How to Talk to a Physicist: Groups, Symmetry, and Topology

Daniel T. Larson

August 4, 2005

Contents

Preface

Metacognition: Thinking about how we think

I recently finished reading a book[1] about teaching and learning that has really inspired me to try to improve my ability to help other students learn. Since I am in the midst of leading this tutorial, you folks are destined to be the guinea pigs for my initial attempts at being a more effective teacher. I apologize in advance for all shortcomings. Please bear with me.

My idea for the first part of the tutorial was based on the frustration I felt as a student trying to piece together an understanding of Lie groups, Lie algebras, and representations in the context of quantum mechanics and particle physics. There are many books on all of these subjects, but it has been hard to assemble the pieces into a succinct packet of understanding. In short, I have struggled to "cut to the chase". Writing up these lecture notes has been a way to collect what I've learned in one place and, honestly, to clarify my understanding of many of the connections. What I hope to do in this tutorial is to encourage you to venture onto this path, provide some guidance past the obstacles I have overcome, and help point the way ahead to deeper insights.

My plan for the second half of the tutorial was initially to show how higher homotopy groups arise in quantum field theory, indicating the existence of anomalies or non-perturbative solutions. I'm a little concerned that this might require asking you to accept too much on faith (even the previous sentence has lots of technical terminology) and not provide a deep enough learning experience to be satisfying. Stimulated by a comment from one of you, I am now rethinking slightly my plan for the second half of the tutorial. (I confess that I haven't started writing up the notes for that yet!) My current

[1] Bain, Ken, "What the best college teachers do," Harvard University Press, 2004.

plan is to focus on the issue of local symmetry, because this is a subtle issue that my fellow Berkeley graduate students and I wrestled with for about a week and still didn't come to a definitive conclusion. Local symmetry does lead naturally to the study of topology so we should be able to bring in homotopy groups as I had originally intended, fleshing out the mathematical side of the tutorial. As this is a work in progress, however, you can influence the outcome by making your personal desires known.

You must be thinking, "Enough of the meta mumbo jumbo already." So I will quickly conclude with the my goals and promises for the tutorial. In this tutorial I will help you:

1. increase your amazement at the way beautiful, abstract mathematics appears in understanding actual experiments in physics,

2. gain a new and unique perspective on the sophisticated mathematical subjects of group theory and topology,

3. prove to yourself that you know what a Lie algebra is and what a local symmetry represents, and be able to explain to anyone who asks the distinction between $SO(3)$ and $SU(2)$,

4. cultivate a desire to study groups, quantum mechanics, topology, or field theory in more detail.

I can help with these goals, but to succeed you will also need to be committed to these goals yourself. Are you?

Now, on with the subject matter.

Chapter 1

Global Symmetries

1.1 Groups

We start with a set of elements G together with a rule (called multiplication) for combining two elements to get a third. The name of the resulting algebraic structure depends on the properties of the multiplication law. If the multiplication is associative, $(g_1 g_2)g_3 = g_1(g_2 g_3) = g_1 g_2 g_3$, then G is called a **semigroup**. If we also require an identity element $1 \in G$ such that $1 \cdot g = g = g \cdot 1 \ \forall g \in G$, then G becomes a **monoid**. Finally, the existence of inverse elements g^{-1} for every element such that $gg^{-1} = 1 = g^{-1}g \ \forall g \in G$ makes G into a **group**. The group multiplication law is not necessarily commutative, but if it is then the group is said to be **Abelian**. Because groups are closely related to symmetries, and symmetries are very useful in physics, groups have come to play an important role in modern physics.

Groups can be defined as purely abstract algebraic objects. For example, the group D_3 is generated by the two elements x and y with the relations $x^3 = 1$, $y^2 = 1$, and $yx = x^{-1}y$. We can systematically list all the elements: $D_3 = \{1, x, x^2, y, xy, x^2y\}$. We should check that this list is exhaustive, namely that all inverses and any combination of x and y appears in the list. Using the relations we see that $x^{-1} = x^2$ and $y^{-1} = y$, so $yx = x^{-1}y = x^2y$, all of which are already in the list. What about yx^2? The **order** of a group is the number of elements it contains and is sometimes written $|G|$. We see that D_3 is a group of order 6, i.e. $|D_3| = 6$. This is a specific example of a **dihedral group** D_n which is defined in general as the group generated by $\{x, y\}$ with the relations $x^n = 1$, $y^2 = 1$, and $yx = x^{-1}y$. We can systematically list the

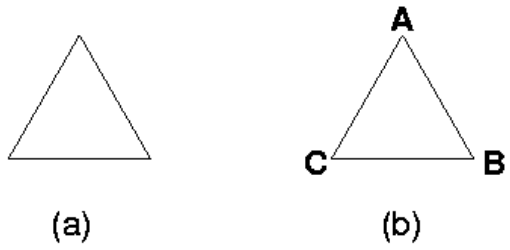

Figure 1.1: (a) An equilateral triangle; (b) the same triangle with labels.

elements: $D_n = \{1, x, x^2, \ldots, x^{n-1}, y, xy, x^2y, \ldots, x^{n-1}y\}$. You should prove that D_n has order $2n$.

Exercise 1 *The symmetric group S_n is the set of permutations of the integers $\{1, 2, \ldots, n\}$. Is S_n Abelian? Determine the order of S_n (feel free to start with simple cases of $n = 1, 2, 3, 4$). Note that $|S_3| = |D_3|$. Are they the same group (isomorphic)?*

1.1.1 Realizations and Representations

The previous example showed how a group can be defined in the abstract. However, groups often appear in the context of physical situations. Consider the rigid motions (preserving lengths and angles) in the plane that leave an equilateral triangle (shown in Figure 1.1(a)) unchanged. To keep track of what's going on, we will need to label the triangle as shown in Figure 1.1(b).

One rigid motion is a rotation about the center by an angle $\frac{2\pi}{3}$ in the counter-clockwise direction as shown in Figure 1.2(a). Let's call such a transformation R for "rotation". We say that R is a **symmetry** of the triangle because the triangle is unchanged after the action of R. Another symmetry is a reflection about the vertical axis, which we can call F for "flip". This is shown in Figure 1.2(b).

Both R and F can be inverted, by a clockwise rotation or another flip, respectively. Clearly combinations of Rs and Fs are also symmetries. For example, Figure 1.3(a) shows the combined transformation FR in the top panel, whereas the lower panel (b) shows the transformation R^2F. Note that our convention is that the right-most operation is done first.

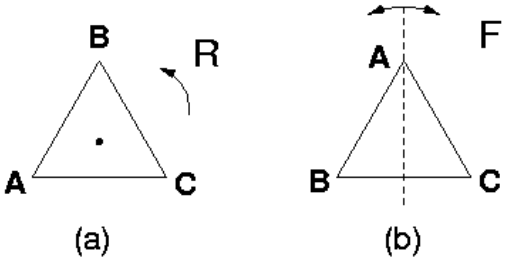

Figure 1.2: (a) R, a counter clockwise rotation by $\frac{2\pi}{3}$; (b) and the "flip" F, a reflection about the vertical axis (b).

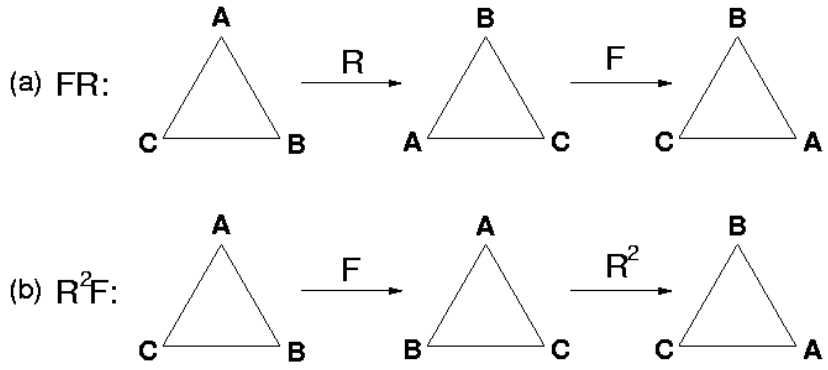

Figure 1.3: (a) The combined transformation FR; (b) the transformation R^2F.

These examples show that transformations on a space can naturally form a group. More formally, a group **realization** is a map from elements of G to transformations of a space M that is a **group homomorphism**, i.e. it preserves the group multiplication law. Thus if $T : G \to T(M) : g \mapsto T(g)$ where $T(g)$ is some transformation on M, then T is a group homomorphism if $T(g_1 g_2) = T(g_1)T(g_2)$. From this you can deduce that $T(1) = I$ where I is the identity ("do nothing") transformation on M, and $T(g^{-1}) = (T(g))^{-1} = T^{-1}(g)$.

Let's go back to the symmetries of our triangle. You probably noticed that $R^3 = I$ and $F^2 = I$, which bears striking similarities to D_3. In fact, with $T(x) = R$ and $T(y) = F$, it isn't hard to prove that T is a realization of D_3. One important thing to check is whether $T(yx) = T(y)T(x) = FR$ and $T(x^{-1}y) = T^{-1}(x)T(y) = R^{-1}F = R^2F$ are the same. But this is exactly what we showed in Figure 1.3.

In this example the map $T : D_3 \to \text{Symmetries}(\triangle)$ is bijective (one-to-one and onto) so in addition to being a homomorphism it is also an **isomorphism**. Such realizations are often called **faithful** because every different group element gets assigned to a different transformation. However, realizations do not need to be faithful. Consider the homomorphism $T' : D_3 \to \text{Symmetries}(\triangle)$ where $T'(x) = I$ and $T'(y) = F$. Then $T'(yx) = T'(y)T'(x) = FI = F$ and $T'(x^{-1}y) = T'(x^2y) = I^2F = F$. Thus the group relations and multiplication still hold so we have a realization, but it is definitely not an isomorphism.

