^\dagger T_b)^* = \mathrm{Tr}(T_a^T T_b^*) = \mathrm{Tr}((T_b^\dagger T_a)^T) = \mathrm{Tr}(T_b^\dagger T_a). \tag{2.7.6}$$

Now define the operator Π_{T_c}, whose action on the states is

$$\Pi_{T_c}|T_a\rangle = |[T_c, T_a]\rangle = i f_{cab}|T_b\rangle. \tag{2.7.7}$$

The adjoint of this operator satisfies $(\Pi_{T_c})^\dagger = \Pi_{T_c^\dagger}$ (show this). Π_{T_c} is obviously a linear operator since

$$\Pi_{T_c}|T_a + T_b\rangle = |[T_c, (T_a + T_b)]\rangle = |[T_c, T_a]\rangle + |[T_c, T_b]\rangle = \Pi_{T_c}|T_a\rangle + \Pi_{T_c}|T_b\rangle. \tag{2.7.8}$$

It is then straightforward to show using the Jacobi identity and the linearity of the operators that

$$[\Pi_{T_a}, \Pi_{T_b}] = \Pi_{[T_a,T_b]} = if_{abc}\Pi_{T_c}, \tag{2.7.9}$$

and so

$$[\Pi_{H_i}, \Pi_{H_j}] = 0 \tag{2.7.10}$$

for H_i and H_j in the Cartan subalgebra. Since these operators commute, the states can be simultaneously eigenstates for all such operators coming from the Cartan subalgebra. Furthermore, if the H_i are Hermitian matrices, then the operator Π_{H_i} is Hermitian. Therefore, its eigenvalues are real.

Thus, let us suppose that we have a Lie algebra with rank n, and so n independent generators in the Cartan subalgebra. Let us write these as a vector

$$(H_1, H_2..., H_n). \tag{2.7.11}$$

Let us also suppose that the H_i are orthonormal, so that

$$\mathrm{Tr}(H_i H_j) = k\delta_{ij}. \tag{2.7.12}$$

Using our arguments from the preceding paragraphs, a basis can be chosen for the generators outside the Cartan subalgebra, with basis vectors $G_{\vec{a}}$ where $\vec{a}$ is an n dimensional vector

$$\vec{a} = (a_1, a_2...a_n) \tag{2.7.13}$$

such that

$$[H_i, G_{\vec{a}}] = a_i G_{\vec{a}}. \tag{2.7.14}$$

The different $\vec{a}$ are called the *roots* and the $G_{\vec{a}}$ are called the *root generators*. Since the H_i are Hermitian, the components of the roots are real. Furthermore, if we take the Hermitian conjugate on (2.7.14), we find that

$$G^\dagger_{\vec{a}} = G_{-\vec{a}} \tag{2.7.15}$$

We can also establish some other properties. For one thing, it is clear from (2.7.14) that commutator of two roots satisfies

$$[G_{\vec{a}}, G_{\vec{b}}] \sim G_{\vec{a}+\vec{b}} \qquad \vec{a} + \vec{b} \neq 0. \tag{2.7.16}$$

If $\vec{a} + \vec{b}$ is not one of the roots, then the commutator must be zero. It is also clear from the hermiticity of Π_{H_i}, that

$$\mathrm{Tr}(G_{\vec{a}} G_{\vec{b}}) = k\delta_{\vec{a}+\vec{b}}, \tag{2.7.17}$$

where the $\delta_{\vec{0}} = 1$, and is zero otherwise. The $G_{\vec{a}}$ have been rescaled so that they have the same factor of k as in (2.7.12). It is also clear that all H_i commute with $[G_{\vec{a}}, G_{-\vec{a}}]$, therefore

$$[G_{\vec{a}}, G_{-\vec{a}}] = \sum \lambda_i H_i. \tag{2.7.18}$$

Now suppose we act with $\Pi_{G_{\vec{a}}}$ on $|G_{\vec{b}}\rangle$, then we have

$$\Pi_{G_{\vec{a}}}|G_{\vec{b}}\rangle \sim |G_{\vec{a}+\vec{b}}\rangle, \tag{2.7.19}$$

Hence $\Pi_{G_{\vec{a}}}$ is a raising operator and $\Pi_{G_{-\vec{a}}}$ is the corresponding lowering operator. If we act with $\Pi_{G_{\vec{a}}}$ on the state $|G_{-\vec{a}}\rangle$, then we have

$$\Pi_{G_{\vec{a}}}|G_{-\vec{a}}\rangle = \sum_i \lambda_i |H_i\rangle. \tag{2.7.20}$$

But we also have that

$$\begin{aligned} \langle H_j|\Pi_{G_{\vec{a}}}|G_{-\vec{a}}\rangle &= \mathrm{Tr}(H_j[G_{\vec{a}}, G_{-\vec{a}}]) \\ &= \mathrm{Tr}([H_j, G_{\vec{a}}]G_{-\vec{a}}) = a_j \mathrm{Tr}(G_{\vec{a}} G_{-\vec{a}}) \end{aligned} \tag{2.7.21}$$

Hence, using the orthogonality of the H_i, we find that

$$[G_{\vec{a}}, G_{-\vec{a}}] = \sum a_i H_i. \tag{2.7.22}$$

One can also check that $G_{\vec{a}}$ is traceless, since

$$\mathrm{Tr}[H_i, G_{\vec{a}}] = 0 = a_i \mathrm{Tr}(G_{\vec{a}}). \tag{2.7.23}$$

Finally, we can show that the root generator $G_{\vec{a}}$ is unique for a given $\vec{a}$. To see this, let us suppose that there exists another root generator $G'_{\vec{a}}$ that is orthogonal to $G_{\vec{a}}$. In other words

$$\langle G'_{\vec{a}}|G_{\vec{a}}\rangle = \mathrm{Tr}[(G'_{\vec{a}})^\dagger G_{\vec{a}}] = 0, \tag{2.7.24}$$

where $(G'_{\vec{a}})^\dagger = G'_{-\vec{a}}$. Now consider the inner product

$$\langle H_i|\Pi^\dagger_{G'_{\vec{a}}}|G_{\vec{a}}\rangle = \langle [G'_{\vec{a}}, H_i]|G_{\vec{a}}\rangle = -a_i \langle G'_{\vec{a}}|G_{\vec{a}}\rangle = 0. \tag{2.7.25}$$

But going the other way, we have that this is

$$\langle H_i|[G'_{-\vec{a}}, G_{\vec{a}}]\rangle. \tag{2.7.26}$$

But we also know that

$$[G'_{-\vec{a}}, G_{\vec{a}}] = \sum_i \sigma_i H_i, \tag{2.7.27}$$

for some constants σ_i. Hence consistency with (2.7.25) requires that

$$[G'_{-\vec{a}}, G_{\vec{a}}] = 0. \tag{2.7.28}$$

Now consider the inner product

$$\langle G_{\vec{a}} | \Pi_{G'_{\vec{a}}} \Pi^\dagger_{G'_{\vec{a}}} | G_{\vec{a}} \rangle = 0. \tag{2.7.29}$$

That this inner product is zero follows from (2.7.28). But this inner product can also be written as

$$\begin{aligned} \langle G_{\vec{a}} | \Pi_{G'_{\vec{a}}} \Pi^\dagger_{G'_{\vec{a}}} | G_{\vec{a}} \rangle &= \langle G_{\vec{a}} | [\Pi_{G'_{\vec{a}}}, \Pi^\dagger_{G'_{\vec{a}}}] | G_{\vec{a}} \rangle + \langle G_{\vec{a}} | \Pi^\dagger_{G'_{\vec{a}}} \Pi_{G'_{\vec{a}}} | G_{\vec{a}} \rangle \\ &= \langle G_{\vec{a}} | a_i \Pi_{H_i} | G_{\vec{a}} \rangle + \langle G_{\vec{a}} | \Pi^\dagger_{G'_{\vec{a}}} \Pi_{G'_{\vec{a}}} | G_{\vec{a}} \rangle \\ &= |\vec{a}|^2 + \langle G_{\vec{a}} | \Pi^\dagger_{G'_{\vec{a}}} \Pi_{G'_{\vec{a}}} | G_{\vec{a}} \rangle > 0. \end{aligned} \tag{2.7.30}$$

Hence we have a contradiction.

2.8 Classifying Groups (Part 1)

In this section we will begin Cartan's proof of the classification of simple compact Lie algebras.

We say that a Lie algebra and its corresponding Lie group are *reducible*, if we can break up the generators into two groups, say T_a and T'_a, such that $[T_a, T'_b] = 0$ for all primed and unprimed generators. We call the Lie algebra *simple* if it cannot be reduced this way. One upshot of this is that the generators in the Cartan subalgebra are traceless. For example, the groups $U(N)$, and hence their Lie algebras are reducible, since $U(N) = U(1) \times SU(N)$ and the generator of the $U(1)$ group commutes with all generators of the $SU(N)$ group. It turns out that the $SU(N)$ groups are simple groups.

Let's now classify the groups. Since the number of generators of the Lie group is finite, we can always choose a root generator $G_{\vec{b}}$, so that the *length* of the root vector, $|\vec{b}|$, is greater than or equal to the length of any other root vector. Let us now act with the operator $\Pi_{G_{\vec{a}}}$ on the state for the root $\vec{b}$

$$\Pi_{G_{\vec{a}}} | G_{\vec{b}} \rangle. \tag{2.8.1}$$

In order for this to be nonzero $|\vec{a}+\vec{b}| \leq |\vec{b}|$ since we have assumed that no root is longer than $\vec{b}$. Now by the triangle inequality, at least one of $|\vec{a}+\vec{b}|$ or $|-\vec{a}+\vec{b}|$ is greater than $|\vec{b}|$. We will assume that this is true for the latter case. Now consider the positive definite inner product

$$\langle G_{\vec{b}}|\Pi_{G_{-\vec{a}}}\Pi_{G_{\vec{a}}}|G_{\vec{b}}\rangle. \tag{2.8.2}$$

Using (2.7.9) and (2.7.22), we have

$$\langle G_{\vec{b}}|\Pi_{G_{-\vec{a}}}\Pi_{G_{\vec{a}}}|G_{\vec{b}}\rangle = \langle G_{\vec{b}}|\Pi_{-\vec{a}\cdot\vec{H}}|G_{\vec{b}}\rangle = -k\vec{a}\cdot\vec{b}. \tag{2.8.3}$$

So up to a phase, we have that

$$[G_{\vec{a}}, G_{\vec{b}}] = \sqrt{-\vec{a}\cdot\vec{b}}\, G_{\vec{a}+\vec{b}}. \tag{2.8.4}$$

So among other things, (2.8.4) tells us that orthogonal roots commute with each other.

Now consider the inner product

$$\langle G_{\vec{b}}|(\Pi_{G_{-\vec{a}}})^n(\Pi_{G_{\vec{a}}})^n|G_{\vec{b}}\rangle = kn!\prod_{m=1}^{n}\left(-\vec{a}\cdot\vec{b}-\frac{m-1}{2}\vec{a}\cdot\vec{a}\right), \tag{2.8.5}$$

where we obtained the righthand side by using the relation

$$[\Pi_{G_{\vec{a}}}, \left(\Pi_{G_{-\vec{a}}}\right)^n] = n\left(\Pi_{G_{-\vec{a}}}\right)^{n-1}\left(-\frac{n-1}{2}\vec{a}\cdot\vec{a}+\Pi_{\vec{a}\cdot\vec{H}}\right). \tag{2.8.6}$$

The relation in (2.8.6) can be proved by induction.

Now we see that there can be a potential problem. The inner product in (2.8.5) is positive definite, but there are terms in the product on the rhs of (2.8.5) which are negative if $m-1$ is big enough. In order to avoid this problem, we must have

$$-\vec{a}\cdot\vec{b}-\frac{n-1}{2}\vec{a}\cdot\vec{a}=0 \tag{2.8.7}$$

for some positive integer n. Even though we started with the state $|G_{\vec{b}}\rangle$, we could have started with the state $|G_{\vec{a}}\rangle$ and obtained a similar result, since there still cannot be a state $|G_{\vec{a}-\vec{b}}\rangle$ since we assumed that no root was longer than $|\vec{b}|$. Hence proceding as in (2.8.5), we can also derive a relation that

$$-\vec{a}\cdot\vec{b}-\frac{n'-1}{2}\vec{b}\cdot\vec{b}=0. \tag{2.8.8}$$

Let us now go through the possible solutions for (2.8.7) and (2.8.8). Since we assumed that $|\vec{b}| \geq |\vec{a}|$, we see that $n \geq n'$.

1) The first possiblity is that $n = n' = 1$. Therefore,

$$\vec{a}\cdot\vec{b}=0, \tag{2.8.9}$$

so the roots are orthogonal.

For the other cases, neither n or n' is equal to 1. Thus, we have that

$$\vec{a}\cdot\vec{a} = \frac{n'-1}{n-1}\vec{b}\cdot\vec{b} \qquad\qquad \vec{a}\cdot\vec{b} = -\frac{\sqrt{(n-1)(n'-1)}}{2}|\vec{a}||\vec{b}| = \cos\theta|\vec{a}||\vec{b}|. \tag{2.8.10}$$

Therefore, it is necessary that

$$\frac{\sqrt{(n-1)(n'-1)}}{2} \leq 1 \tag{2.8.11}$$

2) The second possibility is that $n = n' = 2$. In this case

$$\vec{a}\cdot\vec{b} = -\frac{1}{2}\vec{a}\cdot\vec{a} = -\frac{1}{2}\vec{b}\cdot\vec{b}. \tag{2.8.12}$$

Hence these roots have equal length and are at an angle of 120 degrees from each other.
3) The third possibility is that $n = n' = 3$. Now we have

$$\vec{a}\cdot\vec{b} = -\vec{a}\cdot\vec{a} = -\vec{b}\cdot\vec{b}. \tag{2.8.13}$$

Hence this case has $\vec{a} = -\vec{b}$.
4) The fourth possibility is $n = 3$, $n' = 2$. In this case

$$\vec{a}\cdot\vec{a} = \frac{1}{2}\vec{b}\cdot\vec{b} \qquad\qquad \cos\theta = -\frac{\sqrt{2}}{2}. \tag{2.8.14}$$

5) The fifth last possibility is $n = 4$, $n' = 2$, hence this has

$$\vec{a}\cdot\vec{a} = \frac{1}{3}\vec{b}\cdot\vec{b} \qquad\qquad \cos\theta = -\frac{\sqrt{3}}{2}. \tag{2.8.15}$$

6) The sixth and last possibility is $n = 5$, $n' = 2$. In this case

$$\vec{a}\cdot\vec{a} = \frac{1}{4}\vec{b}\cdot\vec{b} \qquad\qquad \cos\theta = -1. \tag{2.8.16}$$

However, this last case will not be a legitimate choice. This solution requires that $\vec{b} = -2\vec{a}$ and so $\vec{b} + \vec{a} = -\vec{a}$. Consider then

$$\Pi_{G_{-\vec{a}}}\Pi_{G_{\vec{a}}}|G_{\vec{b}}\rangle = \Pi_{-\vec{a}\cdot\vec{H}}|G_{\vec{b}}\rangle = -\vec{a}\cdot\vec{b}|G_{\vec{b}}\rangle = 2|\vec{a}|^2|G_{-2\vec{a}}\rangle. \tag{2.8.17}$$

But we also have, using the fact that root vectors are unique

$$\Pi_{G_{-\vec{a}}}\Pi_{G_{\vec{a}}}|G_{\vec{b}}\rangle = \Pi_{G_{-\vec{a}}}|[G_{\vec{a}}, G_{-2\vec{a}}]\rangle \sim \Pi_{G_{-\vec{a}}}|G_{-\vec{a}}\rangle = |[G_{-\vec{a}}, G_{-\vec{a}}]\rangle = 0. \tag{2.8.18}$$

Therefore, this is a contradiction. This also tells us that different roots cannot be parallel.

In deriving the above, we said that no vector was longer than $\vec{b}$, but we could have derived the same result so long as either $\vec{b}-\vec{a}$ or $\vec{b}+\vec{a}$ is not a root. But what if both *are* roots, how should we proceed? Well let us suppose that $\vec{b}\cdot\vec{a}\leq 0$ (if this were not true, then we could replace $\vec{a}$ with $-\vec{a}$). Then it must be true that $|\vec{b}-m\vec{a}|>|\vec{a}|$ and $|\vec{b}-m\vec{a}|>|\vec{b}|$ where $m\geq 1$. So for some m we will find that $\vec{b}-m\vec{a}-\vec{a}$ is not a root. In which case, we can proceed as before. Hence we have that

$$\begin{aligned} & \vec{a}\cdot(\vec{b}-m\vec{a})+\frac{n-1}{2}\vec{a}\cdot\vec{a}=0 \\ \Rightarrow \quad & \vec{a}\cdot\vec{b}=\frac{2m+1-n}{2}\vec{a}\cdot\vec{a}, \end{aligned} \tag{2.8.19}$$

where n is some integer. But we have already learned that n can only be $1,2,3$ or 4. For $n=1,2$ we have that $\vec{a}\cdot\vec{b}>0$ which violates are previous assumption. For $n=3$, we can have $\vec{a}\cdot\vec{b}=0$ if $m=1$, but for other m it violates the assumption. For $n=4$, we have that $\vec{a}\cdot\vec{b}=-\frac{1}{2}\vec{a}\cdot\vec{a}$ if $m=1$. All other values of m violate the assumption. In this last case, we also have that

$$\vec{a}\cdot(\vec{b}-\vec{a})+\frac{2-1}{2}(\vec{b}-\vec{a})\cdot(\vec{b}-\vec{a})=0, \tag{2.8.20}$$

hence we find that $\vec{a}\cdot\vec{a}=\vec{b}\cdot\vec{b}$. Therefore, these vectors have equal length and are at 120 degrees angle from each other. In other words, we do not get any new possibilities for the relations between root vectors.

Combining all that we know, and allowing for $\vec{a}\to-\vec{a}$, we have that *any* two root vectors must satisfy one of the following (assuming that $|\vec{b}|\geq|\vec{a}|$):

1) $\vec{a}\cdot\vec{b}=0$
2) $\vec{a}\cdot\vec{b}=\pm\frac{1}{2}\vec{b}\cdot\vec{b}$ $\qquad$ $\vec{a}\cdot\vec{a}=\vec{b}\cdot\vec{b}$
3) $\vec{a}\cdot\vec{b}=\pm\frac{1}{2}\vec{b}\cdot\vec{b}$ $\qquad$ $\vec{a}\cdot\vec{a}=\frac{1}{2}\vec{b}\cdot\vec{b}$
4) $\vec{a}\cdot\vec{b}=\pm\frac{1}{2}\vec{b}\cdot\vec{b}$ $\qquad$ $\vec{a}\cdot\vec{a}=\frac{1}{3}\vec{b}\cdot\vec{b}$

2.9 Positive roots and simple roots

Let us write the root vector in component form, where

$$\vec{a}=(a_1,a_2...a_n). \tag{2.9.1}$$

We say a root is a *positive root* if the first nonzero component in the row vector in (2.9.1) is positive. A *simple root* is a positive root that cannot be written as the sum of two other

positive roots. Clearly, any positive root can be written as a linear combination of simple roots with nonnegative coefficients:

$$\vec{b} = \sum n_i \vec{\alpha}_i \tag{2.9.2}$$

where $\vec{\alpha}_i$ refers to one of the simple roots. To prove this, suppose that $\vec{b}$ is simple, then clearly it is equal to a linear combination of simple roots. If $\vec{b}$ is not simple, then we can write it as the sum of two positive roots. These positive roots are either simple or equal to the sums of positive roots. We keep breaking things down until we are left with simple roots.

We also can easily show that if $\vec{\alpha}_1$ and $\vec{\alpha}_2$ are simple, then $\vec{\alpha}_1 - \vec{\alpha}_2$ is not a root. To see this, note that either $\vec{\alpha}_1 - \vec{\alpha}_2$ or $\vec{\alpha}_2 - \vec{\alpha}_1$ is positive. In the first case, we would then have $\vec{\alpha}_1 = (\vec{\alpha}_1 - \vec{\alpha}_2) + \vec{\alpha}_2$, so $\vec{\alpha}_1$ is the sum of two positive roots and is therefore not simple. In the second case we have $\vec{\alpha}_2 = (\vec{\alpha}_2 - \vec{\alpha}_1) + \vec{\alpha}_1$ so $\vec{\alpha}_2$ is not simple. Since $\vec{\alpha}_1 - \vec{\alpha}_2$ is not a root, then we immediately see based on the discussion in the last section that

$$\vec{\alpha}_1 \cdot \vec{\alpha}_2 \leq 0. \tag{2.9.3}$$

We can now show that the number of simple roots is equal to the rank n of the group. To see this, let us first show that the simple roots are linearly independent. If they were not, then it would be possible to write the equation

$$\sum c_i \vec{\alpha}_i = \sum d_j \vec{\alpha}_j, \tag{2.9.4}$$

where the cofficients c_i and d_j are nonnegative and the simple roots on the lhs of the equation are different from those on the rhs. But then this would imply that if I took the scalar product of both sides of the equation with the rhs, then the rhs would be positive definite, but the lhs would be less than or equal to zero, since $\vec{\alpha}_i \cdot \vec{\alpha}_j \leq 0$, if $i \neq j$. Hence we have a contradiction. Then since the roots live in an n-dimensional space and the simple roots generate all positive roots and are linearly independent, there must then be n of them. Note that if the roots spanned a space that was less than n dimensional, it would mean that there is a combination of Cartan generators

$$\sum_{i=1}^{n} c_i H_i \tag{2.9.5}$$

that commutes with all the root generators. This means that the generator in (2.9.5) commutes with all generators of the Lie Algebra, and so this is not a simple Lie algebra.

In any event, we now see that the properties of the groups are determined by the properties of the simple roots, since these generate all the roots, and so it is only necessary to classify the allowed sets of simple roots.

2.10 Classifying groups continued

It is useful to use a picture description to describe the simple roots and their relationship to one another. This is shown in figure 1. In the figure, a simple root is represented by a

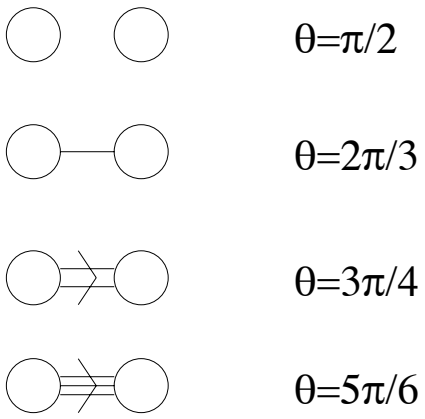


Figure 1: Diagrammatic relations between simple roots

circle. If two simple roots are orthogonal to each other, then the circles are not attached by a line segment. If the simple roots are at an angle of 120 degrees ($2\pi/3$) then we draw one line segment between the circles. If the simple roots are at 135 degrees ($3\pi/4$), then we draw two lines between the circles with the direction of the arrow pointing toward the longer root. Finally, if the roots are at an angle of 150 degrees ($5\pi/6$), then we draw three lines connecting the circles, with the direction of the arrow pointing toward the longer root.

By including all of the simple roots, we can make a chain of these circles with the circles attached to each other by the line segments. Since the group we are considering is simple, we cannot have any disconnected chains. Otherwise this would correspond to having two sets of simple roots spanning orthogonal spaces. In this case we could break up all of the roots and the generators of the Cartan subalgebra into two parts, with every generator in one part commuting with every generator in the other part.

We now show that we cannot have the chains of simple roots shown in figure 2.
If we can find some linear combination of the simple roots such that the square is zero, then it means that the simple roots are not linearly independent. In some cases, we can find linear combinations such that the square is negative. For the chains in figure 2, we have the following results:

$$(\vec{\alpha}_1 + 2\vec{\alpha}_2 + \vec{\alpha}_3)^2 = 0 \qquad (\vec{\alpha}_1 + 2\vec{\alpha}_2 + 3\vec{\alpha}_3)^2 = 0 \qquad (\vec{\alpha}_1 + 2\vec{\alpha}_2 + 3\vec{\alpha}_3)^2 = -(\vec{\alpha}_2)^2$$

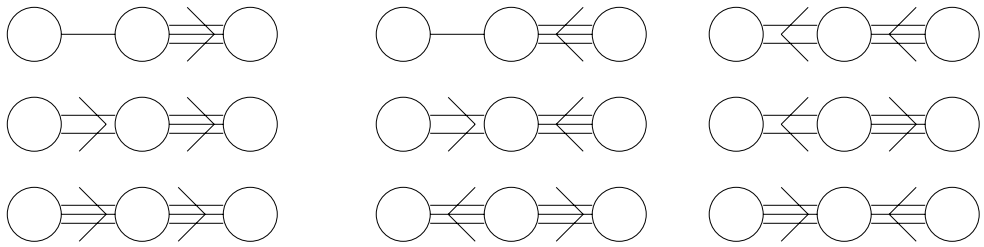
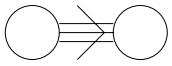


Figure 2: Chains with at least one triple link.

$$\begin{array}{lll} (\vec{\alpha}_1+2\vec{\alpha}_2+\vec{\alpha}_3)^2=-(\vec{\alpha}_1)^2 & (\vec{\alpha}_1+\vec{\alpha}_2+\vec{\alpha}_3)^2 = -\dfrac{(\vec{\alpha}_3)^2}{2} & (\vec{\alpha}_1+2\vec{\alpha}_2+\vec{\alpha}_3)^2=-(\vec{\alpha}_2)^2 \\ (\vec{\alpha}_1+2\vec{\alpha}_2+\vec{\alpha}_3)^2=-2(\vec{\alpha}_1)^2 & (\vec{\alpha}_1+2\vec{\alpha}_2+\vec{\alpha}_3)^2 = -2(\vec{\alpha}_2)^2 & (\vec{\alpha}_1+\vec{\alpha}_2+\vec{\alpha}_3)^2=-(\vec{\alpha}_1)^2 \end{array} \tag{2.10.1}$$

Clearly, we cannot have any chains where these above chains are subchains, since the relations in (2.10.1) do not rely on the possible connections to other simple roots. This gives a very strong constraint on the possible lie algebras. It means that the only allowed chain with a triple line is



that is, the chain with only two simple roots. The corresponding Lie Group is one of the *exceptional* Lie groups and is called G_2. This chain is the first example of a *Dynkin diagram*, a chain that corresponds to a Lie group. We will say more about the group G_2 later.