Exercise 2 *Consider the map $T'' : D_3 \to \text{Symmetries}(\triangle)$ where $T''(x) = R$ and $T''(y) = I$. Is T'' is a realization?*

Exercise 3 *A **normal subgroup** $H \subset G$ is one where $ghg^{-1} \in H$ for all $g \in G$ and $h \in H$. Both T' and T'' map a different subgroup of D_3 to the identity. Are those normal subgroups? In this context, why is the concept of normal subgroup useful?*

Physicists are usually interested in the special class of realizations where M is a vector space and the $T(g)$ are linear transformations. Such realizations are called **representations**.[1] A **vector space** V is an Abelian group with

[1] A warning about terminology: Technically the representation is defined as the map (homomorphism) between G and transformations on a vector space. However, often the term "representation" is used to refer to the vector space on which the elements $T(g)$ act, and sometimes even to the linear transformations $T(g)$ themselves.

elements $|v\rangle$ called vectors and a group operation "+", and it possesses a second composition rule with scalars α which are elements of a field $\mathbb{F}$ (like $\mathbb{R}$ or $\mathbb{C}$) such that $\alpha|v\rangle \in V$ for all $\alpha \in \mathbb{F}$ and all $|v\rangle \in V$. Further we have the properties:

i) $(\alpha\beta)|v\rangle = \alpha(\beta|v\rangle)$

ii) $1 \in \mathbb{F}$ is an identity: $1|v\rangle = |v\rangle$

iii) $(\alpha + \beta)|v\rangle = \alpha|v\rangle + \beta|v\rangle$ and $\alpha(|v\rangle + |w\rangle) = \alpha|v\rangle + \alpha|w\rangle$.

A **linear transformation** on a vector space V is a map $T : V \to V$ such that $T(\alpha|v\rangle + \beta|w\rangle) = \alpha T(|v\rangle) + \beta T(|w\rangle)$. Linear algebra is essentially the study of vector spaces and linear transformations between them. This is important for us because soon we will see how quantum mechanics essentially boils down to linear algebra. Thus the study of symmetry groups in quantum mechanics becomes the study of group representations. But first we should look a some simple examples of representations.

Consider the parity group $P = \{x : x^2 = 1\}$, also known as $\mathbb{Z}_2$, the cyclic group of order 2. The most logical representation of P is by transformations on $\mathbb{R}$ where $T(x) = -1$. Clearly $T(x^2) = T(x)T(x) = (-1)^2 = 1$ so this forms a faithful representation. We could also study the **trivial representation** where $T'(x) = 1$. Again, $T'(x^2) = T'(x)^2 = 1^2 = 1$, so the group law is preserved, but nothing much happens with this representation, so it lives up to its name.

The **dimension** of a representation refers to the dimension of the vector space V on which the linear transformations $T(g)$ act, not to be confused with the order of the group. Let's now consider a two-dimensional representation of P. Take $\mathbb{R}^2$ with basis $|m\rangle, |n\rangle$ and let $T_2(x)|m\rangle = |n\rangle$ and $T_2(x)|n\rangle = |m\rangle$. You can check that $T_2(x^2) = (T_2(x))^2 = I$ because it takes $|m\rangle \to |n\rangle \to |m\rangle$ and $|n\rangle \to |m\rangle \to |n\rangle$. In terms of matrices we have:

$$T_2(x) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \text{and} \qquad T_2(1) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{1.1}$$

and you can check that $T_2(x)^2 = I$ in terms of matrices as well.

Something interesting happens if we change the basis of the vector space

to $|\mu\rangle = \frac{1}{\sqrt{2}}(|m\rangle + |n\rangle)$ and $|\nu\rangle = \frac{1}{\sqrt{2}}(-|m\rangle + |n\rangle)$. Then

$$\begin{aligned} T_2(x)|\mu\rangle &= \frac{1}{\sqrt{2}}(T_2(x)|m\rangle + T_2(x)|n\rangle) = \frac{1}{\sqrt{2}}(|n\rangle + |m\rangle) = |\mu\rangle \\ T_2(x)|\nu\rangle &= \frac{1}{\sqrt{2}}(-T_2(x)|m\rangle + T_2(x)|n\rangle) = \frac{1}{\sqrt{2}}(-|n\rangle + |m\rangle) = -|\nu\rangle \end{aligned}$$

This new basis yields a new representation of P, call it T_2'. The matrices corresponding to the $|\mu\rangle, |\nu\rangle$ basis are:

$$T_2'(x) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \qquad \text{and} \qquad T_2'(1) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{1.2}$$

Of course, since T_2 and T_2' are related by a change of basis their matrices are related by a similarity transformation, $T_2(g) = ST_2'(g)S^{-1}$ and they aren't really different in any substantial way. Representations related by a change of basis are called **equivalent representations**.

Another thing to notice about T_2' is that all the representatives (i.e. both matrices) are diagonal. This means that the two basis vectors $|\mu\rangle$ and $|\nu\rangle$ are acted upon independently, so they each can be considered as separate one-dimensional representations. In fact, the representation on $|\mu\rangle$ is none other than the trivial representation, T', and the representation on $|\nu\rangle$ is the same as our faithful representation T above. A representation that can be separated into representations with smaller dimensions is said to be **reducible**. In general, a representation will be reducible when all of its matrices can be simultaneously put into the same block diagonal form:

$$T_{n+m+p}(g) = \left(\begin{array}{c|c|c} T_1(g) & & \\ \hline & T_2(g) & \\ \hline & & T_3(g) \end{array}\right) \begin{array}{l} \}_{n\times n} \\ \}_{m\times m} \\ \}_{p\times p} \end{array}. \tag{1.3}$$

Then each of the smaller subspaces are acted on by a single block and give a representation of the group G of lower dimension. Because larger representations can be built up out of smaller ones, it is sensible to try to classify the **irreducible** representations of a given group.

Exercise 4 *Can you construct an irreducible 2-dimensional representation of the parity group P? Can you construct a nontrivial 3-dimensional representation of P? If you can, is it irreducible?*

1.2 Lie Groups and Lie Algebras

Now we come to our main example, that of rotations. In $\mathbb{R}^3$ rotations about the origin preserve lengths and angles so they can be represented by 3×3 (real) orthogonal matrices. If we further specify that the determinant be equal to one (avoiding inversions) then we have the **special orthogonal group** called $SO(3)$:

$$SO(3) = \{\mathcal{O} \in 3 \times 3 \text{ matrices} : \mathcal{O}^\dagger = \mathcal{O}^{-1} \text{ and } \det \mathcal{O} = 1\}. \tag{1.4}$$

For real matrices like we have here, the Hermitian conjugate, $\mathcal{O}^\dagger = (\mathcal{O}^*)^T$ is equal to the transpose, $\mathcal{O}^T$, so I've chosen to use the former for later convenience. We can verify that this is a group by checking $(\mathcal{O}_1\mathcal{O}_2)^\dagger = \mathcal{O}_2^\dagger\mathcal{O}_1^\dagger = \mathcal{O}_2^{-1}\mathcal{O}_1^{-1} = (\mathcal{O}_1\mathcal{O}_2)^{-1}$ and $\det(\mathcal{O}_1\mathcal{O}_2) = \det\mathcal{O}_1 \det\mathcal{O}_2 = 1 \cdot 1 = 1$. Here our definition of the group is made in terms of a faithful representation. $SO(3)$ has two important features: it is a continuous group and it is *not* commutative. We will discuss these properties in turn.

Exercise 5 *Prove that orthogonal transformations on $\mathbb{R}^3$ do indeed preserve lengths and angles.*

Exercise 6 *Does the set of general $n \times n$ matrices form a group? If so, prove it. If not, can you add an additional condition to make it into a group.*

1.2.1 Continuous Groups

Rotations about the z-axis are elements (in fact, a subgroup) of $SO(3)$. Since we can imagine rotating by any angle between 0 and 2π it is clear that there are an infinite number of rotations about the z-axis and hence an infinite number of elements in $SO(3)$. We can write the matrix of such a rotation as

$$T_z(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{1.5}$$

demonstrating how such a rotation can be parameterized by a continuous real variable. It turns out that you need 3 continuous parameters to uniquely specify every element in $SO(3)$. (These three can be thought of as rotations about the 3 axes or the three Euler angles.) Because of the continuous parameters we have some notion of group elements being close together.

Such groups have the additional structure of a manifold and are called Lie groups.

A **manifold** is a set of point M together with a notion of open sets (which makes M a topological space) such that each point $p \in M$ is contained in an open set U that has a continuous bijection $\varphi : U \to \varphi(U) \subset \mathbb{R}^n$. Thus an n-dimensional manifold is a space where every small region looks like $\mathbb{R}^n$. If all the functions φ are differentiable then you have a differentiable manifold.[2]

A **Lie group** is a group that is also a differentiable manifold.

Our main example of $SO(3)$ is in fact a Lie group. We already know how to deal with its group properties since matrix multiplication reproduces the group multiplication law. But what kind of manifold is $SO(3)$?

We will begin to answer this question algebraically. More generally, $SO(n)$ consists of $n \times n$ real matrices with $\mathcal{O}^\dagger\mathcal{O} = I$ and $\det \mathcal{O} = 1$. A general $n \times n$ real matrix has n^2 entries so is determined by n^2 real parameters. But the orthogonality condition gives $\frac{n(n+1)}{2}$ constraints (because the condition is symmetric, so constraining the upper triangle of the matrix automatically fixes the lower triangle). Since any orthogonal matrix must have $\det \mathcal{O} = \pm 1$, the constraint for a positive determinant only eliminates half of the possibilities but doesn't reduce the number of continuous parameters. Thus a matrix in $SO(n)$ will be specified by $n^2 - \frac{n(n+1)}{2} = \frac{n(n-1)}{2}$. Thus $SO(3)$ is specified by 3 parameters and is therefore a 3-dimensional manifold.

Knowing the dimension is a start, but we can learn more by using geometric reasoning. Let's specify a rotation about an axis by a vector along that axis with length in the range $[0, \pi]$ corresponding to the counter-clockwise angle of rotation about that axis. The collection of all such vectors is a solid, closed ball of radius π in $\mathbb{R}^3$, call it D^3 (for "disk). However, a rotation by π about some axis $\vec{n}$ is the same as a rotation by π about $-\vec{n}$. So to take this into account we need to specify that opposite points on the surface of ball are actually the same. If this identification is made into a formal equivalence relation $\sim$ then we have $SO(3) \cong D^3/\sim$. So as a manifold $SO(3)$ can be visualized as a three-dimensional solid ball with opposite points on the surface of the ball identified. As a preview of what is to come, note that this shows that $SO(3)$ is *not* simply connected.

Exercise 7 *What is the relationship between $SO(3)$ and the 3-dimensional*

[2]For a general abstract manifold the definition of differentiable is that for two overlapping regions U_i and U_j with corresponding maps φ_i and φ_j the composition $\varphi_i \circ \varphi_j^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ is infinitely differentiable.

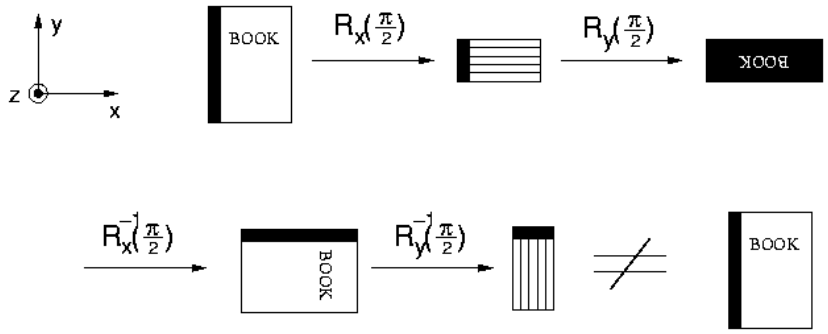

Figure 1.4: Rotating a book by the sequence $R_y^{-1}(\frac{\pi}{2})R_x^{-1}(\frac{\pi}{2})R_y(\frac{\pi}{2})R_x(\frac{\pi}{2})$ does not lead back to the initial configuration, demonstrating the non-Abelian nature of $SO(3)$.

unit sphere, S^3?

Exercise 8 *Does the set of general $n \times n$ matrices form a manifold? If so, what is its dimension? If not, can you add an additional condition to make it into a manifold?*

1.2.2 Non-commutativity

In addition to being a manifold, $SO(3)$ is a non-commutative (non-Abelian) group. If a group is Abelian, then $gh = hg$, which means that $g^{-1}h^{-1}gh = 1$. Let's see what we get for a similar product of rotations in $SO(3)$. Let $R_i(\theta)$ denote a rotation about the i-axis by an angle θ. Let's rotate a book by the sequence $R_y^{-1}(\frac{\pi}{2})R_x^{-1}(\frac{\pi}{2})R_y(\frac{\pi}{2})R_x(\frac{\pi}{2})$. (Note that the right-most rotation is done first.) Such a sequence of group elements is sometimes known as the commutator and is shown graphically in Figure 1.4.

At the end the book is definitely not back at its starting point, demonstrating that this product of group elements is not equal to the identity, proving that $SO(3)$ is non-Abelian. The actual result is some complicated rotation about a new axis. To get a better sense of what's happening it is useful to consider very small rotations by an infinitesimal angle ε. The commutator then becomes $R_y^{-1}(\varepsilon)R_x^{-1}(\varepsilon)R_y(\varepsilon)R_x(\varepsilon) = R_y(-\varepsilon)R_x(-\varepsilon)R_y(\varepsilon)R_x(\varepsilon)$, having used the fact that $R_i^{-1}(\theta) = R_i(-\theta)$. We can work out what this is using the explicit representation by 3×3 matrices and the fact that ε is infinitesimal.

For example,

$$T_z(\varepsilon) = \begin{pmatrix} \cos\varepsilon & -\sin\varepsilon & 0 \\ \sin\varepsilon & \cos\varepsilon & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1-\frac{\varepsilon^2}{2} & -\varepsilon & 0 \\ \varepsilon & 1-\frac{\varepsilon^2}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix} + \mathcal{O}(\varepsilon^3). \quad (1.6)$$

Multiplying out the commutator then gives

$$T_y(-\varepsilon)T_x(-\varepsilon)T_y(\varepsilon)T_x(\varepsilon) = \begin{pmatrix} 1 & \varepsilon^2 & 0 \\ -\varepsilon^2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + \mathcal{O}(\varepsilon^3). \quad (1.7)$$

From the expansion of $T_z(\varepsilon)$ we see that to $\mathcal{O}(\varepsilon^2)$ this is the same as $T_z(-\varepsilon^2)$.

Now we can use the notion of "closeness" in the manifold to get a better handle on the non-commutativity. To this end we write an infinitesimal rotation about the i-axis as $T_i(\varepsilon) = 1 + \varepsilon A_i$. If we want to build up a finite rotation we can simply apply n consecutive small rotations to get a rotation by $\theta = n\varepsilon$.

$$T_i(\theta) = T_i(n\varepsilon) = (T_i(\varepsilon))^n = \left(1 + \frac{\theta}{n}A_i\right)^n \underset{n\to\infty}{\longrightarrow} e^{\theta A_i} \quad (1.8)$$

Expanding the explicit form of T_x, T_y, and T_z leads to

$$A_x = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} \quad A_y = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} \quad A_z = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad (1.9)$$

Exercise 9 *Verify explicitly that $e^{\theta A_z}$ yields the matrix $T_z(\theta)$ shown in Eqn. (1.6).*

The As are called **generators** of the Lie group $SO(3)$, and they behave somewhat differently from the group elements. The $SO(3)$ matrices can be multiplied together without leaving the group, $T_1T_2 \in SO(3)$, but they can't be added: T_1+T_2 is not necessarily in $SO(3)$. (Consider $T_1-T_1 = \mathbf{0}$. The zero matrix is definitely not orthogonal! Or $I + I = 2I$, which has determinant 8 and not 1.) What about the generators? For $T(\theta) = e^{\theta A}$ to be in $SO(3)$ we need

$$T^\dagger(\theta)T(\theta) = \left(e^{\theta A}\right)^\dagger e^{\theta A} = e^{\theta^* A^\dagger} e^{\theta A} = e^{\theta(A^\dagger + A)} = I \quad (1.10)$$

where we have used the fact that the angle θ is a real parameter. This requirement means $A^\dagger + A = \mathbf{0} \Rightarrow A^\dagger = -A$. We say such a matrix A is skew-Hermitian (sometimes also called anti-Hermitian). Also, we need

$$\det T(\theta) = \det e^{\theta A} = e^{\theta \mathrm{Tr} A} = 1 \Rightarrow \mathrm{Tr}(A) = 0. \quad (1.11)$$

So altogether we find that the generators must be skew-Hermitian, traceless matrices. Adding linear combinations of such generators shows

$$\begin{aligned} (aA+bB)^\dagger &= a^*A^\dagger + b^*B^\dagger = -aA - bB = -(aA+bB) \quad (1.12) \\ \mathrm{Tr}\,(aA+bB) &= a\mathrm{Tr}A + b\mathrm{Tr}B = 0 \quad (1.13) \end{aligned}$$

as long as $a, b \in \mathbb{R}$. Thus the generators form a real vector space! But note that the vector space is not closed under matrix multiplication. For instance,

$$A_xA_y = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad (1.14)$$

which is not skew-Hermitian. By considering the generators instead of the group elements we give up the group structure but gain a vector space structure.

Exercise 10 *Prove that* $\left(e^A\right)^\dagger = e^{A^\dagger}$*, which was used in Eqn. (1.10).*

Exercise 11 *Prove that* $\det\left(e^A\right) = e^{\mathrm{Tr}A}$*, which was used in Eqn. (1.11). [Hint: Start by proving it when* A *is a diagonal matrix, and then generalize as much as you can.] Generalize this formula to rewrite the determinant of a general matrix* M *in terms of a trace.*

1.2.3 Lie Algebras

What does the non-commutativity look like in terms of the generators? Reconsider the infinitesimal rotations by ε. Recall that we found

$$T_y(-\varepsilon)T_x(-\varepsilon)T_y(\varepsilon)T_x(\varepsilon) = T_z(-\varepsilon^2). \quad (1.15)$$

Writing this in terms of the As and expanding in ε gives

$$\begin{aligned} e^{-\varepsilon A_y}e^{-\varepsilon A_x}e^{\varepsilon A_y}e^{\varepsilon A_x} &= \left(1-\varepsilon A_y+\frac{\varepsilon^2}{2}A_y^2+\mathcal{O}(\varepsilon^3)\right)\left(1+\varepsilon A_x+\frac{\varepsilon^2}{2}A_y^2+\mathcal{O}(\varepsilon^3)\right) \\ &\quad \left(1-\varepsilon A_y+\frac{\varepsilon^2}{2}A_y^2+\mathcal{O}(\varepsilon^3)\right)\left(1+\varepsilon A_x+\frac{\varepsilon^2}{2}A_y^2+\mathcal{O}(\varepsilon^3)\right) \\ &= 1+\varepsilon^2A_yA_x-\varepsilon^2A_xA_y+\mathcal{O}(\varepsilon^3) \quad (1.16) \end{aligned}$$

while

$$e^{-\varepsilon A_z} = 1-\varepsilon^2A_z+\mathcal{O}(\varepsilon^3). \quad (1.17)$$