We can also rule out the sets of chains with no triple lines but at least one double line shown in figure 3. The numbers inside the circles in figure 3 indicate the number of simple roots in a linear combination whose length squared is zero. For example, for the first chain, the inner product with these coefficients is

$$(2\vec{\alpha}_1+2\vec{\alpha}_2+\vec{\alpha}_3)^2 = 4\vec{\alpha}_1^2+8\vec{\alpha}_1\cdot\vec{\alpha}_2+4\vec{\alpha}_2^2+4\vec{\alpha}_2\cdot\vec{\alpha}_3+\vec{\alpha}_3^2 = (4-8+8-8+4)\vec{\alpha}_1^2. \tag{2.10.2}$$

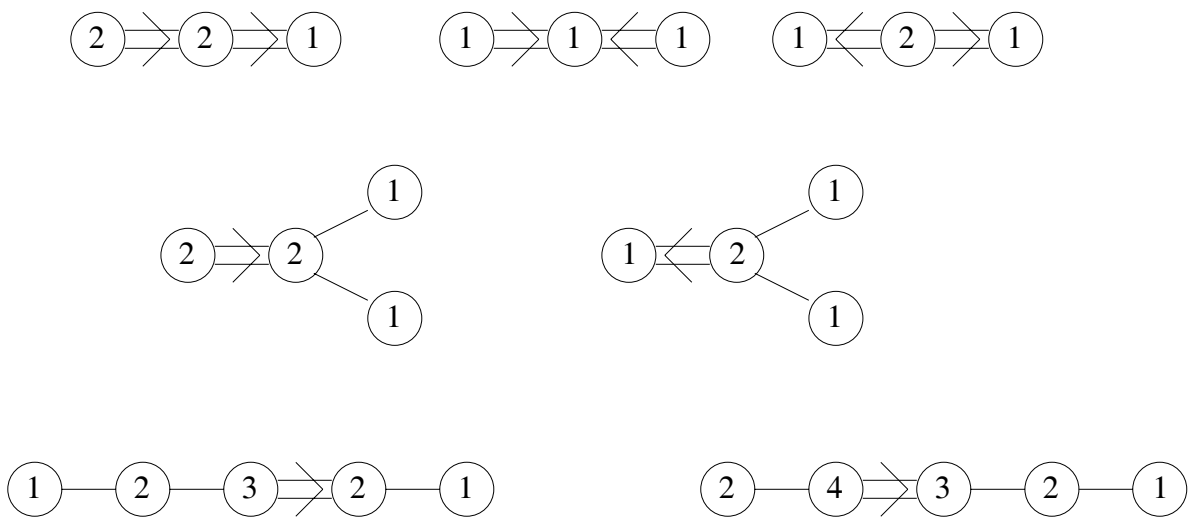
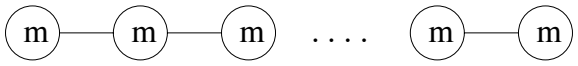


Figure 3: Chains with no triples but at least one double.

Furthermore, if we replace the middle circle in the chain with coefficient m with the chain



where the coefficients of every simple root in the chain are also m, then we find that the linear combination of simple roots still has zero length.

Hence, we learn that the only sets of chains, that is Dynkin diagrams, that do not have linearly dependent simple roots have at *most* one double line. Following our previous arguments, we learn that the allowed chains with one double line are highly restricted and have the form in figure 4.

The first two sets of Dynkin diagrams are called B_n and C_n, each with n simple roots. The corresponding groups are the groups $SO(2n+1)$ and $Sp(n)$. The first group is the special orthogonal group, that is the group of rotations in $2n+1$ directions. The other group is the symplectic group in n complex dimensions. The last Dynkin diagram is for another exceptional group, known as F_4.

To complete the classification, we need to study chains with single lines only. Such sets of chains, and the allowed Dynkin diagrams, are called *simply laced*. It is straightforward to check that the chains in figure 5 are not allowed, since they will lead to linearly dependent simple roots. In the first chain in figure 5, we can replace one simple root with

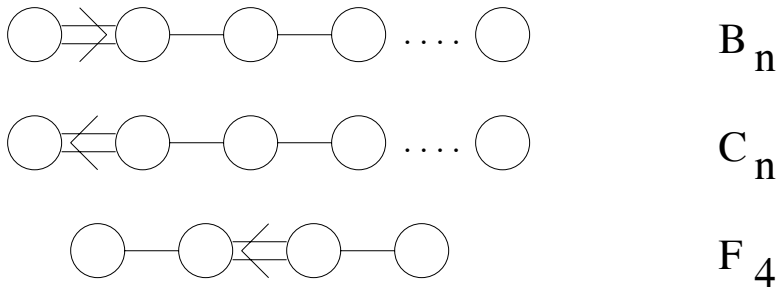


Figure 4: Dynkin Diagrams for root systems with one double line

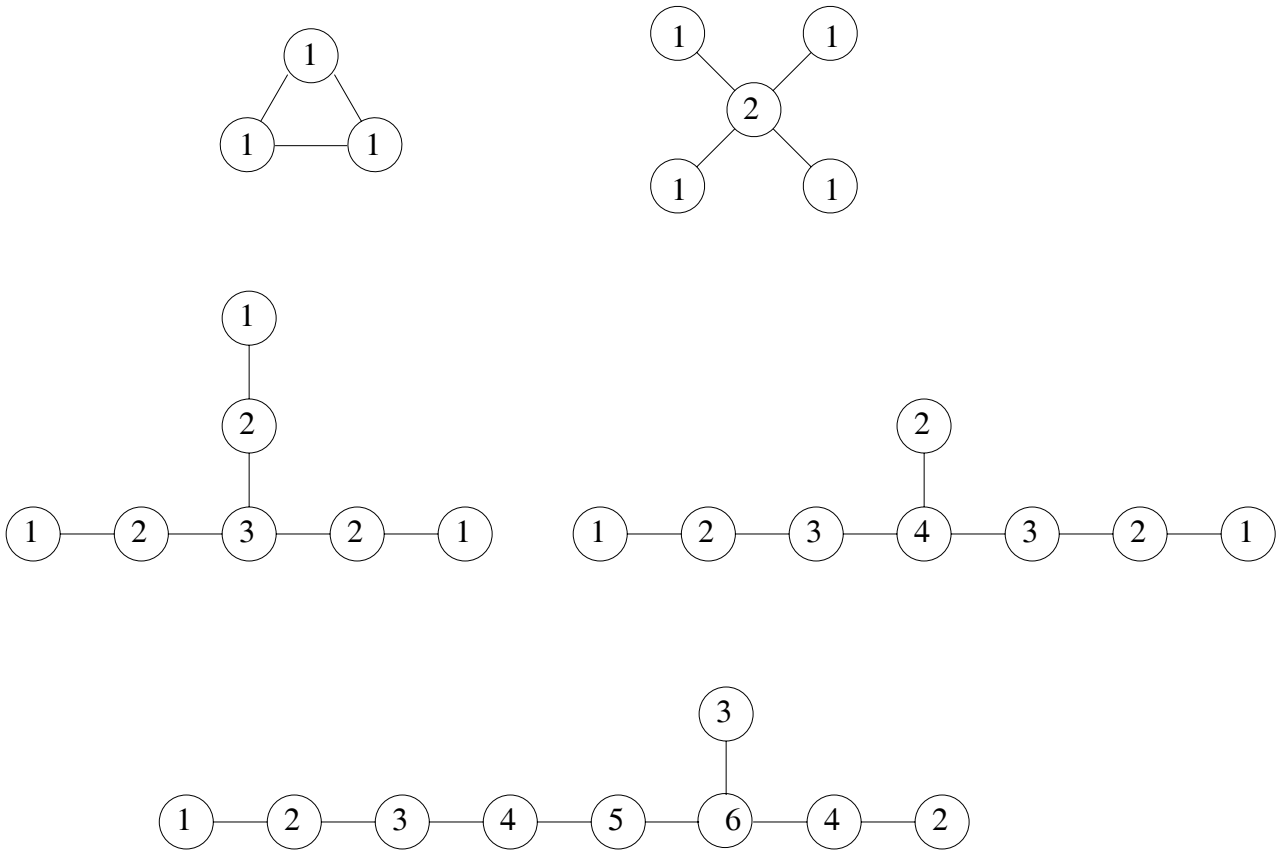


Figure 5: Simply laced diagrams with linear dependencies for the simple roots.

the chain of simple roots in figure 3, and again find that the there is a linear dependence in the simple roots. Hence, it is not possible to have any closed loop in a chain. Likewise, for the second chain in figure 5, we can replace the center simple root with the chain

in figure 3 with coefficient 2, and still find that the simple roots are linearly dependent. The last three diagrams also have linearly dependent simple roots, with the coefficients relating the simple roots given in the circles. Hence, we have shown that the only simply laced Dynkin diagrams are of the form in figure 6. The Dynkin diagrams A_n is for the

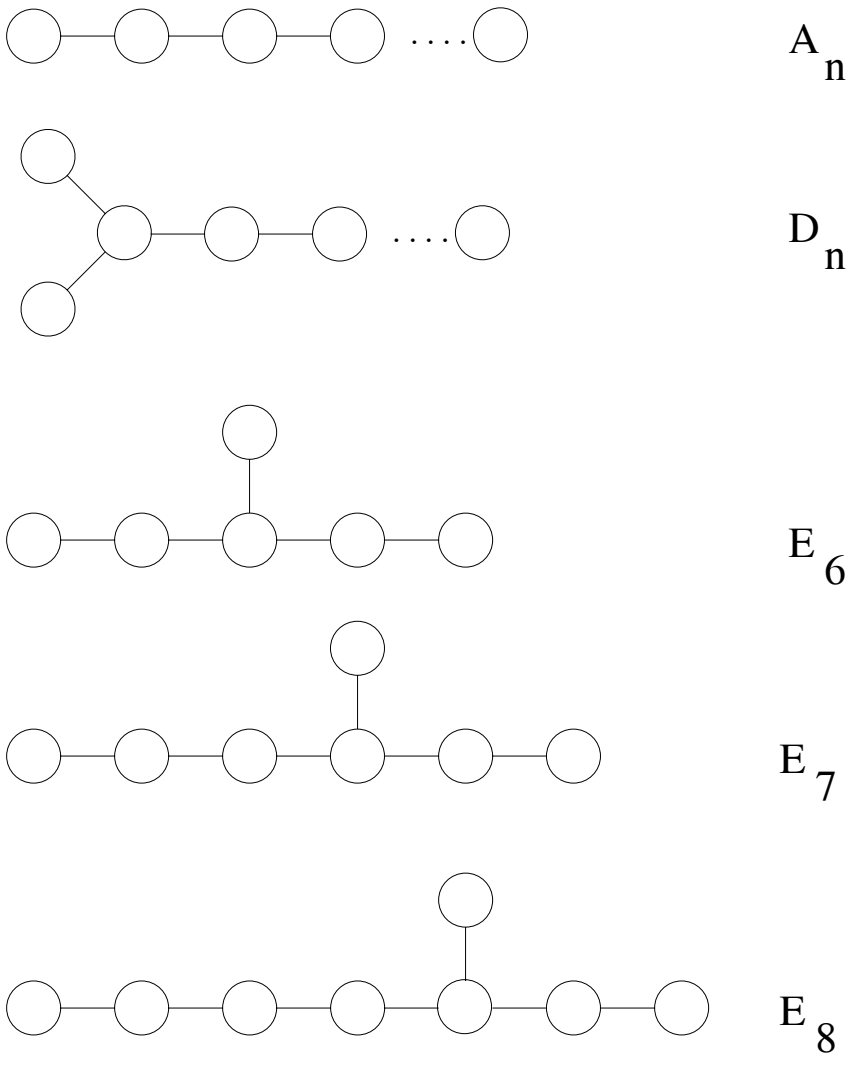


Figure 6: The simply laced Dynkin diagrams

Lie groups $SU(n+1)$. The diagrams D_n are for the groups $SO(2n)$. The last three are for the exceptional groups E_6, E_7, E_8.

So that is it! The only simple compact Lie groups are $SU(n)$, $SO(n)$, $Sp(n)$ and the exceptional groups G_2, F_4, E_6, E_7 and E_8.

One outgrowth of this, is that we can see that some groups are the same, or at least almost the same. For example, we see that $SU(4)$, which has the Dynkin diagram A_3 and $SO(6)$, which has the Dynkin diagram D_3, are the same since the diagrams are the same. Likewise $SO(4)$ is actually a product of $SU(2)$ groups, since D_2 is actually two A_1 diagrams. In this same way, we can see that $Sp(1)$ and $SO(3)$ are very similar to $SU(2)$ and that $Sp(2)$ is the same as $SO(5)$.

2.11 Examples

1) $SU(2)$ (A_1). This group has one simple root, $\vec{\alpha}$, and hence only two roots, $\pm\vec{\alpha}$. There is one element in the Cartan subalgebra, $H_{\vec{\alpha}}$, and so the commutation relations of H with $G_{\pm\vec{\alpha}}$ are

$$[H_{\vec{\alpha}}, G_{\pm\vec{\alpha}}] = \pm\alpha G_{\pm\vec{\alpha}} \qquad\qquad [G_{\vec{\alpha}}, G_{-\vec{\alpha}}] = \alpha H_{\vec{\alpha}}. \tag{2.11.1}$$

Hence, after an appropriate scaling of H and G, this is the algebra for the angular momentum operators, with $H_{\vec{\alpha}} = \alpha J_z$ and $G_{\pm\vec{\alpha}} = \frac{\alpha}{\sqrt{2}} J_\pm$.

It is customary to choose the length of the simple roots in the simply laced diagram to have length squared 2. However, this choice is arbitrary. Some books choose the length squared to be 1, which will be the convention that we follow, unless specifically mentioned otherwise.

2) $SU(3)$ (A_2). This has 2 simple roots of length squared 1 at 120 degrees from each other. If we add the two roots together, we have one more root with length squared 1. These three roots, along with the three negative roots gives 6 roots in total. Combined with the two elements of the Cartan subalgebra, we find 8 generators. This is as expected, since 3 by 3 unitary matrices have 9 generators, but if the the matrices are constrained to have determinant 1, then one of the generators is removed. The roots for $SU(3)$ are shown in figure 7. Note that if we take the subset of roots consisting of one simple root and its negative, then this forms the root system of $SU(2)$, hence $SU(2)$ is a subgroup of $SU(3)$.

3) G_2. This has 2 simple roots, with one root of length squared 1 and the other of length squared 1/3. The root diagram is shown in figure 8. There are 12 roots, and hence 14 generators in all. G_2 is a subgroup of $SO(7)$, the group of rotations in 7 dimensions. The roots are shown in figure 8.

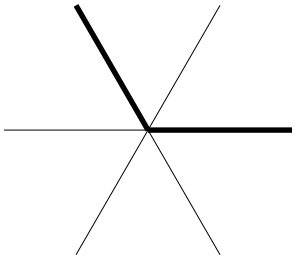


Figure 7: Roots for $SU(3)$ (A_2). The simple roots are denoted by the bold lines.

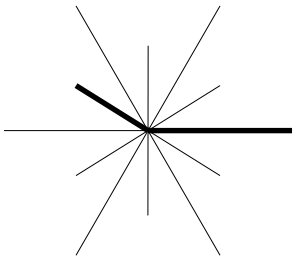


Figure 8: Roots for G_2. The simple roots are denoted by the bold lines. Note that the long roots of G_2 are identical to the $SU(3)$ roots. Hence $SU(3)$ will be a subgroup of G_2.

2.12 The Classical Lie Groups

The groups $SU(n)$, $SO(n)$ and $Sp(n)$ are called the *classical Lie groups*. The others are called the *exceptional Lie groups*. The classical groups are most easily described by considering what is left invariant under their transformations.

1) $SU(n)$. Consider two complex n dimensional vectors z_i and w_i, where i is an index that runs from 1 to n. Consider the quantity $w_i^* z_i = w^\dagger z$. Then if we transform z and w by a unitary transformation, they become $U_{ij} z_j$ and $U_{ij} w_j$, where the repeated j index is summed over. Clearly, $w^\dagger z$ is invariant under this transformation.

Now let us see why the A_{n-1} Dynkin diagrams correspond to the $SU(n)$ Lie algebra. To show this, suppose we consider a basis in an n dimensional space. Let us consider vectors in this space which have length squared 2 and which have the form

$$(1,-1,0...),(0,1,-1,0..0),...(0..0,1,-1,0..0),(0..0,1,-1) \tag{2.12.2}$$

Clearly there are $n-1$ such vectors and that the inner product of these vectors with themselves is described by the A_{n-1} dynkin diagram. Let us assume that these are the

simple roots. Then all of the positive roots have the form

$$(0, 0..0, 1, 0..0, -1, 0..0) \tag{2.12.3}$$

Hence there are $n(n-1)/2$ positive roots and an equal number of negative roots. Note that these roots span an $n-1$ dimensional space, since the sum of the components is always 0, and so all of the roots lie in the plane $x_1 + x_2 + ..x_n = 0$.

Given that these are the root vectors, let us find matrices that have the correct commutation relations. It is not hard to see that the matrix elements for the Cartan subalgebra should have the form $(H_i)_{kl} = \delta_{ik}\delta_{il} - \frac{1}{n}\delta_{kl}$ and that the root vectors should have the form $(G_{ij})_{kl} = \delta_{ik}\delta_{jl}$. The Cartan subalgebra has been constructed to be traceless. The trace piece will commute with all the other generators. Since the Cartan subalgebra has traceless matrices, it has only $n-1$ linearly independent generators. The matrices H_i, $G_{ij}+G_{ji}$ and $iG_{ij}-iG_{ji}$ generate all $n \times n$ traceless hermitian matrices and these generate all $n \times n$ unitary matrices with determinant 1.

2) $SO(m)$. $SO(m)$ transformations leave invariant $x_i y_i$, where x_i and y_i are chosen to be real. The generators of a rotation are given by

$$(M_{ij})_{kl} = i\delta_{ik}\delta_{jl} - i\delta_{il}\delta_{jk} \tag{2.12.4}$$

which generates a rotation in the ij plane. These generators form the Lie algebra

$$[M_{ij}, M_{kl}] = i(\delta_{il}M_{jk} - \delta_{ik}M_{jl} - \delta_{jl}M_{ik} + \delta_{jk}M_{il}). \tag{2.12.5}$$

If $m = 2n$, then we can choose as a Cartan basis

$$H_i = M_{2i-1,2i}. \tag{2.12.6}$$

Using (2.12.5) and (2.12.6), it is straightforward to show that the root generators are given by

$$G_{\pm i \pm j} = M_{2i,2j-1} \pm iM_{2i-1,2j-1} - (\pm)iM_{2i,2j} - (\pm)(\pm)M_{2i-1,2j} \tag{2.12.7}$$

so in terms of the root vectors, these are

$$(0, ..0, \pm 1, ..0, \pm 1, ..0) \tag{2.12.8}$$

where the i and j entry are nonzero. It is then easy to show that the simple roots are

$$(1, -1, 0..0), (0, 1, -1, ..0)..(0, ..0, 1, -1), (0, ..0, 1, 1) \tag{2.12.9}$$

Hence there are n of these roots. It is straightforward to show that the inner product of the simple roots with themselves is given by the D_n Dynkin diagram.

For $m = 2n + 1$, we can have the same generators in the Cartan subalgebra and the root generators in (2.12.7) are also included. In addition, we have the root generators

$$G_{\pm i} = M_{2i,2n+1} \pm iM_{2i-1,2n+1}, \tag{2.12.10}$$

whose vectors are

$$(0..0, \pm 1, 0..0) \tag{2.12.11}$$

where the i index is nonzero. The simple roots, can then be shown to have the form

$$(1, -1, 0..0), (0, 1, -1, ..0)..(0, ..0, 1, -1), (0, ..0, 0, 1) \tag{2.12.12}$$

Note that the last root has a length squared that is $1/2$ the other simple roots. It is then straightforward to show that the inner product of these roots is given by the B_n Dynkin diagram.

3) $Sp(n)$. The symplectic group leaves invariant the symplectic product

$$a_i^1 b_i^2 - a_i^2 b^1 i = (a^1)^T b^2 - (a^2)^T b^1, \tag{2.12.13}$$

where a^1, a^2, b^1 and b^2 are assumed to be n dimensional complex vectors. Notice that the symplectic product is preserved under the transformation $a^\alpha \to \exp(iW)a^\alpha$, $b^\alpha \to \exp(iW)b^\alpha$ where W is any hermitian and *antisymmetric* matrix and $\alpha = 1, 2$. But it is also invariant under $a^\alpha \to \exp(i(-1)^\alpha V)a^\alpha$, $b^\alpha \to \exp(i(-1)^\alpha V)b^\alpha$ where V is any hermitian *symmetric* matrix. Finally, we can have transformations that rotate a^1 into a^2, which have the form $a_1 \to \cos Va_1 + i \sin Va_2$ and $a_2 \to i \sin Va_1 + \cos Va_2$, where V is hermitian and symmetric (and hence real), or have the form $a_1 \to \cos Va_1 + \sin Va_2$ and $a_2 \to -\sin Va_1 + \cos Va_2$. Hence, we can express the complete set of generators as a tensor product

$$1 \otimes W + \sigma_i \otimes V_i, \tag{2.12.14}$$

where the σ_i are the Pauli matrices and W is antisymmetric hermitian $n \times n$ matrix and the V_i are symmetric hermitian $n \times n$ matrices. In matrix form, the generators in (2.12.14) can be written as

$$\begin{pmatrix} W + V_3 & V_1 - iV_2 \\ V_1 + iV_2 & W - V_3 \end{pmatrix}. \tag{2.12.15}$$

There are n elements in the Cartan subalgebra, which in terms of the $2n \times 2n$ matrices have the form

$$(H_i)_{kl} = \delta_{ik}\delta_{il} - \delta_{i+n,k}\delta_{i+n,l}. \tag{2.12.16}$$

There are then 3 distinct types of root vectors. The first have the form

$$(G_{ij})_{kl} = \delta_{ik}\delta_{jl} - \delta_{j+n,k}\delta_{i+n,l}, \tag{2.12.17}$$

the second have the form

$$(G'_i)_{kl} = \delta_{i+n,k}\delta_{il}, \tag{2.12.18}$$

and the third have the form

$$(G''_{ij})_{kl} = \delta_{i+n,k}\delta_{jl} + \delta_{j+n,k}\delta_{i,l}. \tag{2.12.19}$$

The commutators of these root generators with the Cartan subalgebra is given by

$$\begin{aligned}
[H_i, G_{jk}] &= (\delta_{ij} - \delta_{ik})G_{jk} \\
[H_i, G'_j] &= -2\delta_{ij}G'_j \\
[H_i, G''_{jk}] &= -(\delta_{ij} + \delta_{ik})G''_{jk}.
\end{aligned} \tag{2.12.20}$$

Including the adjoints of these root vectors, we see that the complete set of roots have components of the form

$$(0..0 \pm 1, 0..0, \pm 1, 0..0) \qquad \text{or} \qquad (0..0, \pm 2, 0..0). \tag{2.12.21}$$

The n simple roots then are given by

$$(1, -1, 0..0), ...(0, 0..0, 1, -1), (0..0, 2) \tag{2.12.22}$$

Clearly, the last root has length squared that is twice the length of the other roots. It is also clear that the inner products between the simple roots is given by the C_n Dynkin diagram.

2.13 Representations and weights

Up to now, we have been considering the groups in terms the adjoint representation. This is reflected in the fact that the transformation on the ket state $|T_b\rangle$ is

$$\Pi_{T_a}|T_b\rangle = |[T_a, T_b]\rangle = f_{abc}|T_c\rangle. \tag{2.13.1}$$

Hence the number of states in this representation is equal to the number of generators of the Lie algebra. So for example, for $SU(2)$, we would find three states in the adjoint representation. This corresponds to the spin 1 angular momentum states.

But we know from our experience with angular momentum that there are an infinite number of different angular momentum states, labeled by the quantum number j, with j either integer or half integer. These different values for j correspond to different $SU(2)$ irreducible representations, with $j = 1$ being the adjoint representation. In particular, there is a smaller, but nontrivial representation, with $j = 1/2$. This nontrivial representation is known as a *fundamental representation.* In the case of $SU(2)$, it is 2 dimensional (spin "up" and spin "down").

We now show how to find other representations for the other compact Lie groups. In the case of $SU(2)$, remember that the states in the representation were found by acting with raising and lowering operators J_+ and J_-, until eventually the states were annihilated. This put a constraint on the values of j. It should be clear from the discussion in sections (2.7)-(2.9), that the positive roots play the role of the raising operators and the negative roots play the role of the lowering operators.