Thus we determine that

$$A_x A_y - A_y A_x \equiv [A_x, A_y] = A_z, \tag{1.18}$$

where we have defined the bracket operation [,] for generators. Note that even though AB is not in the vector space, $[A, B]$ is:

$$\begin{aligned} [A,B]^\dagger &= (AB - BA)^\dagger = B^\dagger A^\dagger - A^\dagger B^\dagger = (-B)(-A) - (-A)(-B) \\ &= BA - AB = [B, A] = -[A, B] \end{aligned} \tag{1.19}$$

and $\mathrm{Tr}[A, B] = \mathrm{Tr}(AB - BA) = \mathrm{Tr}(AB) - \mathrm{Tr}(BA) = 0$ by the cyclic property of the trace. You can also check that [,] is compatible with the (real) vector space structure. This bracket operation has the three properties that define an abstract **Lie bracket**:

$$\text{i)} \quad [A, B] = -[B, A] \quad \text{antisymmetry} \tag{1.20}$$

$$\text{ii)} \quad [a_1 A_1 + a_2 A_2, B] = a_1 [A_1, B] + a_2 [A_2, B] \quad \text{bilinearity} \tag{1.21}$$

$$[A, b_1 B_1 + b_2 B_2] = b_1 [A, B_1] + b_1 [A, B_2] \tag{1.22}$$

$$\text{iii)} \quad [A, [B, C]] + [B, [C, A]] + [C, [A, B]] = 0 \quad \text{Jacobi identity.} \tag{1.23}$$

A vector space with the additional structure of a Lie bracket is called a **Lie algebra**. In fact, just knowing how the Lie bracket operates on generators is enough to allow you to work out the product of any two group elements. We won't prove that here, but to make it plausible consider the general situation of two elements generated by A and B: $g = e^A$ and $h = e^B$. Then $gh = e^A e^B$, which must in turn be equal to e^C for some C. Presumably C is determined by A and B in some way, but it is complicated if A and B do not commute. The Baker-Campbell-Hausdorff theorem provides the answer:

$$e^A e^B = e^C \quad \text{with} \quad C = A + B + \frac{1}{2}[A, B] + \frac{1}{12}\left([A, [A, B]] + [B, [B, A]]\right) + \cdots . \tag{1.24}$$

Exercise 12 *Derive Eqn. (1.24) and (if you're up for a challenge) work out the next term in the expansion. [Hint: Multiply both A and B by a dummy parameter λ and write C as a series of terms with increasing powers of λ. Then expand both sides and collect powers of λ.]*[3]

[3]For a systematic procedure for calculating the nth term (including a *Mathematica* implementation), see the paper by Matthias Reinsch, arXiv.org: math-ph/9905012.

All higher term in the expansion of C are determined by more nested brackets, demonstrating that you don't even need to be able to define the products AB or BA.

Exercise 13 *Looking back at the proof that the generators are skew-Hermitian, was I justified in writing* $e^{\theta^* A^\dagger} = e^{\theta(A^\dagger + A)}$ *?*

We illustrated the concept of generators and Lie algebras by using the specific, three-dimensional (defining) representation of $SO(3)$. However, the same thing can be done abstractly using a little more machinery and without making any reference to the specific representation of the group or its dimension. The result is that each Lie group has a corresponding Lie algebra that tells you all you need to know about the local structure of the group. In this example we uncovered the Lie algebra as the generators of the group using matrix exponentials. The Lie algebra can also be defined in terms of the manifold structure as the tangent space to the group at the identity, and the exponential of the generators can also be given a more abstract definition. We're interested in connecting to physics applications, so we won't pursue the abstract formulation here, but perhaps it would make a good project for someone.

In our concrete example of $SO(3)$, the Lie algebra is the real vector space of skew-Hermitian, traceless matrices with $[A_x, A_y] = A_z$ and cyclic permutations, often written as $[A_i, A_j] = \epsilon_{ijk} A_k$. Since A_x, A_y, and A_z form a basis for the Lie algebra (as a vector space), knowing how the Lie bracket operates on them is enough to allow us to determine the commutator of any two generators. For a general Lie group G the vector space of generators forms its Lie algebra, often denoted $\mathfrak{g}$. Having a representation of G means that $T : G \to T(V)$ assigns each group element g to a linear operator $T(g)$ where $T(g_1 g_2) = T(g_1)T(g_2)$. Thus if $g = e^A$ and $T(g)$ is a linear operator on some vector space V, we can define $T : \mathfrak{g} \to T(V)$ ("overloading the operator") such that $T(g) = e^{T(A)}$ where $T(A)$ is also a linear operator on V. (To be precise we need to show existence and uniqueness, but we will come back to that later.) In order for the group multiplication to make sense we must have

$$\begin{aligned} T(g_1 g_2) &= T(e^{A_1} e^{A_2}) = T(e^{A_1 + A_2 + \frac{1}{2}[A_1, A_2] + \cdots}) = e^{T(A_1 + A_2 + \frac{1}{2}[A_1, A_2] + \cdots)} \\ &= e^{T(A_1) + T(A_2) + \frac{1}{2} T([A_1, A_2]) + \cdots} \qquad (1.25) \\ &\stackrel{?}{=} e^{T(A_1)} e^{T(A_2)} = T(g_1) T(g_2). \qquad (1.26) \end{aligned}$$

Clearly we will have agreement between the last two lines if $T([A_1, A_2]) = [T(A_1), T(A_2)]$. This last condition is therefore the requirement for $T(A)$ to be a representation of the Lie algebra. Thus we can talk about representations of the group or representations of the algebra, and we can go back and forth between them using exponentiation. Often the Lie algebra is easier to work with, so physicists usually focus on that. We haven't yet discussed the precise relationship between Lie groups and Lie algebras. Our above construction shows how you can find a Lie algebra for every Lie group by determining the generators in a specific (faithful) representation. But we haven't yet dealt with the converse. The result is that it isn't quite a one-to-one relationship between Lie algebras and Lie groups, but we will come back to this later. First we will study quantum mechanics and see how physicists use and understand Lie algebras.

Exercise 14 *Describe the matrices that make up the Lie algebra for the group of all invertible, $n \times n$ real matrices. (This group is called $GL(n, \mathbb{R})$ and its Lie algebra is written $\mathfrak{gl}(n, \mathbb{R})$.*

Exercise 15 *Why is it important that the Lie algebra can be defined without needing to define the products of two generators? [Hint: This is subtle; you won't find the answer in these lecture notes.]*

1.3 Quantum Mechanics

As mentioned earlier, the underlying mathematical structure of quantum mechanics is a complex vector space V together with an inner product $\langle \ | \ \rangle$ that satisfies the usual properties:

$$\begin{aligned} \langle v|w\rangle &= \langle w|v\rangle^* && (1.27)\\ \langle v|v\rangle &\geq 0 \qquad \text{with equality iff } |v\rangle = 0 && (1.28)\\ \langle v|\alpha w + \beta u\rangle &= \alpha\langle v|w\rangle + \beta\langle v|u\rangle && (1.29)\\ \langle \alpha v + \beta u|w\rangle &= \alpha^*\langle v|w\rangle + \beta^*\langle u|w\rangle && (1.30) \end{aligned}$$

More precisely, quantum mechanics is formulated on a Hilbert space. We will use this terminology, but we won't worry about the subtle distinctions between a Hilbert space and a vector space. It is important to understand that each "system" is represented by its own Hilbert space. Presumably one could define the *total* Hilbert space of the universe, but obviously that

would be far more complicated than necessary. Physicists restrict to smaller systems which can be represented by simpler Hilbert spaces.

1.3.1 Postulates of Quantum Mechanics

Here are the postulates of quantum mechanics, according to Steven Weinberg in his field theory book[4]:

1. The state of a system is represented by a **ray** in the Hilbert space, defined as the equivalence class of normalized vectors $\langle v|v\rangle = 1$. The equivalence relation $\sim$ is defined such that $|v\rangle \sim |w\rangle$ if $|v\rangle = \alpha|w\rangle$ for some $\alpha \in \mathbb{C}$ with $|\alpha| = 1$.

2. Observables are linear operators $A : V \to V$ that are also **Hermitian**, i.e. $A^\dagger = A$. The Hermitian conjugate $A^\dagger$ is defined such that $\langle v|Aw\rangle = \langle A^\dagger|w\rangle$. For matrices $A^\dagger = (A^*)^T$.

3. If the system is in the state $|v\rangle$ and we do an experiment to see if it has properties of the state $|w\rangle$, the probability of finding $|v\rangle$ in state $|w\rangle$ is $P(v \to w) = |\langle w|v\rangle|^2$. Furthermore, the **expectation value** of an observable, which is the average value of many measurements on identically prepared systems in the state $|v\rangle$, is given by $\langle A\rangle = \langle v|A|v\rangle$.

Exercise 16 *Prove that the eigenvalues of a Hermitian matrix are real and that eigenvectors corresponding to distinct eigenvalues are orthogonal. Show that this implies that a Hermitian matrix is unitarily diagonalizable.*

In quantum mechanics we want to understand the state of the system $|\psi\rangle$, also called the **wave function**, which has its time evolution governed by the Schrödinger equation

$$i\hbar\frac{\partial}{\partial t}|\psi\rangle = H|\psi\rangle. \tag{1.31}$$

H is a Hermitian operator called the Hamiltonian, which essentially measures the energy of the system. To solve the Schrödinger equation we usually assume that H is independent of time and first solve the simpler equation $H|\psi\rangle = E|\psi\rangle$, with $E \in \mathbb{R}$. Quantum mechanics boils down to solving this

[4]Weinberg, S. *The Quantum Theory of Fields, Vol. I*, Cambridge University Press, 1996, p. 49-50.

eigenvalue equation with various forms for H. Symmetries are very useful for simplifying this procedure.