We can write a state in terms of its *weight vector* $\vec{\mu}$, as $|\vec{\mu}\rangle$ which satisfies

$$\Pi_{H_i}|\vec{\mu}\rangle = \mu_i|\vec{\mu}\rangle \tag{2.13.2}$$

for all elements in the Cartan subalgebra. For any given representation, there must be some state $|\vec{\mu}_{max}\rangle$ such that this state is annihilated by all positive root operators. The weight $\vec{\mu}_{max}$ is called the *highest weight* of the representation. So just as in $SU(2)$, where the representation was labeled by j, which is the maximum eigenvalue of J_z, we have that the representation in this general case is labeled by the value of the highest weight.

Now if the weight is annihilated by all the positive root operators, then it is clearly annihilated by all the simple root operators. In this case, we can proceed as we did in section (2.8) and hence find essentially the same equation as in (2.8.5), that is

$$\langle\vec{\mu}|\left(\Pi_{G_{\vec{\alpha}_i}}\right)^{n_i}\left(\Pi_{G_{-\vec{\alpha}_i}}\right)^{n_i}|\vec{\mu}\rangle = Cn!\prod_{m=1}^{n_i}\left(\vec{\alpha}_i\cdot\vec{\mu} - \frac{m-1}{2}\vec{\alpha}_i\cdot\vec{\alpha}_i\right), \tag{2.13.3}$$

where C is a normalization constant. Thus we find that for every simple root, the highest weights must satisfy an equation of the form

$$\vec{\alpha}_i\cdot\vec{\mu} = \frac{q_i}{2}\vec{\alpha}_i\cdot\vec{\alpha}_i, \tag{2.13.4}$$

where the q_i are nonnegative integers. This is to insure that a manifestly positive definite quantity is not actually negative. A particular useful class of representations are the *fundamental representations*, where the fundamental representation for root $\vec{\alpha}_i$ has $q_j = \delta_{i,j}$.

Let us consider some examples:
1) $SU(2)$: We have only one simple root, so we find for the fundamental representation that the weight has 1/2 the length of the root.
2) $SU(3)$: Now we have two simple roots, hence there are two fundamental represenations. We can write the highest weights as a linear combination of roots with fractional coefficients

$$\vec{\mu}_1 = c_1\vec{\alpha}_1 + c_2\vec{\alpha}_2 \qquad\qquad \vec{\mu}_2 = c_1'\vec{\alpha}_1 + c_2'\vec{\alpha}_2. \tag{2.13.5}$$

In the first case we have

$$(c_1\vec{\alpha}_1 + c_2\vec{\alpha}_2)\cdot\vec{\alpha}_1 = \frac{1}{2}\vec{\alpha}_1\cdot\vec{\alpha}_1, \qquad (c_1\vec{\alpha}_1 + c_2\vec{\alpha}_2)\cdot\vec{\alpha}_2 = 0 \qquad \Rightarrow \vec{\mu}_1 = \frac{2}{3}\vec{\alpha}_1 + \frac{1}{3}\vec{\alpha}_2, \tag{2.13.6}$$

while in the second case we have

$$\vec{\mu}_2 = \frac{1}{3}\vec{\alpha}_1 + \frac{2}{3}\vec{\alpha}_2. \tag{2.13.7}$$

Let us now find the other weights in these representations. We will write the ket states in terms of the components of the simple roots that make up the weights. In the first case, we act on the state $|2/3, 1/3\rangle$ first with $\Pi_{-\vec{\alpha}_1}$ giving $C|-1/3, 1/3\rangle$, where C is a normalization constant. We now act with $\Pi_{-\vec{\alpha}_2}$ on this state, giving $|-1/3, -2/3\rangle$. The three elements $|2/3, 1/3\rangle$, $|-1/3, 1/3\rangle$ and $|-1/3, -2/3\rangle$ make up this representation. This fundamental representation is written as **3** (or sometimes as $\underline{\mathbf{3}}$)

We can derive the weight vectors for the other fundamental representation by interchanging the two simple roots. Hence, the states in this representation, which is called $\overline{\mathbf{3}}$ are $|1/3, 2/3\rangle$, $|1/3, -1/3\rangle$ and $|-2/3, -1/3\rangle$. Comparing these two representations, we see that weights of one are negative the weights of the other representation. Recalling that the adjoint conjugation takes the root vectors to minus themselves, we see that the fundamental $SU(3)$ representations are conjugate to each other. But this also means that the fundamental representations are not real representations. Figure 9 shows the two fundamental representations.

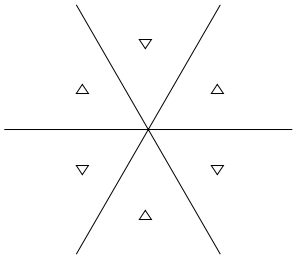


Figure 9: Fundamental representations for $SU(3)$. One representation has triangles pointing up, while the other has the triangles pointing down. The root vectors are also shown.

2.14 A more general weight formula

Let us assume that we have a weight state $|\mu\rangle$ that is not necessarily a highest weight. However, if we act with a root vector $G_{\vec{a}}$ enough times, eventually we will annihilate the state. Hence, there exists a nonnegative integer p such that[3]

$$(G_{\vec{a}})^{p+1}|\vec{\mu}\rangle = 0. \tag{2.14.1}$$

Likewise, there exists a nonnegative integer q such that

$$(G_{-\vec{a}})^{q+1}|\vec{\mu}\rangle = 0. \tag{2.14.2}$$

Consider then the inner product of the commutator

$$\langle\vec{\mu}|[G_{\vec{a}}, G_{-\vec{a}}]|\vec{\mu}\rangle = \langle\vec{\mu}|\vec{a}\cdot\vec{H}|\vec{\mu}\rangle = \vec{a}\cdot\vec{\mu} \tag{2.14.3}$$

where the state $|\vec{\mu}\rangle$ is assumed to be properly normalized. Now we have that

$$G_{\vec{a}}|\vec{\mu}\rangle = N_{\vec{a},\vec{\mu}}|\vec{\mu}+\vec{a}\rangle, \tag{2.14.4}$$

where $|\vec{\mu}+\vec{a}\rangle$ is the normalized state, that is

$$\langle\vec{\mu}+\vec{a}|G_{\vec{a}}|\vec{\mu}\rangle = N_{\vec{a},\vec{\mu}}. \tag{2.14.5}$$

Likewise, we have that

$$G_{-\vec{a}}|\vec{\mu}\rangle = N_{-\vec{a},\vec{\mu}}|\vec{\mu}-\vec{a}\rangle. \tag{2.14.6}$$

Taking the complex conjugate of (2.14.4), we see that

$$(N_{\vec{a},\vec{\mu}})^* = N_{-\vec{a},\vec{\mu}+\vec{a}}. \tag{2.14.7}$$

[3]From now on, we will write Π_{T_a} simply as T_a, where the fact that these are operators is implied.

Now these last sets of equations hold for a general weight in any representation, so in particular they hold for $|\mu + n\vec{a}\rangle$ where n is any integer between $-q$ and $+p$. Hence, using (2.14.3) we can write the series of equations

$$\begin{aligned}
|N_{\vec{a},\vec{\mu}+p\vec{a}}|^2 - |N_{\vec{a},\vec{\mu}+(p+1)\vec{a}}|^2 &= \vec{a}\cdot(\vec{\mu}+p\vec{a}) \\
|N_{\vec{a},\vec{\mu}+(p-1)\vec{a}}|^2 - |N_{\vec{a},\vec{\mu}+p\vec{a}}|^2 &= \vec{a}\cdot(\vec{\mu}+(p-1)\vec{a}) \\
|N_{\vec{a},\vec{\mu}+(p-2)\vec{a}}|^2 - |N_{\vec{a},\vec{\mu}+(p-1)\vec{a}}|^2 &= \vec{a}\cdot(\vec{\mu}+(p-2)\vec{a}) \\
..... &= .... \\
|N_{\vec{a},\vec{\mu}-(q-1)\vec{a}}|^2 - |N_{\vec{a},\vec{\mu}-(q-2))\vec{a}}|^2 &= \vec{a}\cdot(\vec{\mu}-(q-1)\vec{a}) \\
|N_{\vec{a},\vec{\mu}-q\vec{a}}|^2 - |N_{\vec{a},\vec{\mu}-(q-1))\vec{a}}|^2 &= \vec{a}\cdot(\vec{\mu}-q\vec{a})
\end{aligned} \tag{2.14.8}$$

Because of (2.14.1) and (2.14.2) we have that $N_{\vec{a},\vec{\mu}+(p+1)\vec{a}} = N_{\vec{a},\vec{\mu}-q\vec{a}} = 0$. Then, if we add up both sides of the equations in (2.14.8), we see that the left hand side all cancels and we are left with

$$\begin{aligned}
0 &= (p+q+1)\vec{a}\cdot\vec{\mu} + \sum_{n=1}^{p} p\vec{a}\cdot\vec{a} - \sum_{m=1}^{p} q\vec{a}\cdot\vec{a} \\
&= (p+q+1)\vec{a}\cdot\vec{\mu} + \vec{a}\cdot\vec{a}\left(\frac{p(p+1)}{2} - \frac{q(q+1)}{2}\right) \\
&= (p+q+1)\left(\vec{a}\cdot\vec{\mu} + \frac{p-q}{2}\vec{a}\cdot\vec{a}\right).
\end{aligned} \tag{2.14.9}$$

Hence, a general weight μ and a general root $\vec{a}$ satisfy

$$\vec{a}\cdot\vec{\mu} = \frac{q-p}{2}\vec{a}\cdot\vec{a}. \tag{2.14.10}$$

Let us use this to find the weights of one of the fundamental representations of $SU(5)$. Let us write the simple roots as

$$\vec{\alpha}_1 = (1,-1,0,0,0), \quad \vec{\alpha}_2 = (0,1,-1,0,0), \quad \vec{\alpha}_3 = (0,0,1,-1,0), \quad \vec{\alpha}_4 = (0,0,0,1,-1). \tag{2.14.11}$$

The highest weight state for the first fundamental representation has a weight $\vec{\mu}_1$ which satisfies

$$\vec{\mu}_1\cdot\vec{\alpha}_1 = \frac{1}{2}\vec{\alpha}_1\cdot\vec{\alpha}_1 = 1, \quad \vec{\mu}_1\cdot\vec{\alpha}_i = 0, \quad i \neq 1. \tag{2.14.12}$$

Thus,

$$\vec{\mu}_1 = (1,0,0,0,0), \tag{2.14.13}$$

and the next weight in the representation is

$$\vec{\mu}_2 = \vec{\mu}_1 - \vec{\alpha}_1 = (0, 1, 0, 0, 0). \tag{2.14.14}$$

Now, we note that $\vec{\mu}_2 \cdot \vec{\alpha}_2 = -1$. Therefore, for this root and weight, $q - p = 1$. Hence, it must be true that

$$\mu_3 = \mu_2 - \vec{\alpha}_2 = (0, 0, 1, 0, 0) \tag{2.14.15}$$

is a weight. In order for $\mu_2 - 2\vec{\alpha}_2$ to be a weight, it would be necessary for $\mu_2 + \vec{\alpha}_2$ to also be a weight. But notice that

$$\mu_2 + \vec{\alpha}_2 = \mu_1 - (\vec{\alpha}_1 - \vec{\alpha}_2) \tag{2.14.16}$$

so in order for $\mu_2 + \vec{\alpha}_2$ to be a weight $\vec{\alpha}_1 - \vec{\alpha}_2$ would have to be a root. But the difference of simple roots is not a root. We can continue these arguments and generate the other two weights

$$\mu_4 = (0, 0, 0, 1, 0) \quad \mu_5 = (0, 0, 0, 0, 1) \tag{2.14.17}$$

to fill out the representation.

As another example, let us consider the adjoint representation of $SO(9)$. The weights in the representation are the zero weight states corresponding to the elements of the Cartan subalgebra, and the roots. The simple roots are given by

$$\vec{\alpha}_1 = (1, -1, 0, 0), \quad \vec{\alpha}_2 = (0, 1, -1, 0), \quad \vec{\alpha}_3 = (0, 0, 1, -1), \quad \vec{\alpha}_4 = (0, 0, 0, 1). \tag{2.14.18}$$

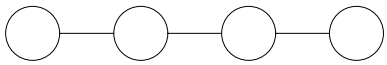
Consider the root $\vec{\alpha}_0 = (1, 0, 0, 0) = -\vec{\alpha}_1 - \vec{\alpha}_2 - \vec{\alpha}_3 - \vec{\alpha}_4$. Its inner product with $\vec{\alpha}_4$ is

$$\vec{\alpha}_0 \cdot \vec{\alpha}_4 = 0. \tag{2.14.19}$$

So we conclude that if $\vec{\alpha}_0 + \vec{\alpha}_4$ is a root, then so is $\vec{\alpha}_0 - \vec{\alpha}_4$ in order that $p - q = 0$. Both combinations are indeed roots.

2.15 Representations for Subgroups

From the simple roots of a group we can determine what the subgroups are. For example, the Dynkin diagram of $SU(5)$ looks like



If we were to remove one of the simple roots, then we would be left with

These are the Dynkin diagrams for $SU(2)$ and $SU(3)$, so they both are subgroups of $SU(5)$. More to the point, they are both subgroups simultaneously, in other words there is an $SU(3) \times SU(2)$ subgroup of $SU(5)$. Moreover, there is one linear combination of the $SU(5)$ Cartan subalgebra that commutes with both the $SU(2)$ roots and the $SU(3)$ roots. Hence there is an additional $U(1)$ subgroup. Hence, $SU(5)$ has an $SU(3) \times SU(2) \times U(1)$ subgroup.

A subgroup is called *semisimple* if it has no $U(1)$ factors, hence $SU(3) \times SU(2)$ is semisimple, but $SU(3) \times SU(2) \times U(1)$ is not semisimple. A semisimple subgroup is called *maximal* if the rank of the subgroup is the same as the original group. Hence $SU(3) \times SU(2)$ is not a maximal subgroup, since its rank is 3, but the rank of $SU(5)$ is 4.

To find maximal subgroups, we need to consider an object known as the *extended Dynkin diagram*. Recall that in our classification of the groups, we found many chains that could not correspond to a Lie algebra since there was a linear relation between the simple roots. However, we can use this to our advantage to find maximal subgroups. An extended Dynkin diagram is found by taking the original Dynkin diagram with n simple roots and adding one circle to the diagram such that the new diagram is connected and has a linear relation between the $n + 1$ simple roots. We have basically determined what these diagrams are already when we ruled out the chains in the previous sections. The extended Dynkin diagrams for the various groups are shown in figure 12. The tildes over the diagram labels indicate that these are the extended diagrams. If we now remove a simple root from the extended diagram, we are left with a diagram that has the same rank as the original Lie group. Furthermore, this is a subgroup of the original group since it is made up of roots from the original Lie algebra. What is more, there is no longer a

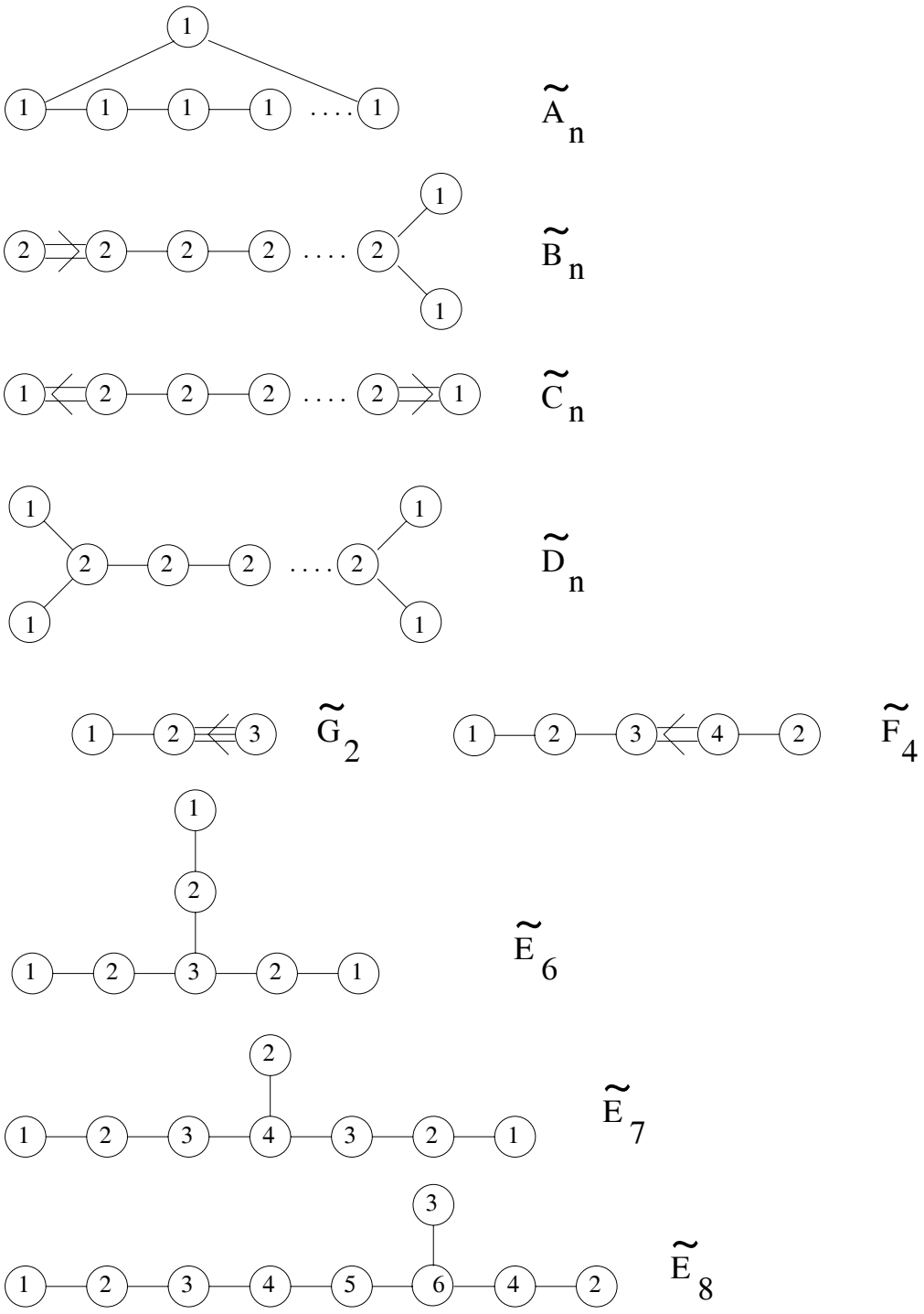


Figure 10: The extended Dynkin diagrams for the compact simple groups. The numbers inside the circles indicate the coefficients for the linear relation.

linear relation between the simple roots, since we have removed one of the simple roots. Hence this is a Dynkin diagram for a semi-simple Lie algebra.

Let us consider a few examples. First for $SU(n)$, it is clear that when we perform this surgery on the extended diagram we end up with the original $SU(n)$ diagram, hence $SU(n)$ has no semi-simple maximal subgroups.

Next consider G_2. If we remove the short root from the extended diagram, we end up with the $SU(3)$ Dynkin diagram. Hence, $SU(3)$ is a maximal subgroup of G_2.

Next consider E_6. If we consider the extended diagram, then if we remove the middle simple root, we are left with three $SU(3)$ diagrams. Hence E_6 has an $SU(3) \times SU(3) \times SU(3)$ maximal subgroup. But it also has another maximal subgroup. If we remove one of the simple roots in the middle of one of the legs, then we are left with an $SU(6) \times SU(2)$ diagram. If we remove one of the outer roots, then we get back the E_6 diagram again.

Next consider E_7. If we remove the simple root second from the left we get $SU(2) \times SO(10)$. (We can also get $SU(3) \times SU(6)$, $SU(3) \times SU(3) \times SU(2)$ and $SU(8)$.)

The group E_8 also has several maximal subgroups. Among others, it has an $SU(2) \times E_7$ subgroup, an $SU(3) \times E_6$ subgroup as well as an $SO(16)$ subgroup.

It is also of interest to determine how the representations transform under the subgroup. One thing that we should definitely expect is that a representation which is irreducible under the original group will be reducible under the subgroup, since the group is now smaller. With the smaller group, there are less transformations available and so it becomes likely that not every element in the representation can be transformed into every other element in the representation.

How the representations break up into the irreducible representations of the subgroup is of intense interest in studying grand unified theories. There is a precise way of determing what these irreducible representations are, but often we can figure out what they are by simply trying to fit the representations of the subgroup into the representations of the larger group.

As an example, consider the fundamental representation of G_2, which we write as **7**. We would like to determine how this representation reduces under the $SU(3)$ subgroup. The most obvious reason why it must reduce is that $SU(3)$ has no 7 dimensional irreducible representation! Since the **7** is a real representation, then under the $SU(3)$ decomposition, if a complex representation appears, then its conjugate representation must also appear. The first few irreducible representations of $SU(3)$ are the singlet **1** (that is the representation that is invariant under all $SU(3)$ transformations), the **3**, the $\overline{\mathbf{3}}$ and

the adjoint **8**. Given this, we see that under the subgroup, the **7** decomposes to

$$\mathbf{7} = \mathbf{1} + \mathbf{3} + \overline{\mathbf{3}}. \tag{2.15.1}$$

2.16 Real, pseudoreal and complex representations

Given the Lie Algebra in (2.6.4), it is simple to show that

$$-T_a^* \tag{2.16.1}$$

satisfies the same algebra, since the structure constants are real. Therefore, if the T_a are a representation of the algebra, then so are the $-T_a^*$. Clearly, the dimensions of the two representations are the same, since the complex conjugate of an $n \times n$ matrix is also an $n \times n$ matrix. But are they the same representation?

We can investigate this question as follows. Since the generators of the Cartan subalgebra H_i are hermitian, it follows that

$$(-H_i)^* = -H_i. \tag{2.16.2}$$

So one of these generators acting on a state of a given weight gives

$$(-H_i)^* |\vec{\mu}\rangle = -\mu_i |\vec{\mu}\rangle. \tag{2.16.3}$$

Hence, if $|\vec{\mu}\rangle$ and $|-\vec{\mu}\rangle$ are both in the representation, then we do not get a new representation by taking the transformation in (2.16.1). Representations which contain a state $|\vec{\mu}\rangle$ but not $|-\vec{\mu}\rangle$ are called *complex*. The representation containing $|-\vec{\mu}\rangle$ is called the *conjugate* of the representation containing $|\vec{\mu}\rangle$. Representations containing both states are called *real* representations.

Let us suppose that we have a real representation containing a state $|\vec{\mu}\rangle$. Since $|-\vec{\mu}\rangle$ is also in the representation, there must exist an invertible linear transformation where

$$R|\vec{\mu}\rangle = |-\vec{\mu}\rangle. \tag{2.16.4}$$

for all weight states in the representation. Thus we find

$$H_i R|\vec{\mu}\rangle = -\mu_i R|\vec{\mu}\rangle \quad \Rightarrow R^{-1} H_i R|\vec{\mu}\rangle = -\mu_i |\vec{\mu}\rangle, \tag{2.16.5}$$

and so

$$R^{-1} H_i R = -H_i^*. \tag{2.16.6}$$

Since the weights are in a representation of the Lie algebra, it must be true for all generators that

$$R^{-1}T_aR = -T_a^* = -T_a^T \tag{2.16.7}$$

where the last equality arises from the hermiticity of the generators.

But we also have that

$$T_a = -(R^{-1}T_aR)^T = -R^TT_a^T(R^{-1})^T \tag{2.16.8}$$

and so

$$R^{-1}R^TT_a^T(R^{-1}R^T)^{-1} = T_a^T. \tag{2.16.9}$$

This last equation means that

$$[R^{-1}R^T, T_a^T] = 0 \tag{2.16.10}$$

for all generators. The only matrix that can commute with all generators in an irreducible representation is one that is proportional to the identity. To see this, suppose that $R^{-1}R^T$ is not proportional to the identity. Then its eigenvalues cannot all be equal. But since it commutes with all T_a, it means that we can block diagonalize the representation such that those states on the upper block have one eigenvalue of $R^{-1}R^T$ while those on the lower block have another eigenvalue. But since $R^{-1}R^T$ commutes with T_a, the T_a cannot mix the states in the different blocks. Hence the representation of the Lie algebra can be block diagonalized, and hence it is reducible. Hence we conclude that

$$R^{-1}R^T = \lambda I \tag{2.16.11}$$

where I is the identity matrix for the representation. Therefore we find that

$$R^T = \lambda R. \tag{2.16.12}$$

Since $(R^T)^T = R$, we see that $\lambda^2 = 1$. Thus, the possible eigenvalues for λ are

$$\lambda = \pm 1. \tag{2.16.13}$$

Representations where $\lambda = +1$ are called *positive real* or sometimes just *real*, while representations where $\lambda = -1$ are called *pseudoreal*.

A positive real representation is one where the hermitian generators are completely imaginary. Therefore, a group element $g = \exp i\theta_aT_a$ is made up of only real matrices. If it is possible to express the generators as entirely imaginary, then there exists a unitary transformation

$$\widetilde{T}_a = UT_aU^{-1}, \tag{2.16.14}$$

where $\widetilde{T}_a = -\widetilde{T}_a^* = -\widetilde{T}^T$. Taking the transpose of (2.16.14) we have

$$\widetilde{T}_a^T = (U^{-1})^T T_a^T U^T = -U T_a U^{-1}. \tag{2.16.15}$$

This then gives

$$T_a = -U^{-1}(U^{-1})^T T_a^T U^T U = (U^T U)^{-1} T_a^T U^T U, \tag{2.16.16}$$

which means that

$$R = U^T U. \tag{2.16.17}$$

But then,

$$R^T = (U^T U)^T = U^T U = R. \tag{2.16.18}$$

Hence, the only representations that can have all real group elements are the positive reals.