If the system possesses a symmetry it is usually manifest in the Hamiltonian. For instance, an atom of hydrogen consists of a negative electron in orbit around a much heavier, positive proton. The (lowest order) Hamiltonian that describes the state of the electron in a hydrogen atom is

$$H_{\text{hydrogen}} = -\frac{\hbar^2}{2m}\nabla^2 - \frac{e^2}{r} \tag{1.32}$$

which is spherically symmetric, meaning there is no preferred axis or direction. So we expect the physical configuration and predictions about this electron will not change under rotations in $SO(3)$. Since H_{hydrogen} is a linear operator on the infinite-dimensional Hilbert space of square-integrable complex functions, we are interested in representations of $SO(3)$ on that same vector (Hilbert) space. But here is where the reducibility of group representations comes in. Obviously an infinite dimensional vector space is a pretty big space to be working in, so it would be nice if the solutions to the eigenvalue equation $H|\psi\rangle = E|\psi\rangle$ broke down into smaller, finite dimensional subspaces. In fact, we can use the finite dimensional irreducible representations of $SO(3)$ to separate the full Hilbert space into smaller, more manageable pieces. But first we need to discuss what it means to have a symmetry in quantum mechanics.

1.3.2 Symmetries in Quantum Mechanics

If we have a system that possesses a symmetry that means that physically observable things such as probabilities for events should not be changed after a symmetry operation. So if g acts on a state $|\psi\rangle$ by an operator $T(g)$, then $|\psi\rangle \to T(g)|\psi\rangle$. If we measure the probability of our state $|\psi\rangle$ being in the state $|\phi\rangle$, the original probability is $P_{\text{orig}} = |\langle\phi|\psi\rangle|^2$. After the transformation, $|\phi\rangle \to T(g)|\phi\rangle \Rightarrow \langle\phi| \to \langle T(g)\phi| = \langle\phi|T^\dagger(g)$. Therefore $P_{\text{after}} = \left|\langle\phi|T^\dagger(g)T(g)|\psi\rangle\right|^2$. For the probability to remain unchanged for any initial and final state we need $T^\dagger T = \pm I$. There is a theorem of Wigner which states that any symmetry of the quantum mechanical system can be represented by a unitary operator $T^\dagger T = I$.[5]

[5]We have implicitly assumed that T is a linear operator. However, time reversal requires the use of an **antilinear** operator, which is defined to include complex conjugation of scalars. We won't pursue any symmetries containing time reversal.

Furthermore, if the system has a specific energy, i.e. it is in an eigenstate of H, we can't have a symmetry operation changing that eigenvalue because a measurement of the energy of that state shouldn't change under the action of a symmetry. So if $H|\psi\rangle = E|\psi\rangle$, before the symmetry transformation we have

$$\langle H\rangle = \langle\psi|H|\psi\rangle = \langle\psi|E|\psi\rangle = E\langle\psi|\psi\rangle = E. \tag{1.33}$$

After doing the symmetry transformation we need

$$\langle H\rangle = \langle\psi|T^\dagger HT|\psi\rangle = E, \tag{1.34}$$

or in other words, $T|\psi\rangle$ also needs to be an eigenstate of H with eigenvalue E: $HT|\psi\rangle = ET|\psi\rangle$. This will be true if H and $T(g)$ commute, because then $HT|\psi\rangle = T(g)H|\psi\rangle = T(g)E|\psi\rangle = E(T(g)|\psi\rangle)$. Note however that $T(g)|\psi\rangle$ need not be equal to $|\psi\rangle$. All we need is for each eigenspace that corresponds to a distinct eigenvalue of H to be an invariant subspace, meaning that it is mapped into itself by the action of all $T(g)$. This means that $T(g)$ is reducible, and that the eigenvalues of H label separate, smaller representations. Eventually we would like to break those smaller representations down further into irreducible representations. So we will focus our study on the irreducible, unitary representations of G. And to do that it is often easier to study the irreducible representations of $\mathfrak{g}$.

Now we come to a place where math and physics make a slight divergence. Recall that for e^A to be unitary, $(e^A)^\dagger e^A = 1$ required A to be skew-Hermitian, $A^\dagger = -A$. But quantum mechanics has a very special role for *Hermitian* operators. So what physicists do is put in an explicit factor of i and define generators that are Hermitian and therefore observables:

$$J \equiv iA \qquad \text{so} \qquad e^A = e^{-iJ}. \tag{1.35}$$

Since $A^\dagger = -A$ we have $(-iJ)^\dagger = iJ^\dagger = iJ \Rightarrow J^\dagger = J$. Recall that for $SO(3)$ we had $[A_i, A_j] = \epsilon_{ijk}A_k$. This now becomes

$$[-iJ_i, -iJ_j] = \epsilon_{ijk}(-iJ_k) \quad \Rightarrow \quad -[J_i, J_j] = -i\epsilon_{ijk}J_k \quad \Rightarrow \quad \boxed{[J_i, J_j] = i\epsilon_{ijk}J_k}.$$

The boxed expression is the famous set of commutation relations for $SO(3)$. It turns out that the J's are observables representing **angular momentum**, so they are also called the angular momentum commutation relations.

Exercise 17 *Using the fundamental commutation between position and momentum, $[x, p] = i$, show that the components of $\vec{L} = \vec{x} \times \vec{p}$ satisfy the angular momentum commutation relations.*[6]

From the mathematical standpoint this is a little weird. The J's still form a real vector space V_J and it possesses a Lie bracket of sorts, with the understanding that $[A, B] = iC$ for $A, B, C \in V_J$. But this clumsiness with the closure of the Lie algebra of the J's is a small price for physicists to pay for getting to work with Hermitian generators.

Back to our main example, we want to study the irreducible, unitary representation of $SO(3)$, so we will focus on the irreducible, traceless, Hermitian representations of the "Lie algebra" $\mathfrak{g}'$ spanned by J_x, J_y, and J_z with the commutation relation $\vec{J} \times \vec{J} = i\vec{J}$.

It turns out that there are finite dimensional representations with each dimension $n = 1, 2, 3, \ldots$. The $n = 1$ representation is simply given by $T(J_i) = [0]$, i.e. the 1×1 0-matrix. Clearly this yields the trivial representation of G: $T(g) = e^{T(A)} = e^{\mathbf{0}} = I$ for all A and thus all g. One could also have trivial representations of any dimension, but this is pretty boring.

For $n = 2$ the non-trivial representation is given by:

$$T_2(J_x) = \frac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad T_2(J_y) = \frac{1}{2}\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad T_2(J_z) = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{1.36}$$

Clearly these matrices are Hermitian and traceless. You should check that they satisfy the commutation relations.

For $n = 3$ we already discussed $A_x = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$ which yields $T_3(J_x) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix}$, etc. By changing basis such that $T_3(J_z)$ is diagonal we arrive

[6] I'm using units where $\hbar = 1$. If you want to retain the dimensions, you need to add an $\hbar$ to the right hand side of both commutation relations.

at the equivalent representation

$$T_3'(J_x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad T_3'(J_y) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & -i & 0 \\ i & 0 & -i \\ 0 & i & 0 \end{pmatrix} \quad (1.37)$$

$$T_3'(J_z) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}. \quad (1.38)$$

Exercise 18 *Show explicitly that T_3 and T_3' are equivalent.*

Similarly, there is a four-dimensional representation:

$$T_4(J_x) = \begin{pmatrix} 0 & \sqrt{\frac{3}{2}} & 0 & 0 \\ \sqrt{\frac{3}{2}} & 0 & 2 & 0 \\ 0 & 2 & 0 & \sqrt{\frac{3}{2}} \\ 0 & 0 & \sqrt{\frac{3}{2}} & 0 \end{pmatrix} \quad (1.39)$$

$$T_4(J_y) = \begin{pmatrix} 0 & -i\sqrt{\frac{3}{2}} & 0 & 0 \\ i\sqrt{\frac{3}{2}} & 0 & -2i & 0 \\ 0 & 2i & 0 & -i\sqrt{\frac{3}{2}} \\ 0 & 0 & i\sqrt{\frac{3}{2}} & 0 \end{pmatrix} \quad (1.40)$$

$$T_4(J_z) = \begin{pmatrix} \frac{3}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & -\frac{1}{2} & 0 \\ 0 & 0 & 0 & -\frac{3}{2} \end{pmatrix} \quad (1.41)$$

There are systematic rules for determining $T_n(J_i)$ for any positive integer n. Note that in each case the $T_n(J_i)$ are the basis vectors of a *real*, three-dimensional vector space, even though as matrices they themselves have complex components and act on a complex vector space of n-dimensions.

1.3.3 $SU(2)$ and Projective Representations

Let's look more carefully at $T_2(J)$. Does this lead to a two-dimensional representation of $SO(3)$? By explicit computation we find

$$e^{-i\theta T_2(J_x)} = \begin{pmatrix} \cos\frac{\theta}{2} & -i\sin\frac{\theta}{2} \\ -i\sin\frac{\theta}{2} & \cos\frac{\theta}{2} \end{pmatrix} \tag{1.42}$$

$$e^{-i\theta T_2(J_y)} = \begin{pmatrix} \cos\frac{\theta}{2} & -\sin\frac{\theta}{2} \\ \sin\frac{\theta}{2} & \cos\frac{\theta}{2} \end{pmatrix} \tag{1.43}$$

$$e^{-i\theta T_2(J_z)} = \begin{pmatrix} e^{-i\theta/2} & 0 \\ 0 & e^{i\theta/2} \end{pmatrix}. \tag{1.44}$$

Note that these are 2×2 *unitary* matrices: $U^\dagger U = 1$. How can we connect them to $SO(3)$ to check whether they do indeed form a representation? Recall our earlier characterization, $SO(3) \cong D^3/\sim$. Just as D^2, the two-dimensional disk (filled circle) is topologically the same as the northern hemisphere of S^2, D^3 is the "northern hemisphere" of S^3. Thus $SO(3)$ can be thought of as the northern hemisphere of S^3 with opposite equatorial points identified, which is the same as all of S^3 with antipodal points identified.

Now consider $SU(2)$, the group of 2×2 unitary matrices. The general matrix of this form can be written

$$U = \begin{pmatrix} \alpha & \beta \\ -\beta^* & \alpha^* \end{pmatrix} \tag{1.45}$$

with $\alpha, \beta \in \mathbb{C}$. Multiplying U and it's conjugate

$$U^\dagger U = \begin{pmatrix} \alpha^* & -\beta \\ \beta^* & \alpha \end{pmatrix}\begin{pmatrix} \alpha & \beta \\ -\beta^* & \alpha^* \end{pmatrix} = \begin{pmatrix} |\alpha|^2+|\beta|^2 & 0 \\ 0 & |\alpha|^2+|\beta|^2 \end{pmatrix} \tag{1.46}$$

we see that $U \in SU(2)$ only if $|\alpha|^2 + |\beta|^2 = 1$. Writing $\alpha = a_1 + ia_2$ and $\beta = b_1 + ib_2$ this condition becomes $|a_1|^2 + |a_2|^2 + |b_1|^2 + |b_2|^2 = 1$. This is nothing but the equation for a three sphere in real four dimensional space. Thus as a manifold $SU(2) \cong S^3$. The connection to S^3 suggests that $SO(3)$ is the same as $SU(2)$ with "opposite" elements identified. This can be made more explicit with the map $\pi : SU(2) \to SO(3)$ defined below.

$$e^{-i\theta_x T_2(J_x)} \mapsto e^{-i\theta_x T_3(J_x)} \tag{1.47}$$

$$\begin{pmatrix} \cos\frac{\theta_x}{2} & -i\sin\frac{\theta_x}{2} \\ -i\sin\frac{\theta_x}{2} & \cos\frac{\theta_x}{2} \end{pmatrix} \mapsto \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_x & -\sin\theta_x \\ 0 & \sin\theta_x & \cos\theta_x \end{pmatrix} \tag{1.48}$$

Note that for $\theta_x = 0$ we have

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \mapsto \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{1.49}$$

whereas for $\theta_x = 2\pi$ we get

$$\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \mapsto \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{1.50}$$

The same is true for mappings of the other generators. This demonstrates that

$$\pi : SU(2) \rightarrow SO(3) \tag{1.51}$$

$$e^{-i(\theta_x T_2(J_x)+\theta_y T_2(J_y)+\theta_z T_2(J_z))} \mapsto e^{-i(\theta_x T_3(J_x)+\theta_y T_3(J_y)+\theta_z T_3(J_z))} \tag{1.52}$$

is a double cover.

Exercise 19 *What is the dimension of $SU(n)$ (as a manifold)? Find a basis for its Lie algebra $\mathfrak{su}(3)$. What is it's dimension (as a vector space)? What is the dimension of $\mathfrak{su}(n)$?*

By construction T_2 gives a representation of $SU(2)$ and its Lie algebra, which indicates that the Lie algebra of $SU(2)$ is the same as that of $SO(3)$. Thus it appears that each Lie algebra doesn't necessarily correspond to a unique Lie group. The true relationship between Lie groups and Lie algebras is the following:

> Every Lie algebra corresponds to a unique **simply-connected** Lie group.

A simply connected space is one where every closed path can be smoothly shrunk to a point. We will (hopefully) talk more about simply-connected manifolds in a later section, but for now we will just consider the case of the sphere S^2. If you imagine any closed path on S^2 you can see that it can be shrunk to a point. (This is somewhat like trying to lasso an orange.) But if we now consider the space which is S^2 with antipodal points identified, we can draw a path from the north pole to the south pole. This is a closed path, because the north and south poles are really the same point, but this path

cannot be shrunk to a point no matter how you deform it. (In fancy language one says that the first homotopy group of the sphere is the cyclic group of order two, or $\pi_1(S^2) = \mathbb{Z}_2$.) A similar situation occurs for $S^3 \cong SU(2)$ and $S^3/\sim \; \cong SO(3)$. S^3 is simply connected, but $SO(3)$ is not, because a path connecting opposite points on S^3 is a closed path that cannot be deformed to a point.

Now that we have ascertained that $SU(2)$ and $SO(3)$ have the same Lie algebra and that $SU(2)$ is a simply-connected double cover of $SO(3)$, we can come back to the question of whether $T_2(J)$ yields a representation $T_2(g)$ of $SO(3)$. As a specific example, first look at the product of two elements, $R_y(\pi)R_x(\pi) = R_z(\pi)$, in the three-dimensional representation T_3.

$$T_3(R_y(\pi))T_3(R_x(\pi)) = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \tag{1.53}$$

$$= \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = T_3(R_z(\pi)) \tag{1.54}$$

So we see that T_3 preserves the group multiplication and follows the rules for representations (at least in this case). What about the same product in the two-dimensional representation?

$$T_2(R_y(\pi))T_2(R_x(\pi)) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & -i \\ -i & 0 \end{pmatrix} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} \tag{1.55}$$

But notice that $T_2(R_z(\pi)) = \begin{pmatrix} -i & 0 \\ 0 & i \end{pmatrix}$. Thus here we have an instance where $T_2(g_1)T_2(g_2) = -T_2(g_1g_2)$. This extra minus sign prevents T_2 from being a representation of $SO(3)$. However, since it *almost* qualifies, such representations "up to a sign" (or more generally, up to a phase) are given the name **projective representations**. Of course, for $SU(2)$ the two-dimensional representation is a normal (non-projective) representation, since the 2×2 unitary matrices define what we mean by $SU(2)$ in the first place. This example shows that the representations of the Lie algebra lead to representations of the corresponding simply-connected Lie group, but they might lead to projective representations of related, non simply-connected Lie groups with the same Lie algebra.

It turns out that all of the even dimensional representations of the Lie algebra $[J_i, J_j] = i\epsilon_{ijk}J_k$ yield projective representations of $SO(3)$, whereas the odd dimensional representations are ordinary representations. So what do we make of these projective representations? Going back to quantum mechanics we recall that each state is designated by a ray R_i in the Hilbert space which is an equivalence class of unit vectors. Abstractly, if the action of the group element g_1 takes $R_1 \to R_2$, and if $|\psi\rangle$ is a vector in R_1, then $T(g_1)|\psi\rangle$ must be a vector in R_2. Similarly, if g_2 takes $R_2 \to R_3$ then $T(g_2)\left(T(g_1)|\psi\rangle\right)$ must lie in R_3. Since the combined transformation g_2g_1 takes $R_1 \to R_3$ consistency requires $T(g_2g_1)|\psi\rangle$ must be a vector in R_3. From these considerations we learn that $T(g_2)\left(T(g_1)|\psi\rangle\right) \sim T(g_2g_1)|\psi\rangle$ or $T(g_2)\left(T(g_1)|\psi\rangle\right) = e^{i\phi(g_1,g_2)}T(g_2g_1)|\psi\rangle$. In order for this to hold for all vectors $|\psi\rangle$ we must have $T(g_2)T(g_1) = e^{i\phi(g_1,g_2)}T(g_2g_1)$. In other words, quantum mechanics only requires that $T(g)$ forms a projective representation of the symmetry group.

Exercise 20 *On the Hilbert space consisting of differentiable functions of angular coordinates $\psi(\theta, \phi)$ with inner product*

$$\langle\psi_1|\psi_2\rangle = \int_0^{2\pi}\int_0^{\pi} \psi_1^*\psi_2 \sin\theta \, d\theta \, d\phi, \tag{1.56}$$

the generator R_z is represented by the operator $L_z = -i\hbar\frac{\partial}{\partial\phi}$. Show that L_z is Hermitian and that that implies $\psi(\theta, 0) = \psi(\theta, 2\pi)$ for any ψ. What does this imply about projective representations in this situation?

1.3.4 The Electron

In quantum mechanics the representations of the angular momentum commutation relations are labeled not by their dimension, as we have done so far, but by the angular momentum quantum number j (sometimes alternatively ℓ or s). The quantum number j can be either an integer or half-integer starting with zero: $j = 0, \frac{1}{2}, 1, \frac{3}{2}, \ldots$. The dimension of the angular momentum j representation is $2j + 1$. Thus the half-integral angular momentum states correspond to even dimensional representations and the integral angular momentum corresponds to odd dimensional representations.

In the previous section we found that quantum mechanics would allow us to use both ordinary and projective representations of $SO(3)$ to describe a system. But we might hope that the mathematical formalism we have

developed might be useful in restricting the physically allowed states of a system. For instance, it would be interesting if for some reason all physical situations turned out to be described by ordinary representations of the rotation group, $SO(3)$. Then we might formulate a new physical principle stating that no realizable quantum systems for projective representations of the relevant symmetry group.

Physics is an experimental science, so we need to test the universe to determine whether our new principle excluding projective representations has any merit. This is where our mathematician friend the electron comes in to play. First we will send our electron into a spherically symmetric hydrogen atom to study the representations of $SO(3)$. The electron reports back that the angular wave-function for an electron in a hydrogen atom can be described by functions called spherical harmonics, $Y_\ell^m(\theta, \phi)$, which correspond to the odd-dimensional (ordinary) representations of $SO(3)$. So far, so good.