2.17 Group invariants

It is often useful to find a set of invariants under the group action, that is the action of the group on its representations. For example, we recall from the study of angular momentum in quantum mechanics that there was an invariant that commuted with all components of the angular momentum operator, namely $\vec{J}^2$. We would like to find such invariants for other groups as well.

To do this, recall that we defined the inner product in the adjoint representation to be the trace of two generators

$$\langle T_a | T_b \rangle = \mathrm{Tr}(T_a^\dagger T_b) \tag{2.17.1}$$

We can always choose an orthornormal basis such that the generators are hermitian and orthonormal, in other words

$$\mathrm{Tr}(T_a T_b) = k\delta_{ab}, \tag{2.17.2}$$

where k is the same constant as in (2.7.12). For example, we can choose this basis by combining the root generators into the two separate terms

$$T_{\vec{a},1} = \frac{1}{\sqrt{2}}[G_{\vec{a}} + G_{-\vec{a}}] \quad \text{and} \qquad T_{\vec{a},2} = \frac{i}{\sqrt{2}}[G_{\vec{a}} - G_{-\vec{a}}]. \tag{2.17.3}$$

Both of these are hermitian and by using eqs. (2.7.21) and (2.7.22) one can easily show that they satisfy

$$\mathrm{Tr}(T_{\vec{a},i} T_{\vec{a},j}) = \delta_{ij} k, \tag{2.17.4}$$

The advantage of this orthornormal basis is that the structure constants have a nice form. Consider the trace over orthornormal generators

$$\mathrm{Tr}(T_a[T_b, T_c]) = \mathrm{Tr}(T_a f_{bcd} T_d) = k f_{bca} = -k f_{cba}. \tag{2.17.5}$$

But using the cyclic properties of the trace, we can also show that

$$\mathrm{Tr}(T_a[T_b, T_c]) = \mathrm{Tr}([T_a, T_b]T_c) = k f_{abc} = -k f_{bac} = \mathrm{Tr}([T_c, T_a]T_b) = k f_{cab} = -k f_{acb}. \tag{2.17.6}$$

In other words, in this basis the structure constants are completely antisymmetric!

Now consider the sum over all generators in this basis

$$I = \sum_a T_a T_a. \tag{2.17.7}$$

This sum is an invariant. To show this, consider the commutator of any generator with the sum (with repeated indices implying the sum)

$$[T_b, I] = [T_b, T_a]T_a + T_a[T_b, T_a] = f_{bad} T_d T_a + f_{bad} T_a T_d = -f_{bda} T_d T_a + f_{bad} T_a T_d = 0, \tag{2.17.8}$$

where we explicitly used the property that $f_{abc} = -f_{acb}$. Hence this sum is an invariant. In the literature this is known as the *quadratic Casimir*

Just as in the case of $SU(2)$, the quadratic Casimir will have different values for different representations. A useful relation can be found between the quadratic casimir and the trace of the square of any generator. These traces are repesentation dependent, so let us explicitly put a subscript on the trace

$$\mathrm{Tr}(T_a T_b) = k_R \delta_{ab} \tag{2.17.9}$$

where R refers to the representation. If we trace the casimir, we find

$$\mathrm{Tr}(I) = d_R I_R \tag{2.17.10}$$

where d_R is the dimension of the representation and I_R is the eigenvalue of the casimir for this representation. But we also have that

$$\mathrm{Tr}(T_a T_a) = \sum_a k_R = D k_R, \tag{2.17.11}$$

where D is the dimension of the algebra. Hence we see that

$$I_R = \frac{D}{d_R} k_R. \tag{2.17.12}$$

One consequence of (2.17.12) is that $I_R = k_R$ if R is the adjoint representation.

Up to now we have been carrying along this factor of k_R that comes from the normalization of the trace without giving a specific value to it. There is actually some freedom in assigning it a value. Notice in (2.6.4) that we could multiply all T_a by a factor of λ and still satisfy the Lie algebra, so long as we also rescale the structure constants by the same factor. Rescaling T_a would then rescale k_R by a factor of λ^2. However, once we fix k_R for one representation of the algebra, then we fix k_R for all representations since the structure constants are the same for all representations.

The standard choice for normalizing k_R is to use

$$\mathrm{Tr}(T_a T_b) = \frac{1}{2}\delta_{ab} \tag{2.17.13}$$

for the fundamental representation of $SU(2)$. Hence, we find that

$$I_f = \frac{D}{d_f}k_f = \frac{3}{2}\frac{1}{2} = \frac{3}{4} = \frac{1}{2}\left(\frac{1}{2}+1\right) \tag{2.17.14}$$

which is the standard result we know from the quantization of angular momentum. With this definition for $SU(2)$, there is a natural way to define k_f for the smallest fundamental representation of $SU(n)$. This representation is n dimensional. $SU(n)$ has an $SU(2)$ subgroup, which means that we can find 3 generators of the $SU(n)$ Lie algebra that form an $SU(2)$ Lie algebra amongst themselves. We can choose these generators to be

$$T_1 = \frac{1}{2}\begin{pmatrix} 0 & 1 & 0 & ... & 0 \\ 1 & 0 & 0 & ... & 0 \\ 0 & 0 & 0 & ... & 0 \\ & .. & ... & .. & \\ 0 & 0 & 0 & ... & 0 \end{pmatrix} \quad T_2 = \frac{1}{2}\begin{pmatrix} 0 & -i & 0 & ... & 0 \\ i & 0 & 0 & ... & 0 \\ 0 & 0 & 0 & ... & 0 \\ & .. & ... & .. & \\ 0 & 0 & 0 & ... & 0 \end{pmatrix} \quad T_3 = \frac{1}{2}\begin{pmatrix} 1 & 0 & 0 & ... & 0 \\ 0 & -1 & 0 & ... & 0 \\ 0 & 0 & 0 & ... & 0 \\ & .. & ... & .. & \\ 0 & 0 & 0 & ... & 0 \end{pmatrix} \tag{2.17.15}$$

Hence we see that we have Pauli matrices in the upper left two by two block and zeroes everywhere else. The group elements coming from these generators then look like

$$U = \exp(i\sum_{a=1}^{3}\theta_a T_a) = \begin{pmatrix} U_{11} & U_{12} & 0 & 0 & ... & 0 \\ U_{21} & U_{12} & 0 & 0 & ... & 0 \\ 0 & 0 & 1 & 0 & ... & 0 \\ 0 & 0 & 0 & 1 & ... & 0 \\ & & .. & .. & .. & .. \\ 0 & 0 & 0 & 0 & ... & 1 \end{pmatrix} \tag{2.17.16}$$

where the two by two matrix in the upper left block is a two dimensional unitary matrix with determinant one and thus is a fundamental representation of $SU(2)$. The 1's along

the diagonal are also representations of $SU(2)$, but they are trivial representations. Hence, we see that the n dimensional irreducible representation of $SU(n)$ is reducible under the $SU(2)$ subgroup to one fundamental representation and $n-2$ trivial representations. These trivial representations are sometimes called *singlets*.

Anyway, now we see that for the three generators of the subgroup, the trace is

$$\mathrm{Tr}(T_a T_b) = \frac{1}{2}\delta_{ab}. \tag{2.17.17}$$

But (2.17.9) holds for all generators of the algebra, so we must get (2.17.17) for all generators of the $SU(n)$ Lie algebra, and so $k_f = 1/2$.

The quadratic casimir is the only invariant for $SU(2)$ (this should come as no surprise since we already know that $\vec{J}^2$ is the only invariant of a representation). However, the other groups have other invariants.

2.18 The Center of the Group

The *center* of a group is made up of all group elements that commute with every other element in the group. The center is clearly a group. For $SU(n)$, the center is Z_n, as should be clear, since the elements

$$\mathrm{diag}(e^{2\pi i m/n}, e^{2\pi i m/n}, e^{2\pi i m/n} ...., e^{2\pi i m/n}) \tag{2.18.1}$$

are the identity multiplied by a constant and the determinant is 1. Of special interest is the coset of the group by its center. The coset itself is still a group. The identification that takes place is between group elements that are not arbitrarily close to each other. This means for group elements constrained to be close to the identity, the group looks the same. What we say is that the local structure of the group is the same, but its global structure has changed.

Since the local structure is the same, the Lie algebra is the same. However, modding out a group by its center could remove some representations. If a representation is not invariant under the center, then it can't be a representation of the coset. One representation that clearly survives is the adjoint representation, since the action by a group element is via a commutator, so an element of the adjoint representation must be invariant under the center since the center commutes with everything.

As an example, let us consider $SU(2)$. This has a Z_2 center. The integer spin representations are invariant under the center, but the half integer representations are not; they change sign when acted upon by the nontrivial element. Hence the group $SU(2)/Z_2$ only

has the integer spin representations, which obviously includes the adjoint. Now consider $SO(3)$, which has the same Dynkin diagram as $SU(2)$, so locally these groups are the same. But $SO(3)$ has no center. (Note that $\mathrm{diag}(-1,-1,-1)$ is not in $SO(3)$ since it has determinant -1. This is in the group $O(3)$.) In fact $SO(3)$ is the same as $SU(2)/Z_2$, which obviously has no center since it was taken out. $SO(3)$ is the group corresponding to the orbital angular momentum, which of course only has integer representations. We also see that the group manifold for $SO(3)$ is S_3/Z_2. This will have interesting consequences.

3 Path Integrals, Wilson Lines and Gauge Fields

We have spent the first few lectures learning about Lie Groups. Lie Groups are of particular importance for the study of gauge theories. However, before we can put this to good use, we need to study something called a *path integral*.

3.1 Path integrals for nonrelativistic particles

In quantum mechanics we often want to compute the probability of finding particles with, say, particular positions and momentums. One particular problem is this: given that a particle is at position $\vec{x}_0$ at time t_0, what is the probability that the particle is at position $\vec{x}_1$ at time t_1? To find the probability, we need to solve the time dependent Schroedinger equation

$$i\hbar\frac{\partial}{\partial t}\Psi(\vec{x},t) = H\,\Psi(\vec{x},t) \tag{3.1.1}$$

where H is the Hamiltonian

$$H = \frac{\vec{p}^2}{2m} + V(\vec{x}). \tag{3.1.2}$$

At time t_0, the wave function is 0 unless $\vec{x} = \vec{x}_0$. (We say that the wave function has *δ function support*). Then, as time elapses, the wave function evolves. At time $t = t_1$, we then compute the probability amplitude for the wavefunction at $\vec{x} = \vec{x}_1$. Formally, the solution to the Schroedinger equation is

$$\Psi(\vec{x},t) = e^{-\frac{i}{\hbar}H(t-t_0)}\Psi(\vec{x},t_0). \tag{3.1.3}$$

When we first learn quantum mechanics, we almost always assume that Ψ is an eigenfunction of the Hamiltonian. However, a wave function with δ function support is certainly not an eigenfunction of the Hamiltonian, since the position operator does not commute with the momentum operator.

Thus to treat this problem, we do the following. Let us use bra-ket notation. Consider the ket state $|\vec{x}_0, t_0\rangle$, where we have explicitly inserted a time variable to indicate that this is the state at time $t = t_0$. The probability amplitude is then found by computing the inner product

$$\langle \vec{x}_1, t_1 | \vec{x}_0, t_0\rangle. \tag{3.1.4}$$

This is a formal way of saying that in order to compute the inner product, we have to let the state evolve over time. But this we can find from (3.1.3), and the inner product in (3.1.4) can be explicitly written as

$$\langle \vec{x}_1, t_1 | \vec{x}_0, t_0\rangle = \langle \vec{x}_1 | e^{-\frac{i}{\hbar}H(t_1-t_0)} | \vec{x}_0\rangle. \tag{3.1.5}$$

To evaluate the expression in (3.1.5), we split up the time interval between t_0 and t_1 into a large number of infinitesimally small time intervals Δt , so that

$$e^{-\frac{i}{\hbar}H(t_1-t_0)} = \prod e^{-\frac{i}{\hbar}H\Delta t}, \tag{3.1.6}$$

where the product is over all time intervals between t_0 and t_1. Since, $\vec{p}$ does not commute with $\vec{x}$, we see that

$$e^{-\frac{i}{\hbar}H\Delta t} \neq e^{-\frac{i}{2m\hbar}\vec{p}^2\Delta t}e^{-\frac{i}{\hbar}V(\vec{x})\Delta t}, \tag{3.1.7}$$

however, if Δt is very small, then it is approximately true, in that

$$e^{-\frac{i}{\hbar}H\Delta t} = e^{-\frac{i}{2m\hbar}\vec{p}^2\Delta t}e^{-\frac{i}{\hbar}V(\vec{x})\Delta t} + \mathrm{O}((\Delta \mathrm{t})^2). \tag{3.1.8}$$

Hence, in the limit that $\Delta t \to 0$, we have that

$$\langle \vec{x}_1, t_1 | \vec{x}_0, t_0 \rangle = \lim_{\Delta t \to 0} \langle \vec{x}_1 | \prod e^{-\frac{i}{2m\hbar}\vec{p}^2\Delta t}e^{-\frac{i}{\hbar}V(\vec{x})\Delta t} | \vec{x}_0 \rangle. \tag{3.1.9}$$

In order to evaluate the expression in (3.1.9), we need to insert a complete set of states between each term in the product, the states being either position or momentum eigenstates and normalized so that

$$\begin{aligned} &\int d^3x |\vec{x}\rangle\langle\vec{x}| = 1 \qquad\qquad \langle\vec{x}_1|\vec{x}_0\rangle = \delta^3(\vec{x}_1 - \vec{x}_0) \\ &\int \frac{d^3p}{(2\pi\hbar)^3} |\vec{p}\rangle\langle\vec{p}| = 1 \qquad\qquad \langle\vec{p}_1|\vec{p}_0\rangle = (2\pi\hbar)^3\delta^3(\vec{p}_1 - \vec{p}_0) \\ &\langle\vec{x}|\vec{p}\rangle = e^{i\vec{x}\cdot\vec{p}/\hbar} \end{aligned} \tag{3.1.10}$$

Inserting the position states first, we have that (3.1.9) is

$$\langle \vec{x}_1, t_1 | \vec{x}_0, t_0 \rangle = \int \prod_{t_0 < t < t_1} d^3x(t) \left[\prod_{t_0 \leq t < t_1} \langle \vec{x}(t+\Delta t) | e^{-\frac{i}{2m\hbar}|\vec{p}|^2\Delta t}e^{-\frac{i}{\hbar}V(\vec{x})\Delta t} | \vec{x}(t) \rangle \right], \tag{3.1.11}$$

where the first product is over each time between $t_0 + \Delta t$ and $t_1 - \Delta t$ and the second product is over each time between t_0 and $t_1 - \Delta t$. We also have that $\vec{x}(t_0) = \vec{x}_0$ and $\vec{x}(t_1) = \vec{x}_1$. Note that for each time interval, we have a position variable that we integrate over. You should think of the time variable in these integrals as a label for the different x variables.

We now need to insert a complete set of momentum states at each time t. In particular, we have that

$$\begin{aligned}
&\langle \vec{x}(t+\Delta t)|e^{-\frac{i}{2m\hbar}|\vec{p}|^2\Delta t}e^{-\frac{i}{\hbar}V(\vec{x})\Delta t}|\vec{x}(t)\rangle \\
&= \int \frac{d^3p}{(2\pi\hbar)^3}\langle \vec{x}(t+\Delta t)|\vec{p}\rangle e^{-\frac{i}{2m\hbar}|\vec{p}|^2\Delta t}\langle \vec{p}|\vec{x}(t)\rangle e^{-\frac{i}{\hbar}V(\vec{x}(t))\Delta t} \\
&= \int \frac{d^3p}{(2\pi\hbar)^3}e^{i\Delta\vec{x}(t)\cdot\vec{p}/\hbar}e^{-\frac{i}{2m\hbar}|\vec{p}|^2\Delta t}e^{-\frac{i}{\hbar}V(\vec{x}(t))\Delta t} \qquad (3.1.12)
\end{aligned}$$

where $\Delta x(t) = x(t+\Delta t) - x(t)$. We can now do the gaussian integral in the last line in (3.1.12), which after completing the square gives

$$\langle \vec{x}(t+\Delta t)|e^{-\frac{i}{2m\hbar}|\vec{p}|^2\Delta t}e^{-\frac{i}{\hbar}V(\vec{x})\Delta t}|\vec{x}(t)\rangle = \left(\frac{m}{2\pi i\hbar\Delta t}\right)^{3/2}\exp\left(\frac{i}{\hbar}\left[\frac{m}{2}\left(\frac{\Delta\vec{x}}{\Delta t}\right)^2 - V(\vec{x}(t))\right]\Delta t\right). \qquad (3.1.13)$$

Strictly speaking, the integral in the last line of (3.1.12) is not a Gaussian, since the coefficient in front of $\vec{p}^2$ is imaginary. However, we can regulate this by assuming that the coefficient has a small negative real part and then let this part go to zero after doing the integral. The last term in (3.1.13) can be written as

$$= \left(\frac{m}{2\pi i\hbar\Delta t}\right)^{3/2}\exp\left(\frac{i}{\hbar}\left[\frac{m}{2}\dot{\vec{x}}^2 - V(\vec{x}(t))\right]\Delta t\right) = \left(\frac{m}{2\pi i\hbar\Delta t}\right)^{3/2}\exp\left(\frac{i}{\hbar}\mathcal{L}(t)\Delta t\right), \qquad (3.1.14)$$

where $\mathcal{L}(t)$ is the lagrangian of the particle evaluated at time t. Hence the complete path integral can be written as

$$\langle \vec{x}_1, t_1|\vec{x}_0, t_0\rangle = \left(\frac{m}{2\pi i\hbar\Delta t}\right)^{3N/2}\int\prod_{t_0<t<t_1} d^3x(t)\exp\left(\frac{i}{\hbar}S\right), \qquad (3.1.15)$$

where S is the *action* and is given by

$$S = \int_{t_0}^{t_1}\mathcal{L}(t)dt. \qquad (3.1.16)$$

N counts the number of time intervals in the path. In the limit that N goes to infinity, we see that the constant in front of the expression diverges. It is standard practice to drop the constant, which is essentially a normalization constant. It can always be brought back later after normalization.

The expression in (3.1.15) was first derived by Feynman and it gives a very intuitive way of looking at quantum mechanics. What the expression is telling us is that to compute the probability amplitude, we need to sum over all possible paths that the particle can take in getting from x_0 to x_1, weighted by $e^{\frac{i}{\hbar}S}$.

It is natural to ask which path dominates the path integral. Since the argument of the exponential is purely imaginary, we see that the path integral is a sum over phases. In general, when integrating over the $x(t)$, the phase varies and the phases coming from the different paths tend to cancel each other out. What is needed is a path where varying to a nearby path gives no phase change. Then the phases add constructively and we are left with a large contribution to the path integral from the path and its nearby neighbors.

Hence, we look for the path, given by a parameterization $x(t)$, such that $x(t_0) = x_0$ and $x(t_1) = x_1$, and such that the nearby paths have the same phase, or close to the same phase. This means that if $x(t)$ is shifted to $x(t) + \delta x(t)$, then the change to the action is very small. To find this path, note that under the shift, to lowest order in δx, the action changes to

$$S \to S + \int_{t_0}^{t_1} dt \left[\frac{\partial \mathcal{L}}{\partial \dot{\vec{x}}} \delta \dot{\vec{x}} + \frac{\partial \mathcal{L}}{\partial \vec{x}} \delta \vec{x} \right] = S + \int_{t_0}^{t_1} dt \left[-\frac{d}{dt} \left(\frac{\partial \mathcal{L}}{\partial \dot{\vec{x}}} \right) + \frac{\partial \mathcal{L}}{\partial \vec{x}} \right] \delta \vec{x}. \tag{3.1.17}$$

Hence there would be no phase change to lowest order in δx if the term inside the square brackets is zero. But this is just the classical equation of motion! A generic path has a phase change of order δx, but the classical path has a phase change of order δx^2.

Next consider what happens as $\hbar \to 0$. Then a small change in the action can lead to a big change in the phase. In fact, even a very small change in the action essentially wipes out any contribution to the path integral. In this case, the classical path is essentially the *only* contribution to the path integral. For nonzero $\hbar$, while the classical path is the dominant contributor to the path integral, the nonclassical paths also contribute, since the phase is finite.

3.2 Path integrals for charged particles

Now suppose that we have a charged particle in an electromagnetic field. Recall that we can write the electric field as

$$\vec{E}(x) = -\nabla \phi(\vec{x}, t) + \frac{d}{dt} \vec{A}(\vec{x}, t) \qquad \vec{B}(\vec{x}, t) = -\vec{\nabla} \times \vec{A}(\vec{x}, t) \tag{3.2.1}$$

We can write these fields as a 4-tensor

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu, \qquad A_0 = \phi. \tag{3.2.2}$$

with $F_{0i} = E_i$ and $F_{ij} = -\epsilon_{ijk} B_k$. A charged particle in such a field feels the Lorentz force

$$F = e\vec{E} + e\vec{v} \times \vec{B}. \tag{3.2.3}$$

To get such a term in the equation of motion for the particle, we can add to the action the term

$$S_{em} = -e \int_{t_0}^{t_1} dt \left[\phi + \dot{\vec{x}} \cdot \vec{A} \right], \tag{3.2.4}$$

so that varying with respect to $\vec{x}(t)$ gives

$$S_{em} \to S_{em} - e \int_{t_0}^{t_1} dt \left[\vec{\nabla}\phi - \dot{\vec{A}} + \dot{x}_i \vec{\nabla} A_i - \dot{x}_i \partial_i \vec{A} \right] \cdot \delta\vec{x}(t), \tag{3.2.5}$$

which is the desired result to get the correct force. By the chain rule, we can rewrite the action in (3.2.4) in terms of the line integral

$$S_{em} = -e \int_{t_0}^{t_1} A_\mu dx^\mu, \tag{3.2.6}$$

which is explicitly Lorentz covariant. Therefore, in the presence of the electromagnetic field, our path integral for a charged particle is modified to

$$\langle \vec{x}_1, t_1 | \vec{x}_0, t_0 \rangle = \int \prod_{t_0 < t < t_1} d^3x(t) \exp\left(\frac{i}{\hbar} S - \frac{ie}{\hbar} \int_{t_0}^{t_1} A_\mu dx^\mu \right), \tag{3.2.7}$$

The gauge field A_μ has a redundancy. Notice that if we transform A_μ to

$$A_\mu \to A_\mu + \partial_\mu \phi \tag{3.2.8}$$

where ϕ is any function, then the electromagnetic fields are unchanged. Notice that this transformation is local – ϕ can take on different values at different space-time points. Since the fields are unchanged, it must be true that the physics for the charged particle does not change under the gauge transformation. If we look at the action in (3.2.6), we see that under the gauge transformation the change in the action is

$$\delta S_{em} = -e \int_{t_0}^{t_1} \partial_\mu \phi dx^\mu = -e(\phi(\vec{x}_1, t_1) - \phi(\vec{x}_0, t_0)). \tag{3.2.9}$$

Hence the path integral in (3.2.7) transforms to

$$\langle \vec{x}_1, t_1 | \vec{x}_0, t_0 \rangle \to e^{-ie\phi(\vec{x}_1, t_1)/\hbar} \langle \vec{x}_1, t_1 | \vec{x}_0, t_0 \rangle e^{+ie\phi(\vec{x}_0, t_0)/\hbar}. \tag{3.2.10}$$

In other words, the path integral only changes by a phase and so the physics does not change, since the amplitude squared is unchanged.

3.3 Wilson lines

Not only is a charged particle acted on by external electromagnetic fields, it is also a source for those same fields. If we assume that the charge particle is very heavy, then it is not influenced by the electromagnetic fields and in fact can be restricted to a specific path. So we could instead consider

$$W(x_0, x_1) = \exp\left(-\frac{ie}{\hbar}\int_{x_0}^{x_1} A_\mu dx^\mu\right), \tag{3.3.1}$$

over a fixed path from x_0 to x_1. In quantum field theory, we will have path integrals where we integrate over the fields A_μ, but let us not worry about that. The expression in (3.3.1) is called a *Wilson line*. As we already saw, the Wilson line is not exactly gauge invariant, it instead changes to

$$W \to e^{-ie\phi(\vec{x}_1,t_1)/\hbar} W e^{+ie\phi(\vec{x}_0,t_0)/\hbar} \tag{3.3.2}$$

An object which transforms like this is said to be *gauge covariant*. Now we notice something about how W transforms. Its endpoints are transformed under two distinct $U(1)$ transformations. That is, a gauge transformation transforms the endpoints of the Wilson line by $U(1)$ group elements. Since W transforms this way, we say that electromagnetism is a $U(1)$ gauge theory. The fact that the $U(1)$ transformations are different at the different endpoints reflects the fact that the gauge transformations are local. We also notice that shifting ϕ by $2\pi\hbar/e$ leaves the endpoints invariant. Hence the gauge transformations are such that ϕ is identified with $\phi + 2\pi\hbar/e$. Because of the explicit $\hbar$ dependence, we see that this identification is a quantum effect.

We can also find an object that is actually gauge invariant, namely the *Wilson loop*, which has the form

$$W = \exp\left(-\frac{ie}{\hbar}\oint A_\mu dx^\mu\right). \tag{3.3.3}$$

Now there is no endpoint, so W must be invariant. Another way to think of this is that the endpoints in (3.3.1) are the same, so the two $U(1)$ factors cancel off with each other.

There is also a discrete way of looking at a Wilson line, which will be useful for us when we consider generalizations to other gauge groups. Suppose that we consider a very short and straight Wilson line that traverses from point x to $x + \Delta x$. The Wilson line may then be approximated by

$$W(x, x + \Delta x) = \exp\left(-ieA_\mu \Delta x^\mu/\hbar\right). \tag{3.3.4}$$

$W(x, x+\Delta x)$ is called the link variable, or just link, between x and $x+\Delta x$. It is clearly a unitary matrix. Under gauge transformations, the link transforms as

$$U(x)W(x, x+\Delta x)U^\dagger(x+\Delta x). \tag{3.3.5}$$

A Wilson line is then given as a product of connected links. The important point here is that we can basically forget the gauge field A_μ and just worry about the link W.

Now we can see how to generalize a Wilson line that transforms under a $U(1)$ gauge transformation to one that transforms under some other Lie group. Let us suppose that we have link variables that take their values in a representation for some Lie group. Let us now consider a product of attached links

$$W(x_0, x_1) = W(x_0, x_0+\Delta x_1)(W(x_0+\Delta x_1, x_0+\Delta x_1+\Delta x_2)..W(x_1-\Delta x_N, x_1). \tag{3.3.6}$$

Under gauge transformations the links transform as

$$W(x, x+\Delta x) \to U(x)W(x, x+\Delta x)U^\dagger(x+\Delta x) \tag{3.3.7}$$

hence the Wilson line transforms as

$$W(x_0, x_1) = U(x)W(x_0, x_1)U^\dagger(x_1) \tag{3.3.8}$$

For a closed Wilson loop, we would find that

$$W(x, x) \to U(x)W(x, x)U^\dagger(x), \tag{3.3.9}$$

hence the Wilson loop for a nonabelian gauge group is not actually gauge invariant, but only gauge covariant. A gauge invariant object would be

$$\mathrm{Tr}(W(x, x)) \to \mathrm{Tr}\left(U(x)W(x, x)U^\dagger(x)\right) = \mathrm{Tr}(W(x, x)). \tag{3.3.10}$$

3.4 The magnetic monopole

As an application let us consider a magnetic monopole. This section will serve as a prelude to our discussion on fiber bundles later in the course.

A monopole will have a $\vec{B}$ field pointing in the radial direction. Let us suppose that the monopole charge is g, then the $\vec{B}$ field is given by

$$\vec{B} = \frac{g}{4\pi r^2}\hat{r} \tag{3.4.1}$$

which insures that the flux through a sphere surrounding the monopole is g (I am choosing units where μ, ϵ and c are all 1.)

$$g = \int_{S_2} d\vec{S} \cdot \vec{B} = \int \sin\theta d\theta d\phi r^2 B_r, \tag{3.4.2}$$

where $d\vec{S}$ is the differential on the surface of S_2.

However, there is a problem. Recall that $\vec{B}$ was given as $-\vec{\nabla} \times \vec{A}$. That means that the first integral in (3.4.2) is given by

$$\int_{S_2} d\vec{S} \cdot \vec{B} = -\int_{S_2} d\vec{S} \cdot \vec{\nabla} \times \vec{A}. \tag{3.4.3}$$

Using Stoke's theorem, which says that

$$\int_{\Sigma} d\vec{S} \cdot \left(\vec{\nabla} \times \vec{A}\right) = \int_{\partial\Sigma} d\vec{x} \cdot \vec{A}, \tag{3.4.4}$$

where Σ is a 2 dimensional surface and $\partial\Sigma$ is its boundary, we see that

$$\int_{S_2} d\vec{S} \cdot \vec{B} = 0, \tag{3.4.5}$$

since S_2 has no boundary.

What went wrong? Our mistake was assuming that $\vec{A}$ is globally defined over the entire S_2. Instead, let us break up the sphere into two hemispheres, with $\vec{A}_N$ on the northern hemisphere which includes the equator and $\vec{A}_S$ on the southern hemisphere which also includes the equator. Hence, on the equator the gauge fields $\vec{A}_N$ and $\vec{A}_S$ overlap. Naively, we would say that the two fields have to equal each other on the equator, but this is too stringent a condition. Instead, we only require that they are equal up to a gauge transformation. Therefore, let us assume that $\vec{A}_N = \vec{A}_S + \vec{\nabla}\phi$. Now, Stokes law tells us that

$$\int_{S_2} d\vec{S} \cdot \vec{B} = -\oint d\vec{x} \cdot \vec{\nabla}\phi, \tag{3.4.6}$$

where the line integral is around the equator. Now normally, we should expect the integral to still be zero, since the rhs in (3.4.6) is a total derivative around a loop. But remember from the previous section that ϕ is identified with $\phi + 2\pi\hbar/e$, so ϕ only needs to come back to itself up to a multiple of $2\pi\hbar/e$ when going around the equator. Therefore we find that the magnetic charge can be nonzero and is given by

$$g = n\frac{2\pi\hbar}{e}, \tag{3.4.7}$$

where n is an integer. Hence we find that the magnetic charge is quantized!

3.5 Gauge fields for nonabelian Lie groups

We saw previously that the *field strength* for a $U(1)$ gauge group is given by

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu. \tag{3.5.1}$$

In this section we construct the corresponding object for a gauge field which has gauge transformations in a nonabelian gauge group.

Recall that the gauge transformation for A_μ is

$$A_\mu \to A_\mu + \partial_\mu \phi = A_\mu - i\frac{\hbar}{e} U^\dagger \partial_\mu U, \tag{3.5.2}$$

where we have written the gauge transformation in terms of the $U(1)$ group element $U = e^{ie\phi/\hbar}$. For the case of a nonabelian gauge group, let us consider the short Wilson line $W(x, x+\Delta x)$. This is some unitary matrix, so we can assume that it can be written as

$$W(x, x+\Delta x) = \exp\left(-i\frac{e}{\hbar} A_\mu^a T^a \Delta x^\mu\right) \tag{3.5.3}$$

where the T_a are the generators of the group. We will write

$$A_\mu = A_\mu^a T^a, \tag{3.5.4}$$

so now A_μ is a hermitian matrix.

To find the gauge transformation of A_μ, note that

$$\exp\left(-i\frac{e}{\hbar} A_\mu \Delta x^\mu\right) \to U(x) \exp\left(-i\frac{e}{\hbar} A_\mu \Delta x^\mu\right) U^\dagger(x+\Delta x) \tag{3.5.5}$$

Hence we have that

$$A_\mu \Delta x^\mu \to U(x) A_\mu U^\dagger(x+\Delta x)\Delta x^\mu + i\frac{\hbar}{e}\left(U(x)U^\dagger(x+\Delta x) - 1\right). \tag{3.5.6}$$

Taking the limit that $\Delta x \to 0$, we have

$$A_\mu \to U(x) A_\mu U^\dagger(x) - i\frac{\hbar}{e}\partial_\mu U(x) U^\dagger(x), \tag{3.5.7}$$

where we used the fact that

$$U^\dagger(x+\Delta x) = U^\dagger(x) + \Delta x^\mu \partial_\mu U^\dagger(x) \tag{3.5.8}$$

and $\partial U^\dagger = -U^\dagger \partial U U^\dagger$.

The field strength should at least be gauge covariant. The naive extension of the $U(1)$ case, $\partial_\mu A_\nu - \partial_\nu A_\mu$ does not satisfy this requirement. However, consider the quantity

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu - i\frac{e}{\hbar}[A_\mu, A_\nu] \tag{3.5.9}$$

Then under a gauge transformation, this becomes

$$\begin{aligned}
&U\left(\partial_\mu A_\nu - \partial_\nu A_\mu - i\frac{e}{\hbar}[A_\mu, A_\nu]\right)U^\dagger \\
&-i\frac{\hbar}{e}[\partial_\mu UU^\dagger, \partial_\nu UU^\dagger] + [\partial_\mu UU^\dagger, UA_\nu U^\dagger] - [\partial_\nu UU^\dagger, UA_\mu U^\dagger] \\
&+i\frac{\hbar}{e}[\partial_\mu UU^\dagger, \partial_\nu UU^\dagger] - [\partial_\mu UU^\dagger, UA_\nu U^\dagger] + [\partial_\nu UU^\dagger, UA_\mu U^\dagger] \\
&= U\left(\partial_\mu A_\nu - \partial_\nu A_m u - i\frac{e}{\hbar}[A_\mu, A_\nu]\right)U^\dagger
\end{aligned} \tag{3.5.10}$$

Therefore, $F_{\mu\nu}$ is the covariant field strength.

3.6 Gauge field actions

Maxwell's equations for the $U(1)$ gauge field with no sources (that is no charges or currents) can be written as

$$\partial_\mu F^{\mu\nu} = 0. \tag{3.6.1}$$

Notice that the indices in $F_{\mu\nu}$ have been raised. This is to guarantee that the expression is Lorentz covariant. (Recall from special relativity that an index is raised using the metric tensor, which in flat space is $\eta_{\mu\nu} = \mathrm{diag}(-1, 1, 1, 1)$.) The repeated indices have an implied sum. Actually, (3.6.1) gives only half the equations, namely

$$\begin{aligned}
\vec{\nabla}\cdot\vec{E} &= 0 \\
\partial_t\vec{E} - \vec{\nabla}\times\vec{B} &= 0.
\end{aligned} \tag{3.6.2}$$

The other two equations are guaranteed by the form of $F_{\mu\nu}$ in (3.5.1). Namely, notice that the equation

$$\partial_\lambda F_{\mu\nu} + \partial_\nu F_{\lambda\mu} + \partial_\mu F_{\nu\lambda} = 0 \tag{3.6.3}$$

is automatic given (3.5.1). The equation in (3.6.3) is an example of a *Bianchi identity*. Using (3.2.1), we can rewrite (3.6.3) as

$$\vec{\nabla}\times\vec{E} + \frac{\partial}{\partial t}\vec{B} = 0, \qquad \vec{\nabla}\cdot\vec{B} = 0. \tag{3.6.4}$$

The first equation is Faraday's Law and the second assumes that there are no magnetic monopoles.

We wish to find a lagrangian that leads to the equations of motion in (3.6.1). Actually, instead of a lagrangian, we want a lagrangian density, where the lagrangian L is given by

$$L = \int d^3x \mathcal{L}. \tag{3.6.5}$$

Let us show that

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} \tag{3.6.6}$$

does the job. Given $\mathcal{L}$, we see that the action is

$$S = -\frac{1}{4} \int d^4x F_{\mu\nu} F^{\mu\nu}, \tag{3.6.7}$$

where d^4x is the measure for the integral over the 4 dimensional space-time. Notice that since $F_{\mu\nu}$ is gauge invariant, the action is also gauge invariant.

To find the equations of motion, we should vary the fields in the action and by doing so, find a field configuration where the action is minimized. Put differently, we want to find a configuration of $A_\mu(x^\sigma)$ such that the action is unchanged to lowest order when the fields are changed by $\delta A_\mu(x^\sigma)$ for all points in space-time. Formally, we wish to take derivatives with respect to the $A_\mu(x^\sigma)$ fields, just as in section (3.1) we took derivatives with respect to $\vec{x}(t)$. For derivatives with respect to $x^i(t)$, we use the fact that

$$\frac{\partial x^i(t')}{\partial x^j(t)} = \delta(t - t')\delta_{ij}, \tag{3.6.8}$$

in other words, the variations at one time are independent of the variations at a different time. In the case of the fields, we have that the variations at a point in space-time are independent of the variations at a different point in space-time. Hence we have,

$$\frac{\partial A_\mu(x^\sigma)}{\partial A_\nu(y^\sigma)} = \delta^4(x^\sigma - y^\sigma)\delta^\nu{}_\mu. \tag{3.6.9}$$

Consider then the variation of the action in (3.6.7). The variation of $\mathcal{L}(y^\sigma)$ is

$$\frac{\partial \mathcal{L}(y^\sigma)}{\partial A_\nu(x^\sigma)} = 4\left(-\frac{1}{4}\right) F^{\mu\nu}(y^\sigma)\partial_\mu \delta^4(x^\sigma - y^\sigma), \tag{3.6.10}$$

and so the variation of the action is

$$\frac{\partial S}{\partial A_\mu(x^\sigma)} = -\int d^4y F^{\mu\nu}(y^\sigma)\partial_\mu \delta^4(x^\sigma - y^\sigma) = \partial_\nu F^{\mu\nu}(x^\sigma) = 0, \tag{3.6.11}$$

where we did an integration by parts to get the final form. This is the equation of motion for the gauge field.

We can see that the funny factor of $-1/4$ in (3.6.6) is chosen so that the end result is (3.5.1). The lagrangian can also be written as

$$\mathcal{L} = \frac{1}{2}\vec{E}\cdot\vec{E} - \frac{1}{2}\vec{B}\cdot\vec{B}, \tag{3.6.12}$$

where $\vec{E}$ serves as the canonical momentum for $\vec{A}$. The Hamiltonian is then

$$H = \frac{1}{2}\vec{E}\cdot\vec{E} + \frac{1}{2}\vec{B}\cdot\vec{B}, \tag{3.6.13}$$

which is the expected energy density for an electromagnetic field.

Now let us assume that there is also a charged particle around that can influence the electromagnetic fields. Hence we want to add to the action in (3.6.7) the action in (3.2.6). However, we have to be a little careful since the action in (3.6.7) has an integration over 4 dimensions but the integral in (3.2.6) is a one dimensional line integral. Anyway, varying the sum of the actions we find

$$\frac{\partial(S+S_{em})}{\partial A_\mu(x^\sigma)} = \partial_\nu F^{\mu\nu} - e\int dy^\nu \delta^4(x^\sigma - y^\sigma) = 0, \tag{3.6.14}$$

and so the modified equation of motion is

$$\partial_\nu F^{\mu\nu} = e\int dy^\nu \delta^4(x^\sigma - y^\sigma). \tag{3.6.15}$$

To understand this equation, let us look at the various components. For example, consider the $\nu = 0$ component equation

$$\partial_\nu F^{0\nu} = \vec{\nabla}\cdot E = e\delta^3(x^i(t) - y^i(t)). \tag{3.6.16}$$

$y^i(t)$ is the position of the particle at time t. Hence the term on the right hand side of this equation is the charge density, and so this is Gauss' law in the presence of a charged particle. If we look at one of the spatial components, then the equation is

$$\partial_\nu F^{i\nu} = \partial_t E^i - (\vec{\nabla}\times\vec{B})^i = e\frac{dy^i}{dt}\delta^3(x^i(t) - y^i(t)). \tag{3.6.17}$$

The right hand side is now the current density for the moving charged particle and so our modified equation of motion is Ampere's law in the presence of a current. From this discussion, we see that not only does S_{em} tell us how an electromagnetic field affects a charged particle, but it also tells us how a charged particle affects the field.

We now wish to generalize the lagrangian in (3.6.6) to nonabelian gauge fields. We saw in the previous section that an appropriate choice for the field strength is the expresssion in (3.5.9). It is natural to choose the lagrangian to be $-\frac{1}{4}F_{\mu\nu}F^{\mu\nu}$. There are two slight problems with this. First, the Lagrangian should be a number, but this is a matrix. Second, this expression is not gauge invariant, but only gauge covariant. Given the transformation properties in (3.5.10), we see that

$$F_{\mu\nu}F^{\mu\nu} \to UF_{\mu\nu}F^{\mu\nu}U^\dagger, \tag{3.6.18}$$

so this is only covariant. To find an *invariant* expression, we can just take the trace. This also gives a lagrangian that is a number and not a matrix. Hence, the Lagrangian is

$$\mathcal{L} = -\frac{1}{4k}\mathrm{Tr}[F_{\mu\nu}F^{\mu\nu}]. \tag{3.6.19}$$

We have the same factor of $-\frac{1}{4}$, so that the kinetic terms of the $U(1)$ subgroups, that is those groups generated by the Cartan subalgebra, have the same normalization as the $U(1)$ gauge fields of electromagnetism. The factor of $1/k$ in (3.6.19) cancels the factor of k from (2.7.12).

3.7 The covariant derivative

There is a convenient way to write the field strength that nicely generalizes to the non-abelian case. Define the *covariant derivative* D_μ to be

$$D_\mu = \partial_\mu - i\frac{e}{\hbar}A_\mu. \tag{3.7.1}$$

Next consider the commutator $[D_\mu, D_\nu]$, which is

$$[D_\mu, D_\nu] = -i\frac{e}{\hbar}\left(\partial_\mu A_\nu - \partial_\nu A_\mu - i\frac{e}{\hbar}[A_\mu, A_\nu]\right) = -i\frac{e}{\hbar}F_{\mu\nu} \tag{3.7.2}$$

Hence, the field strength is proportional to the commutator of two covariant derivatives!

It is also easy to check that the covariant derivative transforms covariantly (hence its name) under gauge transformations. Under the gauge transformation in (3.5.7), we have

$$D_\mu \to \partial_\mu - i\frac{e}{\hbar}UA_\mu U^\dagger - \partial_\mu UU^\dagger = U\left(\partial_\mu - i\frac{e}{\hbar}A_\mu\right)U^\dagger = UD_\mu U^\dagger. \tag{3.7.3}$$

From this, the gauge covariance of $F_{\mu\nu}$ automatically follows.

The equations of motion can be written simply using the covariant derivative. Consider a small transformation of A_μ by δA_μ. Therefore, the leading order change to $F_{\mu\nu}$ is

$$\delta F_{\mu\nu} = \partial_\mu \delta A_\nu - \partial_\nu \delta A_\mu - i\frac{e}{\hbar}[A_\mu, \delta A_\nu] - i\frac{e}{\hbar}[\delta A_\mu, A_\nu] = [D_\mu, \delta A_\nu] - [D_\nu, \delta A_\mu]. \quad (3.7.4)$$

Therefore,

$$\begin{aligned}
\delta S &= -\frac{1}{4k}\int d^4x \; 2\mathrm{Tr}\left[\left(\partial_\mu \delta A_\nu - \partial_\nu \delta A_\mu - i\frac{e}{\hbar}[A_\mu, \delta A_\nu] - i\frac{e}{\hbar}[\delta A_\mu, A_\nu]\right) F^{\mu\nu}\right] \\
&= -\frac{1}{k}\int d^4x \mathrm{Tr}\left[(\partial_\mu \delta A_\nu - i\frac{e}{\hbar}[A_\mu, \delta A_\nu])F^{\mu\nu}\right] \\
&= \frac{1}{k}\int d^4x \mathrm{Tr}\left[\delta A_\nu \left(\partial_\mu F^{\mu\nu} - i\frac{e}{\hbar}[A_\mu, F^{\mu\nu}]\right)\right] \\
&= \frac{1}{k}\int d^4x \mathrm{Tr}\left[\delta A_\nu [D_\mu, F^{\mu\nu}])\right]. \quad (3.7.5)
\end{aligned}$$

In going from line 2 to line 3 in (3.7.5) we integrated by parts and used the cyclic properties of the trace. Hence, the equations of motion are

$$[D_\mu, F^{\mu\nu}] = 0. \quad (3.7.6)$$

Notice that this is a matrix equation and hence must be true for all components of the matrix. Notice also that the equation of motion is gauge covariant, and so the solutions to the equation of motion remain solutions after gauge transformations.

4 Topology and Differential Geometry

In this chapter, we introduce the concepts of topology, differential topology and the notion of a fiber bundle. We will see that fiber bundles are particularly useful in studying the topics of the last chapter.

4.1 Rudimentary Topology

The study of topology is a fascinating subject and is rather fun. Most of you probably have some idea of topology, at least in the sense that you know it when you see it. You are probably aware that a doughnut is topologically distinct from a ball, but not from a coffee cup with a handle. The doughnut shape can be smoothly deformed into the shape of a coffee cup, but not into a ball, since there would be no way to smoothly deform away the hole in the doughnut. When dealing with topology, we often speak of *topological invariants*. In this case, one of the topological invariants is the number of holes in the surface.

In this example it was rather simple to figure out a topological invariant, but to go further, we need to make a precise definition of what we mean by a topology and its invariants.

To this end let X be a set and let $Y = \{X_i\}$ be a collection of subsets of X. Then X is a *topological space* and Y is a *topology of* X if

• $\emptyset \in Y$, and $X \in Y$ where $\emptyset$ is the null set.

• For any finite or infinite collection of sets $X_i \in Y$, $\bigcup X_i \in Y$.

• For any *finite* collection of sets $X_i \in Y$, $\bigcap X_i \in Y$.

The subsets X_i are called the *open sets*.

Now here are a few other definitions.

• A *neighborhood* N of x, is a subset of X, containing an open set X_i in the topology of X where $x \in X_i$. Notice that N does not necessarily have to be an open set.

• *Closed Sets.* A closed set is a set whose complement in X is an open set. Since X and $\emptyset$ are each open sets and are complements of each other, we see that these are both open and closed.

• *Closure of a Set.* Consider a set U and consider all closed sets that contain U. Then the closure of U, $\overline{U}$ is the intersection of all those closed sets.

• *Interior.* The interior of a set U, U^I, is the union of all open sets contained by U.

• *Boundary.* The boundary of a set U, $b(U)$ is the closure of U with the interior removed. $b(U) = \overline{U} - U^I$

• *Open cover.* An open cover of a set U is a collection of open sets X_i, such that $U \subset \bigcup X_i$.

With this last definition, we can now give a precise definition of compactness. A set U is compact, if every open cover of the set has a finite subcovering. What this means is that even if we have an infinite collection of open sets covering the set, we can pick out a finite number of these same open sets and still cover the set.

A set X is *connected* if it cannot be written as $X = X_1 \bigcup X_2$, where $X_1 \bigcap X_2 = \emptyset$.

4.2 Examples

These definitions seem rather abstract, so let us consider several examples. First the trivial topology of a space X is where the only open sets are X and $\emptyset$. The discrete topology is where Y contains *all* subsets of X. This will be the last time we consider such things.

Usual Topology of $\mathbf{R}$. Consider the real numbers $\mathbf{R}$. Define the open sets to be all open intervals (a, b), and there unions, where $a < b$ and a and b are not included in the interval. This then satisfies the properties of a topology over $\mathbf{R}$. Note that if we had allowed infinite intersections to be open sets, then individual points in $\mathbf{R}$ would have been open sets, at which point we would get the discrete topology for $\mathbf{R}$. We can also see that this space is not compact, since we can find open coverings that do not have a finite subcover. For example, we can cover $\mathbf{R}$ with the open sets $(n, n+2)$ for all integer n. But no finite collection of these open sets covers $\mathbf{R}$. We can also see that the closure of an open set (a, b) is given by $[a, b]$, the line interval including the endpoints. Note that the interior of $[a, b]$ is (a, b), while the interior of (a, b) is itself.

The open interval. The open interval (a, b), $a < b$ is also an example of a topological space. For the usual topology the open sets are all open intervals (a_i, b_i) and their unions, with $a_i \geq a$, $b_i \leq b$, $a_i < b_i$. These then satisfy the required union and intersection properties. Like $\mathbf{R}$, this is not a compact space, since there exists open coverings without a finite subcover. For instance, the union of the open intervals $(a + (b-a)/n, b)$ for all n positive integer is a cover for (a, b), but we cannot cover (a, b) with a finite number of these open sets. Note further that finite covers of (a, b) exist. But the criterion for compactness is that *every* cover has a finite subcover. Notice that the closed interval $[a, b]$ is not a topological space for the usual topology, since $[a, b]$ is not an open set (however, it is compact).

$\mathbf{R}^n$. We can also see that $\mathbf{R}^n$ is a topological space in more or less the same way that $\mathbf{R}$ is a topological space. R^n is an example of a *metric space*, in that the space comes

with a distance function

$$d(x, y) \geq 0 \tag{4.2.1}$$

and is 0 only if $x = y$. It also satisfies a triangle inequality

$$d(x, z) \leq d(x, y) + d(y, z). \tag{4.2.2}$$

Instead of open intervals we consider *open balls* about all points x, $B_x(\epsilon)$, $y \in B_x(\epsilon)$ if $d(x, y) < \epsilon$. All open balls and their unions forms the topology.

$CP(1)$. Consider the complex projective space, where $(z_1, z_2) \equiv (\lambda z_1, \lambda z_2)$ with $\lambda \neq 0$ and z_1 and z_2 not both zero. This is a topological space with the topology given by a collection of open two dimensional surfaces. However, this space is compact. To see this, note that all but one point on $CP(1)$ is equivalent to $(z, 1)$ for some z. We can consider an infinite union of subsets, say the disks where $|z| < n$, where n is a positive integer. This has no finite subcovering. On the other hand, this is not a covering of $CP(1)$, because it misses the point equivalent to $(1, 0)$. We can cover this last point with the open set containing the points $(1, w)$, where $|w| < \epsilon$. But now this covering has a finite subcover, since we can cover $CP(1)$ with this open set and the previous sets of disks with $|z| < 1/\epsilon + 1$. This is clearly a finite subcover.

4.3 Homeomorphisms and equivalence classes

Consider two topological spaces X_1 and X_2. A *homeomorphism* is a continuous map

$$f : X_1 \to X_2 \tag{4.3.1}$$

whose inverse map, f^{-1} is also continuous. If a homeomorphism exists between two topological spaces, then we say that the two spaces are homeomorphic. Another term to describe these two spaces is to say that they belong to the same *equivalence class*.

A *topological invariant* is some characteristic of a topological space that is invariant under homeomorphisms. Hence, two spaces with different topological invariants are not of the same equivalence class.

Some examples of topological invariants are compactness and connectedness. Suppose that X is compact and that $f(X) = Y$. If f is a homeomorphism, then $f^{-1}(Y) = X$, and so the open sets of Y map to the open sets of X. Hence an open cover of Y maps to an open cover of X. The open cover of X has a finite subcover, which then maps back to a finite subcover of Y. Connectedness is also easy to show.