Now we send our electron to join a silver atom participating in a Stern-Gerlach experiment. The principle behind this type of experiment is the fact that a charged particle with non-zero angular momentum produces a magnetic moment, $\vec{\mu} = \frac{gq}{2m}\vec{J}$, where m is the particle's mass, q is it's electric charge, and g is dimensionless constant. When a magnetic moment passes through a nonuniform magnetic field it experiences a force. Quantum mechanics predicts that each angular momentum, and thus each magnetic moment, has only a discrete set of allowed positions, specified by the z-component of the angular momentum and given by the diagonal entries of $T_n(J_z)$. For example, in a system with $j = 1$ (3-dimensional representation of $SO(3)$) there can only be three values for the z-component, -1, 0, or $+1$. Classically, however, the angular momentum–and hence the magnetic moment–can have any z-component in a continuous distribution.

So when a beam of particles with magnetic moments passes through a nonuniform magnetic field the individual particles will experience different forces depending on the direction their magnetic moment is pointing. Classically one would expect the beam to spread out in a continuous distribution. Quantum mechanically one expects to get discrete bands corresponding to the allowed values of the z-component. In fact, the number of bands will be given by the dimension of the representation of $SO(3)$. Thus the Stern-Gerlach experiment can help us probe for even-dimensional (projective) representations.

Back in 1921 when the original experiment was carried out by Otto Stern

and Walter Gerlach, they were actually trying to test whether magnetic moments behaved classically (continuous distribution of deflected particles) or quantum mechanically (particles deflected into discrete bands). There are 47 electrons in a neutral silver atom, but they combine their orbital angular momenta together in such a way that all the internal angular momenta average out to zero, and the final valence electron is in an $\ell = 0$ state, i.e. the trivial representation. The result of the experiment was quite surprising. The beam of atoms was clearly split into *two* discrete bands, as shown in the image on the right in Figure 1.5. This means that the total magnetic moment of the silver atoms must be characterized by a two-dimensional representation, $j = \frac{1}{2}$. Since all the inner electrons' angular momenta cancels out, and the valence electron has orbital angular momentum zero, this means that the valence electron must have some additional, half-integral angular momentum. This **intrinsic angular momentum** carried by the electron is called **spin**, though it doesn't have anything to do with rotation because the electron really is understood as a point particle. But regardless, this spin forms a projective representation of $SO(3)$, thus disproving our hypothetical principle put forth above.

1.4 Philosophy

Now the discussion becomes more philosophical, so we'll start with a summary of what we've learned. Classically rotational invariance corresponds to the group $SO(3)$. When we move to quantum mechanics, we naturally want to find a way to use this symmetry. We found that the action of $SO(3)$ in quantum mechanics is mediated by unitary representations acting on the Hilbert space. What's more, those representations could be either ordinary or projective (representations up to a phase). Finally, the Stern-Gerlach experiment has demonstrated that the (even-dimensional) projective representations *do* indeed have a place in nature because they are needed to understand the intrinsic angular momentum of an electron (among other particles).

We also learned that these projective representations of $SO(3)$ are related to the fact that $SO(3)$ is not simply-connected.[7] We also found that the projective representations of $SO(3)$ were ordinary representations of $SU(2)$, a

[7] A possible project would be to flesh out the connection between projective representations and the topology of the group.

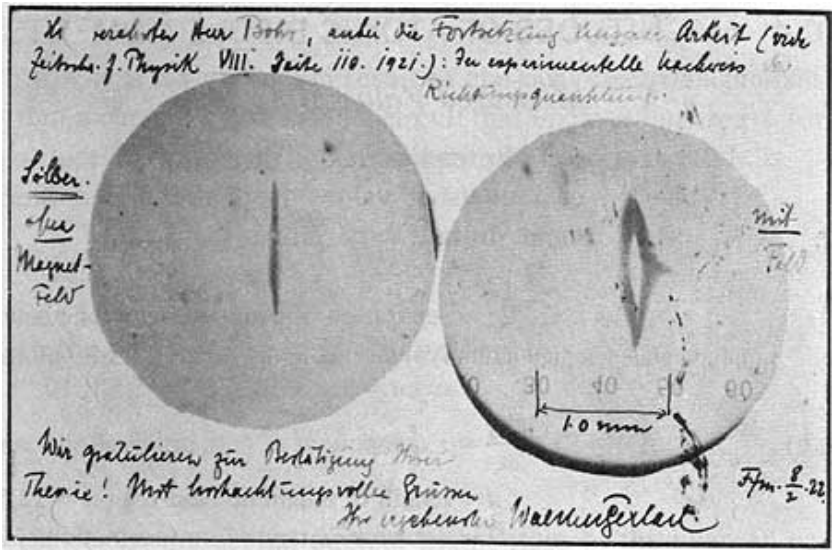

Figure 1.5: The result of the original Stern-Gerlach experiment, showing the splitting of a beam of silver atoms into two distinct bands. This figure was borrowed from a Physics Today article in December 2003, available at http://www.physicstoday.org/vol-56/iss-12/p53.html.

group that is a double cover of $SO(3)$. So maybe instead of worrying about projective representations we should just go ahead and switch to considering the simply-connected group $SU(2)$. Those of you who have taken some quantum mechanics know that this is exactly what physicists do in practice. However, since we're trying to understand the group theory in some detail, we will consider this choice in more detail.

The question we want to answer is the following: Is there any difference between using $SO(3)$ and projective representations and using $SU(2)$ where all the representations are ordinary? The answer is basically "no", so if you're sick of philosophy you can ignore the rest of this section. But I find it interesting to probe further.

Assume our Hilbert space consists of states $|\psi_o\rangle$ that are part of an ordinary representation of $SO(3)$, and other states $|\psi_p\rangle$ that are part of a projective representation. By definition we have

$$T(g_1g_2)|\psi_o\rangle = T(g_1)T(g_2)|\psi_o\rangle \tag{1.57}$$
$$T(g_1g_2)|\psi_p\rangle = e^{i\phi(g_1,g_2)}T(g_1)T(g_2)|\psi_p\rangle \tag{1.58}$$

where the phase $e^{i\phi(g_1,g_2)}$ is equal to ± 1 depending on g_1 and g_2. Now consider what would happen if we make a linear combination $|\psi\rangle = |\psi_o\rangle + |\psi_p\rangle$. Assume we act on the combination with two group elements g_1 and g_2 that have a nontrivial phase. Then we find

$$T(g_1g_2)|\psi\rangle = T(g_1g_2)\left(|\psi_o\rangle + |\psi_p\rangle\right) = T(g_1g_2)|\psi_o\rangle + T(g_1g_2)|\psi_p\rangle \tag{1.59}$$
$$= T(g_1)T(g_2)|\psi_o\rangle - T(g_1)T(g_2)|\psi_p\rangle \tag{1.60}$$
$$= T(g_1)T(g_2)\left(|\psi_o\rangle - |\psi_p\rangle\right). \tag{1.61}$$

This is a problematic equation. It seems that $|\psi\rangle$ doesn't really fit in a representation at all. One way around this difficulty is the use of a **superselection rule** which prohibits the linear combinations of states from different types of representations (ordinary and projective). Another way of phrasing this is to say that $|\psi_o\rangle + |\psi_p\rangle$ and $|\psi_o\rangle - |\psi_p\rangle$ are indistinguishable, which means that the subspaces containing $|\psi_o\rangle$ and $|\psi_p\rangle$ must always remain mutually orthogonal.

What if we were using representations of $SU(2)$? Then there would be no minus sign appearing in Eqn. (1.60) and we would have a nice, ordinary representation respecting the group multiplication law, and no reason to impose superselection rules.

So the question really comes down to whether we can make arbitrary linear superpositions of states. From the perspective of $SO(3)$ we see that linear combinations of states belonging to ordinary and projective representations are prohibited due to a selection rule that arises naturally from requiring a consistent group multiplication law. However, working with $SU(2)$ there is no mathematical reason to impose such a superselection rule. So *if* we were to prepare a state that was a linear combination of even and odd dimensional representations, we would know we were dealing with $SU(2)$ and *not* $SO(3)$.

Unfortunately, the real world isn't so cut and dried. According to the Nobel laureate physicist Steven Weinberg, "it is widely believed to be impossible to prepare a system in a superposition of two states whose total angular momenta are integers and half-integers, respectively".[8] So maybe the "true" quantum mechanical group is $SO(3)$ and the absence of certain linear combinations are the result of a mathematical requirement resulting in superselection rules. On the other hand, the group might still be $SU(2)$ and there is simply a *physical* principle that forbids those same linear combinations. Since there is no observable distinction between these two options (that I can see), we are left being unable to distinguish between $SO(3)$ and $SU(2)$. More generally, Weinberg calls the issue of superselection rules (and thus implicitly projective representations) a "red herring", because whether or not you can make arbitrary linear combinations one can't determine it by using symmetry arguments alone, since any group with projective representations can be replaced by another larger group (the universal covering group) that gives the same physical results but does not have any projective representations.

Exercise 21 *Here's a fun challenge: The power series for an exponential e^A looks similar to the Taylor series expansion of a function:*

$$f(x_0 + a) = f(x_0) + a \left.\frac{\partial}{\partial x}\right|_{x_0} f(x) + \frac{1}{2!} a^2 \left.\frac{\partial^2}{\partial x^2}\right|_{x_0} f(x) + \cdots . \tag{1.62}$$

Can you frame this similarity in terms of Lie algebras?

Exercise 22 *Another open-ended question: Recast the first postulate of quantum mechanics in terms of one-dimensional projection operators, i.e. $\rho =$*

[8] Weinberg, S. *The Quantum Theory of Fields, Vol. I*, Cambridge University Press, 1996, page 53.