One can also show that the dimension of the space is a topological invariant. For example, let us consider $\mathbf{R}^n$. Start with $\mathbf{R}$ and assume that $\mathbf{R}$ is homeomorphic to $\mathbf{R}^2$. That is, there is an invertible map that takes $\mathbf{R}$ to $\mathbf{R}^2$. If such a map exists, then it maps a point in $\mathbf{R}$ to a point in $\mathbf{R}^2$ and the inverse maps this point back to the original point. Let us remove a point p from $\mathbf{R}$ and remove the point $f(p)$ from $\mathbf{R}^2$. Then if $\mathbf{R}$ is homeomorphic to $\mathbf{R}^2$, then $R - \{p\}$ is homeomorphic to $\mathbf{R}^2 - \{f(p)\}$. However, $R - \{p\}$ is not connected but $\mathbf{R}^2 - \{f(p)\}$ is connected. Hence these two spaces cannot be homeomorphic. Therefore, we have a contradiction, so $\mathbf{R}$ is not homeomorphic to $\mathbf{R}^2$. We can keep continuing this argument for the higher $\mathbf{R}^n$, showing that $\mathbf{R}^m$ is not homeomorphic to $\mathbf{R}^n$ if $m \neq n$.

4.4 Differential Manifolds

Let us try and make the discussion of topological spaces even more concrete. To this end, let us define a *differential manifold* $\mathcal{M}$.

Let $\mathcal{M}$ be a topological space which has the open cover

$$\mathcal{M} \subset \bigcup M_\alpha \tag{4.4.1}$$

The open sets M_α are homeomorphic to open subsets O_α of $\mathbf{R}^n$, through the invertable map

$$\phi_\alpha : M_\alpha \to O_\alpha. \tag{4.4.2}$$

In the intersection region where $M_\alpha \cap M_\beta \neq \emptyset$, we have the map

$$\phi_\beta \circ \phi_\alpha^{-1} : \phi_\alpha(M_\alpha \cap M_\beta) \to \phi_\beta(M_\alpha \cap M_\beta). \tag{4.4.3}$$

If this map is infinitely differentiable (also known as C^∞), then $\mathcal{M}$ is a differentiable manifold. The dimension of the manifold is given by n.

Examples Let us consider 1 dimensional connected manifolds. The only two examples are the real line $\mathbf{R}$ and the circle S_1. Clearly, $\mathbf{R}$ is a manifold. To show that S_1 is a manifold, let us write the circle in terms of its x, y coordinates $(\cos\theta, \sin\theta)$. The maps to the open subsets of $\mathbf{R}$ map the (x, y) coordinate to θ. We can cover the circle with the two open subsets, $0 < \theta < 3\pi/2$ and $\pi < \theta < 5\pi/2$. Then the maps $\phi_2 \circ \phi_1^{-1}$, map $\theta \to \theta$ in the first intersection region and $\theta \to \theta + 2\pi$ in the second intersection region. Clearly these maps are C^∞, therefore, S_1 is a manifold.

4.5 Differential forms on manifolds

Since ϕ_α maps M_α into $\mathbf{R}^n$, the function $\phi_\alpha(p)$ where p is a point on $\mathcal{M}$, has n components. Hence we can write this as

$$\phi_\alpha(p) = (x^1_\alpha(p), x^2_\alpha(p), ...x^n_\alpha(p)) \tag{4.5.1}$$

The components x^i_α are called the *local coordinates* on $\mathcal{M}$. The open region M_α is called the *coordinate patch* for these local coordinates. With these local coordinates, we can describe derivatives and integrals on manifolds. Since the components are understood to be local, we will drop the α subscript on them.

Suppose that we have a curve parametrized by a variable τ on our manifold $\mathcal{M}$. Hence we have a function where $p(\tau)$ corresponds to the point on the curve for parameter τ. The local coordinates on the curve are given by $x^i(p(\tau))$. The tangent to the curve is found by taking the derivative with respect to τ

$$\frac{d}{d\tau}x^i(p(\tau)). \tag{4.5.2}$$

Suppose that we generalize this to any function $f(p)$, which we reexpress in terms of the local coordinates. Then the rate of change of this function along the curve is

$$\frac{\partial f}{\partial x^i}\frac{dx^i}{d\tau} \equiv a^i\frac{\partial f}{\partial x^i} \tag{4.5.3}$$

Since the function f is arbitrary, we can instead refer to a *tangent vector* over p,

$$a^i\frac{\partial}{\partial x_i}, \tag{4.5.4}$$

where the a^i are the components of the vector. If we consider all possible curves through a point p, then the allowed values for a^i span $\mathbf{R}^n$. Hence the tangent vectors over p make up a linear vector space called the *tangent space* of p. This is usually written as

$$T_p(\mathcal{M}). \tag{4.5.5}$$

The operators $\partial_i \equiv \frac{\partial}{\partial x^i}$ are a basis for the tangent space, hence the tangent space over p is n dimensional, the same dimension as the manifold.

Given this vector space, we next construct what is known as a *dual space*. To this end consider a vector space V and a linear map α such that

$$\alpha : V \to \mathbf{R} \tag{4.5.6}$$

where $\mathbf{R}$ refers to the real numbers. The fact that this is linear means that $\alpha(v_1 + v_2) = \alpha(v_1) + \alpha(v_2)$. But the space of *all* such linear maps is itself a vector space. This space is called the *dual space* and is written as V^*.

To find the dual of the tangent space, we note that for any curve parametrized by τ passing through point p, and any real function $f(p)$, the tangent $\frac{d}{d\tau}f$ is itself a real number. We also note that the differential df can be written as

$$df = \frac{\partial f}{\partial x^i} dx^i \tag{4.5.7}$$

Allowing for a general real function, we see that $\frac{\partial f}{\partial x^i}$ spans $\mathbf{R}^n$ at *each point* p. Since $d(f + g) = df + dg$, we see that this too defines a linear vector space and the differentials dx^i form a basis of this space. Hence, this space has the same dimension as $\mathcal{M}$. Moreover, this defines the dual space to $T_p(\mathcal{M})$, $T_p^*(\mathcal{M})$ since $\frac{df}{d\tau}$ is a linear map from $\frac{dx^i}{d\tau}\partial_i$ to $\mathbf{R}$ at each point p. We can write these relations in bra-ket notation, with

$$\langle dx^i | \partial_j \rangle = \delta^i{}_j \tag{4.5.8}$$

so that

$$\langle df | \frac{dx^i}{d\tau}\partial_i \rangle = \frac{\partial f}{\partial x^i}\frac{dx^i}{d\tau} = \frac{df}{d\tau}. \tag{4.5.9}$$

The space $T_p^*(\mathcal{M})$ is called the cotangent space at point p. Elements of this space df are called *1-forms*.

It turns out that we can generalize 1-forms to m-forms, where $m \leq n$. To this end, we define the *wedge product* between two 1-forms to satisfy

$$dx^i \wedge dx^j = -dx^j \wedge dx^i. \tag{4.5.10}$$

A *2-form* $\boldsymbol{\omega}$ is then given by

$$\boldsymbol{\omega} = \omega_{ij} dx^i \wedge dx^j \tag{4.5.11}$$

where $\omega_{ij} = -\omega_{ji}$. We have written this 2-form in terms of the local coordinates on the coordinate patch, but this can be mapped back to the open region in $\mathcal{M}$. The 2-forms also form a linear vector space with basis vectors $dx^i \wedge dx^j$. This space is $n(n-1)/2$ dimensional and is written as $\Omega^2(\mathcal{M})$. We can also take several wedge products to make higher forms . In fact, we can continue doing this all the way up to n-forms. However, there cannot be m-forms with $m > n$, since the space of 1-forms is n-dimensional but the wedge product antisymmetrizes between the forms. All m-forms form a linear vector space, $\Omega^m(\mathcal{M})$.

We can also combine these forms together to make a one bigger vector space for all m-forms,

$$\Omega(\mathcal{M}) = \Omega^0(\mathcal{M}) \oplus \Omega^1(\mathcal{M}) \oplus ...\Omega^n(\mathcal{M}), \tag{4.5.12}$$

where $\Omega^0(\mathcal{M})$ are the 0-forms, which are basically functions.

There is a very useful operator that maps m-forms to $m+1$-forms known as the *exterior derivative*, $\boldsymbol{d}$. By definition, we have that $\boldsymbol{d}$ acting on a m-form $\boldsymbol{\Lambda}$ is

$$\boldsymbol{d\Lambda} = \partial_i \left(\Lambda_{j_1 j_2 ... j_m}\right) dx^i \wedge dx^{j_1} \wedge ... \wedge dx^{j_m}, \tag{4.5.13}$$

which is clearly an $m+1$-form. It then follows that $\boldsymbol{d}^2 = 0$, since $\boldsymbol{d}^2$ acting on any m-form $\boldsymbol{\Lambda}$ is

$$\boldsymbol{d}^2\boldsymbol{\Lambda} = \partial_i\partial_j \left(\Lambda_{k_1 k_2 ... k_m}\right) dx^i \wedge dx^j \wedge dx^{k_1} \wedge ... \wedge dx^{k_m} = 0, \tag{4.5.14}$$

since $\partial_i\partial_j = \partial_j\partial_i$. Any operator whose square is 0 is said to be *nilpotent*.

Any form $\boldsymbol{\Phi}$ that satisfies $\boldsymbol{d\Phi} = 0$ is said to be *closed*. Any form that can be written as $\boldsymbol{d\Lambda}$ is said to be *exact*. Obviously, every exact form is closed, but the converse is not true. However, it is almost true, in that *locally* every closed form can be written as the exterior derivative of another form.

Let us now consider some examples. First consider the circle S_1. The 0-forms are then functions on the circle, that is functions that are periodic under $\theta \to \theta + 2\pi$. The 1-forms are of the form $g(\theta)d\theta$ where $g(\theta)$ is periodic. The exact 1-forms are $df(\theta)$. However, not all closed 1-forms are exact. For example $d\theta$ is closed, but not exact, since θ is not a 0-form on the circle (it is not periodic). However, $d\theta$ is locally exact. If we consider the two open regions of the previous section, then θ, or $\theta + 2\pi$ is a 0-form in $\mathbf{R}^1$.

For the next example, let us consider $U(1)$ gauge fields in 4-d space-time. This space-time is homeomorphic to $\mathbf{R}^4$. Recall that the gauge fields are 4-vectors, and that the relevant term in the action for a charged particle in an electromagnetic field is $A_\mu dx^\mu$. Hence the gauge field can be thought of as a 1-form, $\boldsymbol{A}$. From (3.5.1), it is also clear that the field strength satisfies

$$F_{\mu\nu} dx^\mu \wedge dx^\nu = \boldsymbol{F} = \boldsymbol{dA}, \tag{4.5.15}$$

hence $\boldsymbol{F}$ is an exact 2-form. Thus, we see that $\boldsymbol{dF} = 0$, which is the Bianchi identity in (3.6.3). We now see that the Bianchi identity is a consequence of $\boldsymbol{d}^2 = 0$. We also see that gauge tranformations can be written as

$$\boldsymbol{A} \to \boldsymbol{A} + \boldsymbol{d\Phi} \tag{4.5.16}$$

where the gauge parameter $\boldsymbol{\Phi}$ is a 0-form. It is then obvious that $\boldsymbol{F}$ is invariant under the gauge transformations because of the nilpotence of $\boldsymbol{d}$.

4.6 Integrals of forms

As you might have guessed by the dx^i terms, m-forms may be integrated over m dimensional surfaces, or over $\mathcal{M}$ itself. The definition of an integral of an m-form over an m-dimensional surface is as follows. Let Σ be a manifold and let $\{\Sigma_\alpha\}$ be an open covering where the Σ_α have invertible maps ϕ_α into $\mathbf{R}^m$. We also assume that the covering is locally finite, which means that each point in Σ is covered a finite number of times by the covering. An integral of an m-form $\mathbf{\Lambda}$ over an open region Σ_α is then given by

$$\int_{\Sigma_\alpha} \mathbf{\Lambda} = \int_{\phi_\alpha(\Sigma_\alpha)} \Lambda_{123\ldots m} dx^1 dx^2 .. dx^m. \tag{4.6.1}$$

To find the integral over Σ, we consider a partition of unity, where we have the functions $e_\alpha(p)$, with the properties

$$0 \leq e_\alpha(p) \leq 1, \qquad e_\alpha(p) = 0 \text{ if } p \notin \Sigma_\alpha \qquad \sum_\alpha e_\alpha(p) = 1. \tag{4.6.2}$$

Hence, we can define the full integral to be

$$\int_\Sigma \mathbf{\Lambda} = \sum_\alpha \int_{\phi_\alpha(\Sigma_\alpha)} e_\alpha(p) \Lambda_{123\ldots m} dx^1 dx^2 .. dx^m. \tag{4.6.3}$$

One important result is Stokes Law. Suppose we have an open region Σ_α which has a boundary $\partial\Sigma_\alpha$, then the integral of the exact form $\boldsymbol{d\Lambda}$ on this region satifies

$$\int_{\Sigma_\alpha} \boldsymbol{d\Lambda} = \int_{\partial\Sigma_\alpha} \mathbf{\Lambda}. \tag{4.6.4}$$

4.7 The monopole revisited

Let us recall the magnetic monopole, with a field strength given by (3.4.1). To find the charge, we integrated F_{ij} over the surface of S_2. As we previously saw, we could not define F_{ij} in terms of the curl of $\vec{A}$ over the entire sphere, instead we had to break the sphere up into two regions, with a different gauge field in each region.

We have already seen that $\boldsymbol{F}$ is a 2-form. For the monopole, it is a closed 2-form on S_2, but it is not an exact 2-form. We know that it is closed because it is locally exact. On the northern hemisphere we have that $F = \boldsymbol{dA}_1$ and on the southern hemisphere it is $F = \boldsymbol{dA}_2$. The monopole charge is given by the integral

$$\int_{S_2} \boldsymbol{F} = \int_N \boldsymbol{dA}_1 + \int_S \boldsymbol{dA}_2 = -\int_{S_1} \boldsymbol{d\Phi}. \tag{4.7.1}$$

4.8 Fiber bundles

We have seen in the previous sections that a manifold has associated with it a tangent space and cotangent space at *each* point p in the manifold. One might ask whether or not one could combine these spaces with the underlying manifold $\mathcal{M}$ to make one big topological space. The answer to this is yes. In fact we can combine other spaces with the manifold to make a bigger topological space. The whole space is known as a *fiber bundle*.

A fiber bundle is defined as follows:

• Let E be a topological space, called the *total space* or the *bundle space*.

• There is a projection $\Pi : E \to X$ of E onto X, where X is also a topological space, called the *base space* or just the *base*.

• There exists another topological space F called the *fiber*, along with a group G of homeomorphisms of F into itself.

• There is a cover $\{X_\alpha\}$ of X as well as a set of homeomorphisms ϕ_α such that

$$\phi_\alpha : \Pi^{-1}(X_\alpha) \to X_\alpha \times F. \tag{4.8.1}$$

The expression $\Pi^{-1}(X_\alpha)$ refers to that part of E that projects down to X_α.

• The inverse function for ϕ_α satisfies

$$\Pi\phi_\alpha^{-1}(x, f) = x, \qquad \text{where } x \in X_\alpha \text{ and } f \in F. \tag{4.8.2}$$

In other words, we assume that the maps take us from the same point in x in the base space into the same point in x that appears in the product space.

In the overlapping regions where $X_\alpha \cap X_\beta \neq \emptyset$ we have that

$$\phi_\alpha \circ \phi_\beta^{-1} : (X_\alpha \cap X_\beta) \times F \to (X_\alpha \cap X_\beta) \times F. \tag{4.8.3}$$

Since the point $x \in X_\alpha \cap X_\beta$ maps to itself, these maps define homeomorphisms of F, given by $g_{\alpha\beta}(x)$. These homeomorphisms are called *transition functions* and they are required to live in the group G, also known as the *structure group* of the bundle. The bundle is usually given by the data (E, Π, F, G, X)

A bundle E is said to be *trivial* if E is homeomorphic to the product space $X \times F$, where X is the base. Based on the construction of a fiber bundle, we see that all bundles are at least locally trivial. We also should note that there is some similarity in the definition of a fiber bundle and the definition of a manifold. For the manifold we had local maps to $\mathbf{R}^n$, whereas for the bundle, we had local maps to a product space. In the case of the manifold, we considered maps from $\mathbf{R}^n$ to $\mathbf{R}^n$ over the intersection regions. For the bundle we considered maps from the product space to the product space in the intersecting regions.

4.9 Examples of fiber bundles

The simplest example of a nontrivial bundle is the Mobius strip. Let the base be S_1 and the fiber be the line segment $[-1, 1]$. We cover S_1 with two overlapping open regions that intersect in two seperate regions. In the first intersecting region, $g_{12}(x) = 1$, in other words, the map takes the fiber to itself. In the other region, because of the twisting of the strip as we go around the circle, the map takes every point in the fiber $f \to -f$. Hence $g_{12}(x) = -1$ in this region. Hence the structure group consists of two group elements, the identity and a nontrivial element g where $g^2 = 1$. Hence the structure group is Z_2.

The next example we consider is the *tangent bundle* for $\mathcal{M}$, $T(\mathcal{M})$. The bundle is given by the union

$$T(\mathcal{M}) = \bigcup T_p(\mathcal{M}), \tag{4.9.1}$$

for all $p \in \mathcal{M}$. The base space is $\mathcal{M}$ itself and the fiber for any point p is its tangent space $T_p(\mathcal{M})$. To find the transition functions, we consider the intersection of two coordinate patches M_α and M_β, and a point p that lives in this intersecting region. A tangent vector at p then satisfies

$$a^i_\alpha(p)\frac{\partial}{\partial x^i_\alpha} = a^i_\beta(p)\frac{\partial}{\partial x^i_\beta} = a^i_\beta\frac{\partial x^j_\alpha}{\partial x^i_\beta}\partial x^i_\beta. \tag{4.9.2}$$

Hence we have that

$$a^i_\alpha = a^j_\beta\frac{\partial x^i_\alpha}{\partial x^j_\beta}, \tag{4.9.3}$$

Hence the transition functions are given by

$$g_{\alpha\beta}(p) = \frac{\partial x^i_\alpha}{\partial x^j_\beta}. \tag{4.9.4}$$

These are elements of the group $GL(n, \mathbf{R})$, the group of *general linear transformations* on an n dimensional real vector space. Hence the structure group is $GL(n, \mathbf{R})$.

We next consider the cotangent bundle, $T^*(\mathcal{M})$. This proceeds in a fashion similar to the tangent bundle. Now the elements of the fibers in a particular coordinate patch have the form

$$b^\alpha_i dx^i_\alpha. \tag{4.9.5}$$

Hence the fibers in the intersecting regions have the relation

$$b^\alpha_i = b^\beta_j\frac{\partial x^j_\beta}{\partial x^i_\alpha}. \tag{4.9.6}$$

Hence the transition functions have the form

$$g_{\alpha\beta} = \frac{\partial x^j_\beta}{\partial x^i_\alpha}, \tag{4.9.7}$$

and the group is once again $GL(n, \mathbf{R})$. Note the difference in the transition functions for $T^*(\mathcal{M})$ and $T(\mathcal{M})$. The bundles $T(\mathcal{M})$ and $T^*(\mathcal{M})$ are examples of *vector bundles*, bundles whose fibers are vector spaces.

The next examples we consider are called *principle bundles*. A principle bundle is a bundle whose fibers are actually the transition functions. In other words, points in the fiber are elements of the structure group. It is straightforward to show that the homeomorphisms of the fiber are actually the group elements. This is because for any $g \in G$, $g : G \to G$ and the map is invertible since every group element has a unique inverse. Hence the transition functions are elements of the structure group.

The first example of a nontrivial principle bundle is one where we have Z_2 over the circle. The fiber then has two group elements, the identity e and g, where $g^2 = e$. Breaking up the circle into two open regions and two intersection regions, we see that in one of the intersecting regions $e(e, g)$ is mapped to $e, g)$ and in the other intersecting region (e, g) is mapped to $(g, e) = g(e, g)$, hence g is the transition function in this region.

The second principle bundle we shall consider is known as a *frame bundle*, where the fibers are the space of all frames for $T_p(\mathcal{M})$ over a point p. This means the fibers are all possible bases for the tangent space. Since one basis is rotated into another basis through a general linear transformations, the frames can be defined as a general linear transformation of some fixed reference frame. Hence the fibers are essentially elements of $GL(n, \mathbf{R})$.

We will later see that gauge theories are closely associated with principle bundles of compact Lie groups.

4.10 More on fiber bundles

A *section* of a fiber bundle is a continuous map s from the base space to the bundle, which satisfies

$$s : X \to E \qquad \Pi(s(x)) = x \qquad x \in X. \tag{4.10.1}$$

Sometimes this is called a *global section*. Sections (that is global sections) don't always exist.

In fact, for a principle bundle, the existence of a section means that the bundle is trivial. To see this, note that a section satisfies $s(x) \in G$, where G is the structure group.

Hence the fiber over a point x is generated by taking $gs(x)$ for all $g \in G$ and the bundle is generated by taking all g and all points x. Hence an element of the bundle E is given by $gs(x)$. But given this, there clearly is a continuous map from E to $X \times G$, since $s(x)$ is continuous and g ranges over G. Namely, the map takes $gs(x) \to (x, g)$. The inverse map back to E is just $gs(x)$, so E is homeomorphic to $X \times G$ and so the bundle is trivial.

So for the example of the (E, Π, Z_2, Z_2, S_1), the only possible sections are $s(x) = g$ or $s(x) = e$. But for the nontrivial bundle, it is clear that such an identification is not possible over the whole circle.

For any bundle E, we can always construct a principle bundle $P(E)$, by replacing the fibers in E with the transition functions of E, while keeping the transition functions the same. We now show that if $P(E)$ is trivial, then E itself is trivial.

First we argue that if a bundle E is equivalent to another bundle E' with the same base, fiber and structure group then the transition functions of E' are related to those of E by

$$g'_{\alpha\beta}(x) = g_\alpha^{-1}(x) g_{\alpha\beta}(x) g_\beta(x), \tag{4.10.2}$$

where the $g_\alpha(x)$ live in the structure group. Suppose that X is covered by X_α and for E we have the maps ϕ_α into the product space and for E' we have the maps ψ_α. This just constitutes a relabeling of the fiber coordinates. Then

$$\phi_\alpha \circ \psi_\alpha^{-1} : X_\alpha \times F \to X_\alpha \times F. \tag{4.10.3}$$

Since the map fixes x, this defines a homeomorphism of F, $g_\alpha(x)$, which is in the structure group. It is then clear that the relation between the structure constants is as in (4.10.2). We can also argue the converse, that given (4.10.2), then the bundles E and E' are equivalent since we can explicitly construct the homeomorphism between the two bundles.

To show that E is trivial if $P(E)$ is trivial, we note that for any trivial bundle, the transition functions have the form

$$g_{\alpha\beta}(x) = g_\alpha^{-1}(x) g_\beta(x), \tag{4.10.4}$$

since there is a homeomorphism that maps all the transition functions to the identity. But since the principle bundle is trivial, it means that it has transition functions of this form, and since E has the same transition functions, it too has functions of this form. Hence if $P(E)$ is trivial then E is trivial.

4.11 Connections and curvatures on bundles

In this section we would like to look at gauge theories from a geometric perspective. We will assume that the gauge theory lives on some manifold $\mathcal{M}$ and that the group associated with it is G.

Let us consider a principle bundle P over a manifold $\mathcal{M}$ with structure group G. Let us further assume that G is continuous. Locally, the coordinates in the bundle look like (x, g), where $x \in \mathcal{M}$ and $g \in G$. Now since $\mathcal{M}$ is a manifold and G is continuous, then P itself is a manifold and as such it has a tangent and a contangent space over each point in the bundle. If n is the dimension of the manifold and d is the dimension of the group, then $n + d$ is the dimension of the tangent and cotangent spaces.

We can also construct a tangent bundle for P, $T(P)$ and a cotangent bundle $T^*(P)$ and consider vectors and 1-forms that live in these bundles. To this end, let us consider the following 1-form $\boldsymbol{\omega}$ which we write as[4]

$$\boldsymbol{\omega} = ig^{-1}\boldsymbol{d}g + g^{-1}\boldsymbol{A}g \tag{4.11.1}$$

where $g \in G$ and $\boldsymbol{A} = A^a{}_\mu T^a dx^\mu$. In general, g and $\boldsymbol{A}$ are matrices, hence $\boldsymbol{\omega}$ is matrix valued. So for instance, the first term in (4.11.1) has components

$$\left[g^{-1}dg\right]_{ij} = [g^{-1}]_{ik} dg_{kj}. \tag{4.11.2}$$

Hence we see that $\boldsymbol{A}$ is a 1-form on $\mathcal{M}$, while $\boldsymbol{\omega}$ is 1-form for P. Note that this is given for a particular point in P, (x, g). If $A = 0$, then we see that $\boldsymbol{\omega}$ points along the direction of the fiber, the only differentials are for g and not x.

The differential in terms of a group element may seem a little strange. One could think about this is in terms of local coordinates, where for a small change in the group element, we have

$$g \to g \exp i\epsilon^a T^a, \tag{4.11.3}$$

where the T_a are the generators of the Lie algebra and the $\epsilon_a \to 0$. The ϵ_a can be thought of as the limiting values for the local coordinates θ_a on the group manifold. Hence $g^{-1}\boldsymbol{d}g = T^a d\theta^a$. This is very similar to the $1-forms$ on $\mathcal{M}$, which we often write in terms of its local coordinates. So for a $U(1)$ group, the group manifold is a circle and $g^{-1}\boldsymbol{d}g = d\theta$. For $SU(2)$, we saw that the group manifold was S_3, so the differentials can be thought of as differentials for the local coordinates on S_3. Having said this, it is more convenient to leave the form as in (4.11.1).

[4]We have dropped the factors of $e/\hbar$. This can be done under a rescaling of the gauge fields.

One problem with the local coordinates is the noncommutativity of g with dg. In fact, we could have defined the differentials so that $g^{-1}dg = ig^{-1}T^a d\theta^a g$. One further advantage of leaving the differentials as $g^{-1}dg$ is that they are invariant under a global transformation, $g \to hg$, where h is a group element. This invariance is called a *left invariance*.

Next consider elements of $T(P)$ which in terms of the group coordinates and local coordinates on $\mathcal{M}$ is given by

$$C^\mu X_\mu = C^\mu \left(iB_{\mu_{ij}} \frac{\partial}{\partial g_{ij}} + \frac{\partial}{\partial x^\mu} \right). \tag{4.11.4}$$

We now break up the tangent space into two pieces, a vertical piece $V(P)$ and a horizontal piece $H(P)$. By definition, we define the vertical part to be tangent vectors of the form

$$\gamma_{ij} \frac{\partial}{\partial g_{ij}}, \tag{4.11.5}$$

hence these are vectors that point along the fiber and not $\mathcal{M}$. The horizontal piece should be chosen such that it is orthogonal to $V(P)$ and so that $T(P) = V(P) \oplus H(P)$. But recall that inner products are not taken between tangent vectors, but between a tangent vector and a cotangent vector. We will define $H(P)$ then to be those tangent vectors whose inner product with $\boldsymbol{\omega}$ is zero. If $\boldsymbol{A} = 0$, then clearly $H(P)$ is $\mathcal{M}$ itself, in other words the horizontal space is just the base of the fiber. If $\boldsymbol{A} \neq 0$ then the result is more interesting.

Why bother defining a horizontal space? The reason is that given a point in the bundle which is locally (x, g), we would like to see how the group element g changes as we move along a curve in $\mathcal{M}$. In other words, the curve on $\mathcal{M}$ can be *lifted* onto a curve in the bundle where the curve passes through one point on the fiber. The lifted curve is determined by $H(P)$, in that we find the curve by lifting a point in x to a point in the fiber, and then moving along the curve such that we lie in $H(P)$. So let a curve be locally parameterized by $x(\tau)$, where τ is a value that parameterizes the curve. Then the lifted curve is given by $g(x(\tau))$. If $\boldsymbol{A}$ is zero, then $g(x) = g$ is contant in x. The possible lifts over a curve in the base space are then parallel since g does not change along the curve.

If $\boldsymbol{A}$ is not zero, then we say that the lifts *parallel transport* the fibers along the curve. $\boldsymbol{A}$ is called the *connection*, since it tells us how to parallel transport a fiber at a point x to another fiber at x'. Given $\boldsymbol{\omega}$ in (4.11.1), then the elements $X \in H(P)$ satisfy

$$\begin{aligned} \langle \boldsymbol{\omega}_{ij} | X_\mu \rangle &= \langle [ig^{-1}]_{ik} \boldsymbol{d} g_{kj} + \left[g^{-1} A^a{}_\nu T^a g \right]_{ij} dx^\nu | \frac{\partial}{\partial x^\mu} + iB_{\mu_{lm}} \frac{\partial}{\partial g_{lm}} \rangle \\ &= \left[g^{-1} A^b{}_\mu T^b g - g^{-1} B_\mu \right]_{ij} = 0. \end{aligned} \tag{4.11.6}$$

Hence we find that

$$B_{\mu ij} = \left[A^b{}_\mu T^b g\right]_{ij}. \tag{4.11.7}$$

This then determines B^a_μ and hence the horizontal space $H(P)$.

With the above constraint, a tangent vector in the horizontal piece of the bundle is given by

$$D_\mu = \frac{\partial}{\partial x^\mu} + i\left[A^a{}_\mu T^a g\right]_{ij} \frac{\partial}{\partial g_{ij}}. \tag{4.11.8}$$

As this is written, this does not look particularly enlightening. Let us however, note the following. Suppose we are considering some matrix valued quantity whose dependence on g is of the form $\widetilde{W} = gWg^{-1}$ but which is otherwise independent of g. Under the transformation $g \to hg$, we see that $\widetilde{W} \to h\widetilde{W}h^{-1}$. Recall from the previous section that this is how a gauge covariant object transformed under a gauge transformation h. If we act with D_μ in (4.11.8) on gWg^{-1}, we find

$$D_\mu\left(gWg^{-1}\right) = \frac{\partial}{\partial x^\mu}\widetilde{W} + iA^a{}_\mu T^a gWg^{-1} - igWg^{-1}A^a{}_\mu T^a = [\partial_\mu + iA^a{}_\mu T^a, \widetilde{W}], \tag{4.11.9}$$

where we used that $\frac{\partial}{\partial g_{ij}} g_{kl} = \delta^i{}_k \delta^j{}_l$. In other words, this is the covariant derivative of the previous chapter. Hence we see that the covariant derivatives in gauge theories can be thought of as the covariant derivatives on the principle bundles and that the gauge fields are the connections. The covariant derivative in (4.11.8) can be used for quantities that transform other ways under the gauge transformations. For example, in particle physics, one often considers fields ψ that transform as $\psi \to h\psi$. It is straightforward to determine how the covariant derivative acts on these fields.

Let us now show that the left multiplication of g by h are indeed the gauge transformations. Note that h need not be constant over $\mathcal{M}$. Actually to do this, we need to say something more about $\boldsymbol{\omega}$. $\boldsymbol{\omega}$ is assumed to be invariant under a change of bundle coordinates that leaves x fixed. In other words, $\boldsymbol{\omega}$ is invariant under $g \to hg$. With this assumption, we see that $\boldsymbol{A}$ must change. In particular, we have that

$$ig^{-1}\boldsymbol{d}g + g^{-1}\boldsymbol{A}g = ig^{-1}h^{-1}\boldsymbol{d}(hg) + g^{-1}h^{-1}\boldsymbol{A}'hg. \tag{4.11.10}$$

Then this gives that

$$\boldsymbol{A}' = -i\boldsymbol{d}h\ h^{-1} + h\boldsymbol{A}h^{-1} \tag{4.11.11}$$

which is our gauge transformation from the previous chapter. Recall that in the last section we argued that two bundles are equivalent if $g \to hg$ for all elements of the fibers.

Hence we learn that gauge transformations does not change the equivalency of a bundle. In other words, equivalent bundles are equal up to a gauge transformation.

Now consider the commutator of two covariant derivatives. This is

$$[D_\mu, D_\nu] = i(\partial_\mu A_\nu - \partial_\nu A_\mu) g\frac{\partial}{\partial g} - A^a{}_\mu A^b\nu \left(T^a g\frac{\partial}{\partial g} T^b g\frac{\partial}{\partial g} - T^b g\frac{\partial}{\partial g} T^a g\frac{\partial}{\partial g} \right). \quad (4.11.12)$$

The quantity in the last parentheses satisfies

$$T^a_{ij} g_{jk}\frac{\partial}{\partial g_{ik}} T^b_{lm} g_{mn}\frac{\partial}{\partial g_{ln}} - T^b_{ij} g_{jk}\frac{\partial}{\partial g_{ik}} T^a_{lm} g_{mn}\frac{\partial}{\partial g_{ln}} = (T^b T^a - T^a T^b) g\frac{\partial}{\partial g}. \quad (4.11.13)$$

Therefore, we have

$$[D_\mu, D_\nu] = i(\partial_\mu A_\nu - \partial_\nu A_\mu - i[A_\mu, A_\nu]) g\frac{\partial}{\partial g} = iF_{\mu\nu} g\frac{\partial}{\partial g}. \quad (4.11.14)$$

Thus we see that the field strength is associated with the curvature on the principle bundle.

Note that the 2-form $\boldsymbol{F}$ can be written as

$$\boldsymbol{F} = \boldsymbol{dA} - i\boldsymbol{A} \wedge \boldsymbol{A}. \quad (4.11.15)$$

But we also see that we can find a nice relation in terms of $\boldsymbol{\omega}$. Note that

$$\begin{aligned} \boldsymbol{d\omega} - i\boldsymbol{\omega} \wedge \boldsymbol{\omega} &= \boldsymbol{d}(ig^{-1}\boldsymbol{d}g + g^{-1}\boldsymbol{A}g) - i(ig^{-1}\boldsymbol{d}g + g^{-1}\boldsymbol{A}g) \wedge (ig^{-1}\boldsymbol{d}g + g^{-1}\boldsymbol{A}g) \\ &= -ig^{-1}\boldsymbol{d}g \wedge g^{-1}\boldsymbol{d}g - g^{-1}\boldsymbol{d}g \wedge g^{-1}\boldsymbol{A}g + g^{-1}\boldsymbol{dA}g - g^{-1}\boldsymbol{A} \wedge \boldsymbol{d}g \\ &\quad + ig^{-1}\boldsymbol{d}g \wedge g^{-1}\boldsymbol{d}g + g^{-1}\boldsymbol{d}g \wedge g^{-1}\boldsymbol{A}g + g^{-1}\boldsymbol{A}\boldsymbol{d}g - ig^{-1}\boldsymbol{A} \wedge \boldsymbol{A}g \\ &= g^{-1}(\boldsymbol{dA} - i\boldsymbol{A} \wedge \boldsymbol{A})g. \end{aligned} \quad (4.11.16)$$

In deriving this last equation, we used the fact that $\boldsymbol{d}(\boldsymbol{A}g) = \boldsymbol{dA}g - \boldsymbol{A}\boldsymbol{d}g$. Hence, we see that $\boldsymbol{F} = g(\boldsymbol{d\omega} - i\boldsymbol{\omega} \wedge \boldsymbol{\omega})g^{-1}$.

4.12 Cohomology

We have argued that principle bundles are equivalent, if the fibers are relabled through the transformation $g \to hg$. We also showed that this corresponds to a gauge transformation for the connection 1-form. So it is important to determine whether two principle bundles are equivalent or not. It turns out that there is a nice way of doing this using something called *charateristic classes*. But before doing this, we need to learn a little cohomology.

Let us consider the differential manifold $\mathcal{M}$ with dimension n and the vector space of forms on this

$$\Omega(\mathcal{M}) = \sum_{r=0}^{n} \oplus \Omega^r(\mathcal{M}). \tag{4.12.1}$$

Recall that an r-form $\mathbf{\Lambda}$ is given in terms of local coordinates as

$$\mathbf{\Lambda}_{i_1 \ldots i_r} dx^{i_1} \wedge \ldots \wedge dx^{i_r}. \tag{4.12.2}$$

We also recall that the exterior derivative $\boldsymbol{d}$ takes an r-form to an $(r+1)$ form and that $\boldsymbol{d}^2 = 0$. We say that $\boldsymbol{d}$ is nilpotent and it turns out that with such nilpotent operators one can find topological invariants.

In particular, let us consider all r-forms $\boldsymbol{\lambda}$ that satisfy $\boldsymbol{d}\boldsymbol{\lambda} = 0$. Such forms are called closed. Certainly forms that satisfy $\boldsymbol{\lambda} = \boldsymbol{d}\Phi$ are themselves closed because of the nilpotence of $\boldsymbol{d}$. Such forms are called exact. The question is, are all closed forms exact? The answer is no, but it is this fact that makes them interesting.

Instead of considering all closed forms, we will instead consider all closed forms modded out by exact forms. In other words, two closed forms are said to be equivalent (or cohomologous) if

$$\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_2 + \boldsymbol{d}\boldsymbol{\Phi} \qquad \rightarrow \qquad [\boldsymbol{\lambda}_1] = [\boldsymbol{\lambda}_2] \tag{4.12.3}$$

for any $\boldsymbol{\Phi}$. This equivalence among closed forms determines a *cohomology* and the elements of the cohomology are the closed forms modded out by all exact forms. We write this as

$$H^r(\mathcal{M}, \mathbf{R}) = Z^r(\mathcal{M})/B^r(\mathcal{M}), \tag{4.12.4}$$

where Z^r refers to all closed r-forms and B^r refers to all exact r-forms. The $\mathbf{R}$ in $H^r(\mathcal{M}, \mathbf{R})$ means that the cohomology takes its values in the reals (it turns out that there can be cohomologies that take values in the integers, or in Z_2, but we will not consider them here.) The forms should also be nonsingular, in that integrating over them should not lead to any divergences.

As an example, let us consider the circle S_1. Since S_1 is one dimensional, we see that all 1 forms are closed. Suppose that the circle is parameterized by θ which runs from $0 \rightarrow 2\pi$. Then an exact 1-form can be written as $\boldsymbol{d}f(\theta)$, where $f(\theta)$ is periodic in θ. An example of a closed 1-form that is not exact is $\boldsymbol{d}\theta$. While $\boldsymbol{d}\theta$ is single valued around the circle and hence is an allowed form, θ is not single valued. In any event, any 1-form can be written as

$$\boldsymbol{\lambda} = f(\theta)\boldsymbol{d}\theta. \tag{4.12.5}$$

Since $f(\theta)$ is periodic and nonsingular, it must have the form

$$f(\theta) = c_0 + \sum_{n=1} [c_n \cos(n\theta) + b_n \sin(n\theta)] \tag{4.12.6}$$

where the coefficients can take any real values. But we note that

$$f(\theta)\boldsymbol{d}\theta = c_0 \boldsymbol{d}\theta + \sum_{n=1} \left[\frac{c_n}{n} \boldsymbol{d} \sin(n\theta) - \frac{b_n}{n} \boldsymbol{d} \cos(n\theta) \right] = c_0 \boldsymbol{d}\theta + \boldsymbol{d}(g(\theta)). \tag{4.12.7}$$

Hence any 1-form can be written as $c_0 \boldsymbol{d}\theta$ plus an exact form. Since c_0 is any real number, we see that $H^1(S_1, \mathbf{R}) = \mathbf{R}$.

Next consider $H^0(S_1, \mathbf{R})$. In this case no closed forms are exact, since we don't have -1 forms. The 0-forms are functions, and the closed forms are constants. Since the constants live in $\mathbf{R}$, we see that $H^0(S_1, \mathbf{R}) = \mathbf{R}$. In fact for any connected space $H^0(\mathcal{M}, \mathbf{R}) = \mathbf{R}$.

Let me now state something that I will not prove. This is the Poincaré Lemma, which states that all closed forms on a contractible space are exact. A space is contractible if it can be smoothly deformed to a point. The canonical way to contract a space is to consider maps from $\mathcal{M} \times [0,1] \to \mathcal{M}$, where $(x, t = 0) = x$, $(x, t = 1) = x_0$. If a C^∞ map exits then the flow from $t = 0$ to $t = 1$ describes $\mathcal{M}$ smoothly shrinking down to the point x_0. So for example, in 1 dimensions, the closed line segment $[-1, 1]$ is contractible, since there is the smooth map $[-1 + t, 1 - t]$ that shrinks the segment to the point $x = 0$. However, S_1 is not contractible. There is no smooth map that takes $[0, 2\pi]$ to say $[0, 0]$, because the end points of the interval have to differ by an integer times $[2\pi]$ and there is no way that an integer can smoothly jump to another integer.

A manifold $\mathcal{M}$ is said to be *simply connected* if every closed curve is contractible to a point. For example, the 2-sphere is simply connected since a loop going around the equator, can be smoothly deformed by pulling it into the northern hemisphere and then shrinking smoothly to a point at the north pole. Anyway, if the space is simply connected, then the integral of any closed 1-form over a closed loop is 0.

$$\oint \boldsymbol{\omega} = 0. \tag{4.12.8}$$

To see this, we note that a small deformation of a closed loop does not change the integral. This is because a small deformation can be written as

$$\oint \boldsymbol{\omega} + \oint_{loop} \boldsymbol{\omega} \tag{4.12.9}$$

where *loop* refers to a small contractible loop. (See figure). Since a closed form is locally

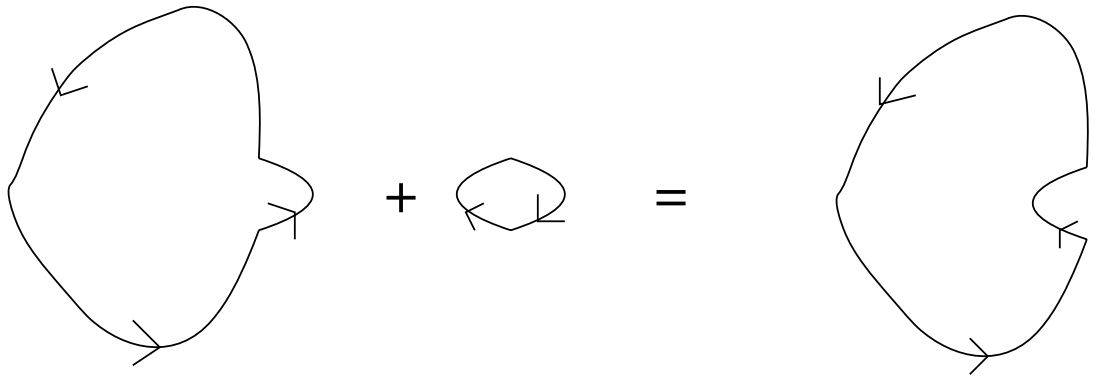


Figure 11: A loop is deformed by adding a small loop to it. Note the orientations of the loops.

exact, we can write the integral over the small loop as a total derivative. Since the loop is closed, the integral is zero. Hence, we can keep adding small deformations to the loop and not change the value of $\oint \boldsymbol{\omega}$. But since the loop is contractible, we can smoothly shrink it to a point by smoothly changing the curve by adding small loops. Since $\boldsymbol{\omega}$ is nonsingular, the integral of the loop when shrunken to a point is zero. (We can also use this argument to say that for any two noncontractible loops that can be smoothly deformed into one another, the integrals of $\boldsymbol{\omega}$ over the loops are the same.) Since the integral is zero, this means that the integral is conservative and so $\boldsymbol{\omega} = \boldsymbol{d\Lambda}$. Hence, for any simply connected space, $H^1(\mathcal{M}, \mathbf{R}) = 0$.

Let us consider the cohomology for the n dimensional sphere S_n. Since these are simply connected, we have that $H^1(S_n, \mathbf{R}) = 0$ if $n > 1$. We also have that $H^0(S_n, \mathbf{R}) = \mathbf{R}$ since the space is connected. Furthermore, one can show that $H^p(S_n, \mathbf{R}) = H^{p-1}(S_{n-1}, \mathbf{R})$. The rough proof of this is as follows. Consider a closed p form $\boldsymbol{\omega}$ on S_n. If we consider the two spaces A_1 and A_2, where A_1 has the south pole removed and A_2 has the north pole removed, then these spaces are contractible and so $\boldsymbol{\omega}$ is exact on each space. Let us say that $\boldsymbol{\omega} = \boldsymbol{d\Phi}_1$ on A_1 and $\boldsymbol{\omega} = \boldsymbol{d\Phi}_2$ on A_2. Now consider the intersecting region $A_1 \cap A_2$. This space is contractible to an S_{n-1} sphere. If $\boldsymbol{\omega}$ is not exact, then $\boldsymbol{\Phi} = \boldsymbol{\Phi}_1 - \boldsymbol{\Phi}_2 \neq 0$ on the intersecting region. But since $\boldsymbol{\omega}$ is equal on the two regions, we see that $\boldsymbol{\Phi}$ is a closed form on this region, which contracts to a closed form on S_{n-1}. We can also mod out by the exact forms in this region, since we are free to add exact forms to $\boldsymbol{\Phi}_1$ and $\boldsymbol{\Phi}_2$ and still satisfy $\boldsymbol{\omega} = \boldsymbol{d\Phi}_i$ in the two regions. Hence we see that the cohomology descends down. Given this, we see that for S_n, we have that $H^p(S_n, \mathbf{R}) = 0$ if $0 < p < n$ and that

$H^n(S_n, \mathbf{R}) = \mathbf{R}$. In general, for an n-dimensional orientable connected compact space, the surface itself is not contractible, so $H^n(\mathcal{M}, \mathbf{R}) = \mathbf{R}$. These sets of closed forms are often called the *volume forms*.

Roughly speaking, the cohomology $H^p(\mathcal{M}, \mathbf{R})$ counts the number of noncontractible p-dimensional surfaces in $\mathcal{M}$. So for example, the cohomology of the torus T^2, which is the product space of two circles is $H^0(T^2, \mathbf{R}) = \mathbf{R}$, $H^1(T^2, \mathbf{R}) = \mathbf{R} \oplus \mathbf{R}$ and $H^2(T^2, \mathbf{R}) = \mathbf{R}$. The torus has two noncontractible loops, hence the result for H^1. Since T^2 is compact and orientable, we have $H^2(T^2, \mathbf{R}) = \mathbf{R}$. Specifically, we can parameterize the torus by two angles, θ_1 and θ_2. Elements of $H^1(T^2, \mathbf{R})$ are $c_1 \boldsymbol{d}\theta_1 + c_2 \boldsymbol{d}\theta_2$ and the elements of $H^2(T^2, \mathbf{R})$ are $c \boldsymbol{d}\theta_1 \wedge \boldsymbol{d}\theta_2$.

4.13 Characteristic classes for principle bundles

We would now like to take what we learned for cohomology on manifolds and apply it to the study of principle bundles. In doing so, we will want to construct closed forms that are invariant under gauge transformations.

For a group G there there is a straightforward way of finding the invariants, although it is not always practical. Consider the determinant

$$\det(tI + c_a T^a), \tag{4.13.1}$$

where I is the identity matrix and t and c_a are ordinary coefficients. If the T^a are $m \times m$ matrices, then this determinant is given by

$$\det(tI + c_a T^a) = \sum_{i=0}^{m} t^i P_{m-i}(c_a), \tag{4.13.2}$$

where $P_{m-i}(c_a)$ is a polynomial of order $m - i$ in the coefficients c_a. Assume that the c_a are objects that transform covariantly, in other words, under a gauge transformation, $c_a T^a \to g c_a T^a g^{-1}$. Then the determinant transforms as

$$\det(tI + c_a T^a) \to \det(tI + g c_a T^a g^{-1}) = \det(g(tI + c_a T^a) g^{-1}) = \det(tI + c_a T^a). \tag{4.13.3}$$

Hence, the determinant is invariant. But since it is invariant for arbitrary t, it must be true that the individual polynomials in (4.13.2) are invariant.

Now instead of c_a, let us put in the curvature two forms $\boldsymbol{F} = F^a_{\mu\nu} dx^\mu \wedge dx^\nu T^a = \boldsymbol{F}^a T^a$. Under gauge transformations we have that $\boldsymbol{F}$ transforms covariantly, so the polynomials in (4.13.2) will be invariant. The lorentz indices of $F^a_{\mu\nu}$ do not effect the group transformation

properties, so we can take them anyway we please and the resulting polynomial will still be invariant. We therefore choose the polynomials to be

$$c_i(P) = P_i\left(\frac{\boldsymbol{F}}{2\pi}\right) \tag{4.13.4}$$

where the products of the terms in the P_i are assumed to be wedge products. The factor of 2π is put in for later covenience. If we choose the group G to be $SU(n)$, then the resulting polynomials $c_i(P)$ are called the *Chern classes* for the bundle P.

To evaluate these Chern classes, we note that the determinant of a matrix is the product of its eigenvalues. Suppose for a given $\boldsymbol{F}$, the eigenvalues are $\boldsymbol{f}_j$ with n eigenvalues for $G = SU(n)$. Hence the determinant satisfies

$$\det\left(tI + \frac{\boldsymbol{F}}{2\pi}\right) = \left(t + \frac{\boldsymbol{f}_1}{2\pi}\right) \wedge \left(t + \frac{\boldsymbol{f}_2}{2\pi}\right) \wedge ... \wedge \left(t + \frac{\boldsymbol{f}_n}{2\pi}\right) \tag{4.13.5}$$

Hence we have that

$$c_i(P) = \frac{1}{(2\pi)^i} \sum_{j_1<j_2..<j_i\leq n} \boldsymbol{f}_{j_1} \wedge \boldsymbol{f}_{j_2} \wedge ... \wedge \boldsymbol{f}_{j_i}. \tag{4.13.6}$$

Finally, we use the fact that the trace of a matrix is the sum of its eigenvalues, and that the trace of a matrix to a power m is the sum of the eigenvalues to the power m. We can then work through the individual values of i and find the characters. In the end, let me just state the formula:

$$c_i(P) = \frac{(-1)^i}{(2\pi)^i} \sum_{\sum j i_j = i} \prod_j (-1)^{i_j} \frac{1}{i_j! j^{i_j}} \left(\mathrm{Tr}\boldsymbol{F}^j\right)^{i_j} \tag{4.13.7}$$

where all products of $\boldsymbol{F}$ are assumed to be wedge products and where the sum means over all possible i_j such that

$$\sum_j j i_j = i. \tag{4.13.8}$$

Let me give the first few Chern classes. The lowest one has

$$c_0(P) = 1. \tag{4.13.9}$$

The next one, known as the *first Chern class* is

$$c_1(P) = \frac{1}{2\pi}\mathrm{Tr}\boldsymbol{F}. \tag{4.13.10}$$

For $SU(n)$, since the generators are traceless $c_1(P) = 0$. For $U(1)$, we have that $c_1(P) = \frac{1}{2\pi}\boldsymbol{F}$. The *second Chern class* is given by

$$c_2(P) = \frac{1}{4\pi^2}\left(\frac{1}{2}(\mathrm{Tr}\boldsymbol{F}) \wedge (\mathrm{Tr}\boldsymbol{F}) - \frac{1}{2}\mathrm{Tr}(\boldsymbol{F} \wedge \boldsymbol{F})\right) = \frac{1}{8\pi^2}\left((\mathrm{Tr}\boldsymbol{F}) \wedge (\mathrm{Tr}\boldsymbol{F}) - \mathrm{Tr}(\boldsymbol{F} \wedge \boldsymbol{F})\right) \tag{4.13.11}$$

For $U(1)$, it is clear that this Chern class is 0. For $SU(n)$, this reduces to

$$c_2(P) = -\frac{1}{8\pi^2}\mathrm{Tr}(\boldsymbol{F} \wedge \boldsymbol{F}). \tag{4.13.12}$$

The *third Chern class* is

$$c_3(P) = \frac{1}{8\pi^3}\left(\frac{1}{6}(\mathrm{Tr}\boldsymbol{F}) \wedge (\mathrm{Tr}\boldsymbol{F}) \wedge (\mathrm{Tr}\boldsymbol{F}) - \frac{1}{2}\mathrm{Tr}(\boldsymbol{F} \wedge \boldsymbol{F}) \wedge (\mathrm{Tr}\boldsymbol{F}) + \frac{1}{3}\mathrm{Tr}(\boldsymbol{F} \wedge \boldsymbol{F} \wedge \boldsymbol{F})\right). \tag{4.13.13}$$

For $U(1)$, it is simple to check that this is 0, as it should be. For $SU(2)$, it should also be 0. The nontrivial term to check is the $\mathrm{Tr}\boldsymbol{F} \wedge \boldsymbol{F} \wedge \boldsymbol{F}$ term, but it can be shown that this is zero based on the properties of the Pauli matrices.

The Chern classes have two important properties. The first one is that they are closed forms. To see this, we note that the $\boldsymbol{F}$ is proportional to $\boldsymbol{F} \sim \boldsymbol{D} \wedge \boldsymbol{D}$, where $\boldsymbol{D}$ is the covariant derivative form $\boldsymbol{D} = \boldsymbol{d} - i\boldsymbol{A}$. Note that before we used a commutator of covariant derivatives to find $\boldsymbol{F}$, but the commutator is essentially built into the wedge product, since

$$\boldsymbol{D} \wedge \boldsymbol{D} = D_\mu D_\nu dx^\mu \wedge dx^\nu = \frac{1}{2}[D_\mu, D_\nu]dx^\mu \wedge dx^\nu. \tag{4.13.14}$$

Therefore, we have

$$\boldsymbol{D} \wedge \boldsymbol{D} = -i(\boldsymbol{dA} - i\boldsymbol{A} \wedge \boldsymbol{A}) = -i\boldsymbol{F}. \tag{4.13.15}$$

This is the nonabelian version of the bianchi identity. Then the covariant derivative acting on $\boldsymbol{F}$ is

$$[\boldsymbol{D}, \boldsymbol{F}] = \boldsymbol{dF} - i[\boldsymbol{A}, \boldsymbol{F}] = (\boldsymbol{D}) \wedge i(\boldsymbol{D} \wedge \boldsymbol{D}) - i(\boldsymbol{D} \wedge \boldsymbol{D}) \wedge (\boldsymbol{D}) = 0. \tag{4.13.16}$$

Now consider the trace of the product of any number of $\boldsymbol{F}$. Then since the covariant derivative acting on $\boldsymbol{F}$ is 0, we have

$$\begin{aligned} 0 &= \mathrm{Tr}((\boldsymbol{dF} - i[\boldsymbol{A}, \boldsymbol{F}]) \wedge \boldsymbol{F} ... \wedge \boldsymbol{F}) + \mathrm{Tr}(\boldsymbol{F} \wedge (\boldsymbol{dF} - i[\boldsymbol{A}, \boldsymbol{F}]) \wedge ... \wedge \boldsymbol{F}) + ... \\ &\qquad + \mathrm{Tr}(\boldsymbol{F} \wedge \boldsymbol{F} \wedge ... \wedge (\boldsymbol{dF} - i[\boldsymbol{A}, \boldsymbol{F}])) \\ &= \mathrm{Tr}((\boldsymbol{dF}) \wedge \boldsymbol{F} ... \wedge \boldsymbol{F}) + \mathrm{Tr}(\boldsymbol{F} \wedge (\boldsymbol{dF}) \wedge ... \wedge \boldsymbol{F}) + ... \mathrm{Tr}(\boldsymbol{F} \wedge \boldsymbol{F} \wedge ... \wedge (\boldsymbol{dF})) \\ &= \boldsymbol{d}\ \mathrm{Tr}(\boldsymbol{F} \wedge \boldsymbol{F} \wedge ... \wedge \boldsymbol{F}), \end{aligned} \tag{4.13.17}$$

where we used the cyclic properties of the trace to eliminate the $[\boldsymbol{A}, \boldsymbol{F}]$ terms. Since the Chern classes are made up of products of traces, it automatically follows that they are closed.

Another important feature of the Chern classes is that the closed forms only change by an exact form when the gauge connection is varied. Let me stress that this is not a gauge transformation, but an actual change in the connection that can change the field strengths. To show this, suppose that I change the connection globally by a small amount $\boldsymbol{\epsilon}$. Then the change in $\boldsymbol{F}$ is $\boldsymbol{d\epsilon} - i(\boldsymbol{\epsilon} \wedge \boldsymbol{A} + \boldsymbol{A} \wedge \boldsymbol{\epsilon})$. Therefore the change in $\mathrm{Tr}\boldsymbol{F}^m$ to lowest order in $\boldsymbol{\epsilon}$ is

$$\begin{aligned}
\delta \mathrm{Tr}\boldsymbol{F}^m &= m\mathrm{Tr}((\boldsymbol{d\epsilon} - i(\boldsymbol{\epsilon} \wedge \boldsymbol{A} + \boldsymbol{A} \wedge \boldsymbol{\epsilon}) \wedge F^{m-1}) \\
&= m\boldsymbol{d}\ \mathrm{Tr}(\boldsymbol{\epsilon} \wedge F^{m-1}) + m\mathrm{Tr}(\boldsymbol{\epsilon} \wedge (\boldsymbol{d}(\boldsymbol{F}^{m-1}) - i[\boldsymbol{A}, \boldsymbol{F}^{m-1}])) \\
&= m\boldsymbol{d}\ \mathrm{Tr}(\boldsymbol{\epsilon} \wedge F^{m-1}) + m\mathrm{Tr}(\boldsymbol{\epsilon} \wedge \boldsymbol{D}(\boldsymbol{F}^{m-1})) \\
&= m\boldsymbol{d}\ \mathrm{Tr}(\boldsymbol{\epsilon} \wedge F^{m-1}). \qquad (4.13.18)
\end{aligned}$$

Therefore, we see that for an infinitesimal change, the trace changes by an exact form.

To see what happens for a finite change in $\boldsymbol{A}$, let us replace $\boldsymbol{\epsilon}$ by $t\boldsymbol{\eta}$, where t will vary from 0 to 1. The gauge fields as a function of t are then $\boldsymbol{A}(t) = \boldsymbol{A} + t\boldsymbol{\eta}$. For a small change from $\boldsymbol{A}(t)$ to $\boldsymbol{A}(t + \Delta t)$ then we have that

$$\delta \mathrm{Tr}(\boldsymbol{F}^m(t)) = m\boldsymbol{d}\ \mathrm{Tr}(\boldsymbol{\eta} \wedge F^{m-1}(t))\Delta t, \qquad (4.13.19)$$

where $\boldsymbol{F}$ is now t dependent. Hence we have that

$$\frac{d}{dt}\mathrm{Tr}(\boldsymbol{F}^m(t)) = m\boldsymbol{d}\ \mathrm{Tr}(\boldsymbol{\eta} \wedge F^{m-1}(t)) \qquad (4.13.20)$$

and so

$$\Delta \mathrm{Tr}(\boldsymbol{F}^m) = \boldsymbol{d}\ \left[m \int_0^1 dt \mathrm{Tr}(\boldsymbol{\eta} \wedge F^{m-1}(t))\right]. \qquad (4.13.21)$$

Since all traces differ by an exact form, then all products of traces also differ by exact forms. As an example consider the product

$$\begin{aligned}
\mathrm{Tr}\boldsymbol{F}^m \wedge \mathrm{Tr}\boldsymbol{F}^l \rightarrow \quad & (\mathrm{Tr}\boldsymbol{F}^m + \boldsymbol{d\Phi}_1) \wedge (\mathrm{Tr}\boldsymbol{F}^l + \boldsymbol{d\Phi}_2) \\
& = \mathrm{Tr}\boldsymbol{F}^m \wedge \mathrm{Tr}\boldsymbol{F}^l + \boldsymbol{d\Phi}_1 \wedge \mathrm{Tr}\boldsymbol{F}^l + \boldsymbol{d\Phi}_2 \wedge \mathrm{Tr}\boldsymbol{F}^m + \boldsymbol{d\Phi}_1 \wedge \boldsymbol{d\Phi}_2 \\
& = \mathrm{Tr}\boldsymbol{F}^m \wedge \mathrm{Tr}\boldsymbol{F}^l + \boldsymbol{d}(\boldsymbol{\Phi}_1 \wedge \mathrm{Tr}\boldsymbol{F}^l + \boldsymbol{\Phi}_2 \wedge \mathrm{Tr}\boldsymbol{F}^m + \boldsymbol{\Phi}_1 \wedge \boldsymbol{d\Phi}_2), \\
& \qquad (4.13.22)
\end{aligned}$$

where we used the fact that $\boldsymbol{d}^2 = 0$ and that the traces are closed.

4.14 The magnetic monopole (yet again), integer cohomology and the Hopf fibration

We can use the first Chern class to describe magnetic monopoles. Consider the

$$c_1(P) = \frac{1}{2\pi}\boldsymbol{F} \tag{4.14.1}$$

integrated over S_2. Locally, $\boldsymbol{F}$ is exact, so $\boldsymbol{F} = \boldsymbol{dA}_1$ in one hemisphere of the sphere and $\boldsymbol{F} = \boldsymbol{dA}_2$ on the other hemisphere. The two hemispheres intersect along the S_1 at the equator. Therefore

$$\frac{1}{2\pi}\int_{S_2}\boldsymbol{F} = \frac{1}{2\pi}\int_{S_1}\boldsymbol{A}_1 - \frac{1}{2\pi}\int_{S_1}\boldsymbol{A}_2 = \frac{1}{2\pi}\int_{S_1}\boldsymbol{d\Phi} = n, \tag{4.14.2}$$

where n is an integer. The fact that we can have nonzero values for n is a consequence of the nontrivial cohomology of $H^2(S_2, \mathbf{R})$.

Since the quantities we are computing turn out to be integers, it is often convenient to consider the *integer cohomology* as opposed to the real cohomology. They are essentially the same thing. The only difference is that for the integer cohomology, we assume that the coefficients are integers. What we would say is that $H^2(S_2, \mathbf{Z}) = \mathbf{Z}$. The advantage of this is that often times we are integrating things that do have coefficients in the integers. In this case, it is the closed form $\frac{1}{2\pi}\boldsymbol{F}$. The integer coefficient is n.

If n is nonzero, then we have a nontrivial fiber bundle. Let me make this explicit, with the example of $n = 1$. In our case, the base space is S_2 and the fiber is the $U(1)$ group. A group element is given by $e^{i\phi}$ where $\phi \equiv \phi + 2\pi$. Hence, the fiber space is essentially a circle S_1. If the principle bundle P is trivial then it is equivalent to a product space $P \simeq S_2 \times S_1$.

Let us express the sphere in terms of $CP(1)$, that is we have two complex numbers and an identification $(z_1, z_2) \equiv (\lambda z_1, \lambda z_2)$, where λ is any nonzero complex number. Let us now include the fiber which is given by ϕ. Region 1 contains the point $(1, 0)$ and region 2 contains the point $(0, 1)$. Since z_1 is never zero in region 1 and z_2 is never zero in region 2, let us attempt to identify the fiber to be the phase of z_1, θ_1, in region 1 and the phase of z_2, θ_2, in region 2. In the overlap region, both phases are well defined, since neither z_1 nor z_2 are zero here. In the overlap region, from (4.14.2), we see that $\theta_1 = \theta_2 + n\theta$ where θ is the angle around the equator. But the angle around the equator is given by $\arg(z_1/z_2)$, hence we can make this identification of fibers if $n = 1$. The whole bundle is therefore described by the space $(z_1, z_2) \equiv (\rho z_1, \rho z_2)$ where ρ is positive real. Note that the identification under the phases is gone, since a change in the phase is considered a

change along the fiber. By rescaling, we can always reexpress our space as (z_1, z_2) with the constraint that

$$z_1 z_1^* + z_2 z_2^* = 1. \tag{4.14.3}$$

But this is S_3! Hence the principle bundle corresponding to a single monopole is topologically the same as the three sphere. This is clearly not the same as $S_2 \times S_1$, for one thing the cohomology is different. This fibration is known as the Hopf fibration.

4.15 Instantons and the second chern class

In this final section, we consider an important phenomenon in gauge theories called instantons. Instantons are essentially nontrivial solutions to the equations of motion in *Euclidean space*. We will consider such solutions on S_4, the 4 sphere. S_4 is almost topologically the same as $\mathbf{R}^4$, but not quite. What S_4 is equivalent to is $\mathbf{R}^4 \cup \{\infty\}$, that is $\mathbf{R}^4$ and the point at infinity. However, if we restrict our forms on $\mathbf{R}^4$ to be well behaved at infinity, then this is essentially the same as considering the space to be S_4.

Let us consider the second Class for $SU(2)$, which is given by

$$c_2 = -\frac{1}{8\pi^2}\mathrm{Tr}(\boldsymbol{F} \wedge \boldsymbol{F}). \tag{4.15.1}$$

As we have already seen this is a closed form. If we integrate this over S_4, then the result is a topological invariant.

Since c_2 is closed, it is locally exact. Hence we can break up S_4 into two regions that intersect at the equator which is an S_3. Just as in the monopole case, we have that the integral of the Chern class is

$$\int_{S_4} c_2 = \int_{S_3} \boldsymbol{\Lambda} \tag{4.15.2}$$

where $\boldsymbol{\Lambda}$ is a closed 3 form. What is this 3-form? Just as in the monopole case, it is basically a "pure gauge" piece.

To see this, let us look more closely at the $\mathrm{Tr}\boldsymbol{F} \wedge \boldsymbol{F}$ piece. If we expand this in terms of the $\boldsymbol{A}$ field as shown in (4.11.15), we find

$$\mathrm{Tr}(\boldsymbol{F}\boldsymbol{F}) = \mathrm{Tr}(\boldsymbol{dA} \wedge \boldsymbol{dA} - 2i\boldsymbol{dA} \wedge \boldsymbol{A} \wedge \boldsymbol{A} - \boldsymbol{A} \wedge \boldsymbol{A} \wedge \boldsymbol{A} \wedge \boldsymbol{A}). \tag{4.15.3}$$

We next note that the last term in the parentheses is zero. To see this, we use the antisymmetry properties of the wedge product and the cyclic property of the trace. You can readily verify by expanding out the matrices that

$$\mathrm{Tr}(\boldsymbol{B}_p \wedge \boldsymbol{B}_q) = (-1)^{pq}\mathrm{Tr}(\boldsymbol{B}_q \wedge \boldsymbol{B}_p), \tag{4.15.4}$$

where $\boldsymbol{B}_p$ is a matrix valued p-form and $\boldsymbol{B}_q$ is a matrix valued q-form. Hence we have that

$$\mathrm{Tr}([\boldsymbol{A}\wedge\boldsymbol{A}\wedge\boldsymbol{A}]\wedge\boldsymbol{A}) = -\mathrm{Tr}(\boldsymbol{A}\wedge[\boldsymbol{A}\wedge\boldsymbol{A}\wedge\boldsymbol{A}]) = -\mathrm{Tr}(\boldsymbol{A}\wedge\boldsymbol{A}\wedge\boldsymbol{A}\wedge\boldsymbol{A}) = 0. \quad (4.15.5)$$

We can then show that the remaining terms are

$$\mathrm{Tr}(\boldsymbol{dA}\wedge\boldsymbol{dA} - 2i\boldsymbol{dA}\wedge\boldsymbol{A}\wedge\boldsymbol{A}) = \boldsymbol{d}\left[\mathrm{Tr}\left(\boldsymbol{F}\wedge\boldsymbol{A} + \frac{i}{3}\boldsymbol{A}\wedge\boldsymbol{A}\wedge\boldsymbol{A}\right)\right]. \quad (4.15.6)$$

To verify this last statement, we note that

$$\begin{aligned}
&\boldsymbol{d}\left[\mathrm{Tr}\left(\boldsymbol{F}\wedge\boldsymbol{A} + \frac{i}{3}\boldsymbol{A}\wedge\boldsymbol{A}\wedge\boldsymbol{A}\right)\right] \\
&\qquad = \mathrm{Tr}\left(\boldsymbol{dF}\wedge\boldsymbol{A} + \boldsymbol{F}\wedge\boldsymbol{dA} + \frac{i}{3}\left(\boldsymbol{dA}\wedge\boldsymbol{A}\wedge\boldsymbol{A} - \boldsymbol{A}\wedge\boldsymbol{dA}\wedge\boldsymbol{A} + \boldsymbol{A}\wedge\boldsymbol{A}\wedge\boldsymbol{dA}\right)\right) \\
&\qquad = \mathrm{Tr}\left(i[\boldsymbol{A},\boldsymbol{F}]\wedge\boldsymbol{A} + \boldsymbol{F}\wedge\boldsymbol{dA} + i\boldsymbol{dA}\wedge\boldsymbol{A}\wedge\boldsymbol{A}\right) \\
&\qquad = \mathrm{Tr}(\boldsymbol{dA}\wedge\boldsymbol{dA} - 2i\boldsymbol{dA}\wedge\boldsymbol{A}\wedge\boldsymbol{A}), \qquad (4.15.7)
\end{aligned}$$

where in going from the second line to the third line, we used the Bianchi identity in (4.13.15). Hence, in the two hemispheres of S_4, we have that

$$c_2 = \boldsymbol{d\Phi} \qquad \boldsymbol{\Phi} = -\frac{1}{8\pi^2}\mathrm{Tr}\left(\boldsymbol{F}\wedge\boldsymbol{A} + \frac{i}{3}\boldsymbol{A}\wedge\boldsymbol{A}\wedge\boldsymbol{A}\right). \quad (4.15.8)$$

Next we note that under a gauge transformation, $\boldsymbol{A}\to U\boldsymbol{A}U^\dagger - i\boldsymbol{d}UU^\dagger$, $\boldsymbol{\Phi}$ transforms as

$$\begin{aligned}
\boldsymbol{\Phi} \quad &\to \boldsymbol{\Phi} - \frac{1}{8\pi^2}\Big[\mathrm{Tr}(\boldsymbol{F}\wedge(-i\boldsymbol{d}UU^\dagger)) + \frac{i}{3}3\mathrm{Tr}(-iU^\dagger\boldsymbol{d}U\wedge\boldsymbol{A}\wedge\boldsymbol{A}) \\
&\qquad -\frac{i}{3}3\mathrm{Tr}(U^\dagger\boldsymbol{d}U\wedge U^\dagger\boldsymbol{d}U\wedge\boldsymbol{A}) - \frac{1}{3}\mathrm{Tr}(U^\dagger\boldsymbol{d}U\wedge U^\dagger\boldsymbol{d}U\wedge U^\dagger\boldsymbol{d}U)\Big] \\
&= \boldsymbol{\Phi} - \frac{1}{8\pi^2}\left[-i\boldsymbol{d}(\mathrm{Tr}(\boldsymbol{F}\wedge(\boldsymbol{d}UU^\dagger))) - \frac{1}{3}\mathrm{Tr}(U^\dagger\boldsymbol{d}U\wedge U^\dagger\boldsymbol{d}U\wedge U^\dagger\boldsymbol{d}U)\right]. \quad (4.15.9)
\end{aligned}$$

Therefore, we find that on the S_3 equator, if the two gauge fields from the northern and southern hemispheres differ by a gauge transformation, then

$$\int_{S_4} c_2 = \frac{1}{8\pi^2}\int_{S_3}\frac{1}{3}\mathrm{Tr}(U^\dagger\boldsymbol{d}U\wedge U^\dagger\boldsymbol{d}U\wedge U^\dagger\boldsymbol{d}U), \quad (4.15.10)$$

where we have thrown away the total derivative term.

To find this integral, let us recall that the group manifold of $SU(2)$ is actually the 3-sphere. In other words, a general element U is given by

$$U = b_0 + b_1\sigma_1 + b_2\sigma_2 + b_3\sigma_3, \qquad b_0^2 + b_1^2 + b_2^2 + b_3^2 = 1. \quad (4.15.11)$$

Since the group manifold is S_3, there exists a one to one and onto map from the physical S_3 to the S_3 of the group manifold. For such a map, the trace term in (4.15.10) is 12 times the volume form of S_3. The factor of 12 has a factor of 3! from the combinatorics in the wedge product and a factor of 2 from the two dimensional trace. For this particular map, we then have that

$$\int_{S_4} c_2 = \frac{1}{2\pi^2} \int_{S_3} \boldsymbol{\omega}_3 = 1 \tag{4.15.12}$$

where $\boldsymbol{\omega}_3$ is the volume 3-form of S_3. A more general class of maps would wind the S_3 n times around the group manifold. In this case

$$\int_{S_4} c_2 = n \qquad\qquad n \in \mathbf{Z}. \tag{4.15.13}$$

If the second Chern class is nonzero, then the bundle over S_4 is nontrivial. In the case where $n = \pm 1$, the bundle is a Hopf fibration which is topologically the same as S_7.

Why should we care about these objects? The reason is that we can use them to find nontrivial minima for gauge field configurations. In doing physical computations, it is often convenient to rotate the time variable t to be imaginary, in other words t satisfies $t = i\tau$, where τ is real. The effect this has on the action that appears in the path integral is that

$$S = \int dt L \to i \int \tau L = i S_E. \tag{4.15.14}$$

S_E is called the *Euclidean action*. This is because the normal minkowski metric is rotated to a Euclidean metric:

$$ds^2 = -dt^2 + d\vec{x}^2 \to d\tau^2 + d\vec{x}^2. \tag{4.15.15}$$

Under the rotation, the phase factor in the path integral becomes

$$e^{iS} \to e^{-S_E} \tag{4.15.16}$$

Hence for the Euclidean action, we want to find field configurations that minimize S_E.

Let us consider the Euclidean action for a gauge field,

$$S_E = \frac{1}{4k} \int d^4x \mathrm{Tr}(F_{\mu\nu} F^{\mu\nu}), \tag{4.15.17}$$

where the integral is over $\mathbf{R}^4$. The traces are normalized so that $k = 1/2$. It is now convenient to express this using something called the *dual form*. On $\mathbf{R}^4$, the dual form is constructed by taking the ϵ tensor and writing

$$\widetilde{F}_{\mu\nu} = \frac{1}{2} \epsilon_{\mu\nu\lambda\rho} F^{\lambda\rho}. \tag{4.15.18}$$

$\widetilde{F}_{\mu\nu}$ is antisymmetric in its two indices and from it we can write a two tensor, which is usually written as $*\boldsymbol{F}$,

$$*\boldsymbol{F} = \frac{1}{2}\widetilde{F}_{\mu\nu}\boldsymbol{d}x^{\mu}\boldsymbol{d}x^{\nu}. \tag{4.15.19}$$

Then it is not too hard to see that the Euclidean action is

$$S_E = \int \mathrm{Tr}\boldsymbol{F} \wedge *\boldsymbol{F}. \tag{4.15.20}$$

We also note that

$$\mathrm{Tr}\boldsymbol{F} \wedge \boldsymbol{F} = \mathrm{Tr} * \boldsymbol{F} \wedge *\boldsymbol{F}. \tag{4.15.21}$$

Hence we have that

$$\mathrm{Tr}(\boldsymbol{F} \wedge *\boldsymbol{F}) = \frac{1}{2}\mathrm{Tr}([\boldsymbol{F} \pm *\boldsymbol{F}] \wedge [*\boldsymbol{F} \pm \boldsymbol{F}]) \mp \mathrm{Tr}(\boldsymbol{F} \wedge \boldsymbol{F}). \tag{4.15.22}$$

Now the first term on the rhs of (4.15.22) is positive definite and the second term when integrated over $\mathbf{R}^4$ is a topological invariant. Hence we have a local minimimum of S_E in terms of the field configurations if we set the first term on the rhs to zero. This means that

$$\boldsymbol{F} = \pm * \boldsymbol{F} \tag{4.15.23}$$

These configurations are localized in the $\mathbf{R}^4$ in the sense that $\boldsymbol{F}$ falls to zero as we approach infinity. These configurations can be thought of solutions in space-time, but where the time component is also space-like. Since the solution is also localized in euclidean time, they are call "instantons". Actually the solution with the plus sign in (4.15.23) is called an instanton and the one with a minus sign is called an anti-instanton. We can easily see what S_E is at the minima using (4.15.22) and (4.15.13), namely it is

$$S_E = 8\pi^2|n| \tag{4.15.24}$$

The absolute value arises because the lhs of (4.15.22) is positive definite. Hence the instantons can only have positive n and the anti-instantons can only have negative n. Hence we see that for $n \neq 0$, the instantons are nontrivial configurations since the action is nonzero.