$|v\rangle\langle v|$. *This leads to a formulation of quantum mechanics in terms of a* ***density matrix*** *which is a little more general that what we discussed above.*

Chapter 2

Local Symmetries

In this second chapter we discuss the concept of **local symmetry** that appears in physics. It seems to me that local symmetry represent is the way a **gauge symmetry** manifests itself in quantum field theory, so our story will begin with gauge symmetries as they appear in classical electrodynamics. Since gauge field configurations have interesting topological properties, we will explore homotopy groups in some detail as we work our way through the path integral formulation of quantum mechanics up to quantum field theory. Since all of these interesting topics are huge subjects in and of themselves, we cannot hope to cover them all in any detail. Therefore I have chosen to restrict the rigor (such as it is) to the discussion of homotopy theory and will, unfortunately, have to ask you to take most of the physics formalism and results on faith. I hope, however, that this is not too frustrating and that I can whet your appetite for further study of quantum mechanics and field theory.

2.1 Gauge Symmetries in Electromagnetism

The forces of electricity and magnetism are described by special vector fields. The "E-field", $\mathbf{E}(\mathbf{x})$ is a vector at each point in space that describes the electric field, while the "B-field", $\mathbf{B}(\mathbf{x})$ similarly describes the magnetic field. A particle carrying electric charge q that is moving in the presence of E- and B-fields feels a force given by

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}). \tag{2.1}$$

We will focus on electro- and magnetostatics, which means we only consider time-independent electric and magnetic fields. The E- and B-fields are "special", as mentioned above, because they satisfy the time-independent Maxwell equations:

$$\nabla \cdot \mathbf{B} = \rho/\epsilon_0 \qquad \nabla \times \mathbf{E} = 0 \tag{2.2}$$

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} \qquad \nabla \cdot \mathbf{B} = 0. \tag{2.3}$$

Here ρ is a static charge distribution and $\mathbf{J}$ is a current density, both of which act as sources that produce the electric and magnetic fields. So $\mathbf{E}$ and $\mathbf{B}$ are determined by ρ and $\mathbf{J}$, but they are also subject to the constraints that $\mathbf{E}$ be curl-free ($\nabla \times \mathbf{E} = 0$) and that $\mathbf{B}$ be divergenceless ($\nabla \cdot \mathbf{B} = 0$).

There is a slick way to take care of these constraints on $\mathbf{E}$ and $\mathbf{B}$ once and for all. If we write $\mathbf{E} = -\nabla\varphi$ where $\varphi : \mathbb{R}^3 \to \mathbb{R}^3$ is an arbitrary function, we find that $\nabla \times \mathbf{E} = \nabla \times (-\nabla\varphi) = 0$ always, since the curl of a gradient is identically zero. Then we just need to solve

$$\nabla \cdot \mathbf{E}(\mathbf{x}) = \nabla \cdot (-\nabla\varphi(\mathbf{x})) = -\nabla^2\varphi(\mathbf{x}) = \rho(\mathbf{x})/\epsilon_0 \tag{2.4}$$

to fine $\mathbf{E}(\mathbf{x})$ for a given charge distribution $\rho(\mathbf{x})$. This is a well studied problem for which many different techniques have been developed.

Thus we have simplified the problem from finding a 3-component vector field $\mathbf{E}(\mathbf{x})$ to finding a scalar field $\varphi(\mathbf{x})$. (That's great!) In addition, $\varphi(\mathbf{x})$ is called the **electric potential** because it is related to the potential energy a charged particle possesses when it is sitting in the E-field: potential energy $= q\varphi(\mathbf{x})$. However, there is some redundancy in using this "potential" formulation. If we change the potential by a constant, $\varphi'(\mathbf{x}) = \varphi(\mathbf{x}) + c$ it does not change the electric field:

$$\mathbf{E}'(\mathbf{x}) = -\nabla\varphi'(\mathbf{x}) = -\nabla(\varphi(\mathbf{x}) + c) = -\nabla\varphi(\mathbf{x}) + 0 = \mathbf{E}. \tag{2.5}$$

Physically this means that we can change the zero of energy arbitrarily; only energy *differences* matter.

A similar situation holds for magnetism, but it isn't quite as clean. For the B-field we want to impose the constraint $\nabla \cdot \mathbf{B} = 0$, so to follow the example of electricity, we want something whose divergence is always zero. Thinking back to vector calculus you might recall that the divergence of a curl always vanishes, so we will write $\mathbf{B} = \nabla \times \mathbf{A}$, where $\mathbf{A}$ is called the **magnetic vector potential**. Then

$$\nabla \cdot \mathbf{B} = \nabla \cdot (\nabla \times \mathbf{A}) = 0 \tag{2.6}$$

so we are left having to solve

$$\nabla \times \mathbf{B} = \nabla \times (\nabla \times A) = \mu_0 \mathbf{J}. \tag{2.7}$$

This is not quite as nice as the electrostatic case because

1. $\mathbf{A}$ is still a vector so we have three components to find,
2. there is no simple energy interpretation, and
3. there is even more redundancy.

It is this redundancy in the B-field that will occupy us below. We already mentioned that the curl of a gradient vanishes, which means that we can add any gradient to the vector potential, $\mathbf{A}'(\mathbf{x}) = \mathbf{A}(\mathbf{x}) + \nabla\lambda(\mathbf{x})$ without changing the magnetic field:

$$\mathbf{B}'(\mathbf{x}) = \nabla \times \mathbf{A}'(\mathbf{x}) = \nabla \times (\mathbf{A}(\mathbf{x}) + \nabla\lambda(\mathbf{x})) = \nabla \times \mathbf{A}(\mathbf{x}) + \nabla \times \nabla\lambda(\mathbf{x}) = \mathbf{B} \tag{2.8}$$

for any scalar function $\lambda(\mathbf{x})$.

Physical results and effects depend only on $\mathbf{E}(\mathbf{x})$ and $\mathbf{B}(\mathbf{x})$, however it is often nicer to study $\varphi(\mathbf{x})$ and $\mathbf{A}(\mathbf{x})$, especially in quantum mechanics. But we have seen above that the inherent redundancy means that there are many possible φ and $\mathbf{A}$ that describe the same physical situation ($\mathbf{E}$ and $\mathbf{B}$). In particular,

$$\varphi'(\mathbf{x}) = \varphi(\mathbf{x}) + c \qquad \text{and} \qquad \mathbf{A}'(\mathbf{x}) = \mathbf{A}(\mathbf{x}) + \nabla\lambda(\mathbf{x}) \tag{2.9}$$

yield identical results to $\varphi(\mathbf{x})$ and $\mathbf{A}(\mathbf{x})$. These changes to the potentials are called **gauge transformations**. The name doesn't mean anything, it is a historical holdover from people thinking about redefining the length scale, or "gauge". In practice, a gauge transformation or gauge symmetry refers to this change in potentials (i.e. change in the mathematical description) that leaves the physical fields unchanged.

Sometimes this gauge freedom can be used to simplify problems. Often we choose c such that

$$\lim_{|\mathbf{x}|\to\infty} \varphi(\mathbf{x}) = 0 \tag{2.10}$$

so that particles far away have no potential energy. Sometimes for $\mathbf{A}$ we require $\nabla \cdot \mathbf{A} = 0$. For instance, this simplifies Eqn. (2.7). However, there are plenty of other gauge choices that are more appropriate for other situations.

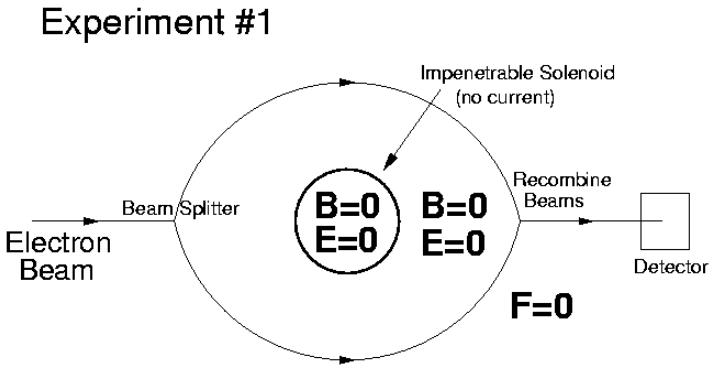

Figure 2.1: An idealized experiment to measure the Aharonov-Bohm effect. A beam of electrons is split in two and directed to either side of an impenetrable solenoid, then recombined and measured by a detector. In Experiment #1 there is no current in the solenoid, so there are no E- or B-fields anywhere.

Exercise 23 *Show how the gauge choice* $\nabla \cdot \mathbf{A} = 0$ *simplifies Eqn. (2.7).*

In summary, a gauge symmetry is a redundancy in the mathematical formulation of the problem, perhaps you could say it is a "symmetry in the formalism", but it is qualitatively different from the global symmetries we studied in the first chapter, which were symmetries of the physical system, like the triangle that could be rotated by 120° without being changed.

2.1.1 The Aharonov-Bohm Effect: Experiment

In 1959 Aharonov and Bohm described an interesting quantum mechanical effect of magnetic fields on charged particles. In an idealized experiment to measure the Aharonov-Bohm effect, one takes an electron beam, splits it into two, and sends the two beams on either side of a solenoid, a long coil of wire. The electron beams are then recombined and enter a detector. The solenoid is furthermore impenetrable to the electron beams. This setup is shown schematically in Figure 2.1. With no current flowing through the wires, there are no electric or magnetic fields anywhere, neither inside nor outside the solenoid. The electrons feel no force, and they recombine and yield a signal in the detector. This is the situation in Experiment #1.

In a second version of the same experiment, a steady current flows through the wire in the solenoid, producing a magnetic field inside the solenoid, as

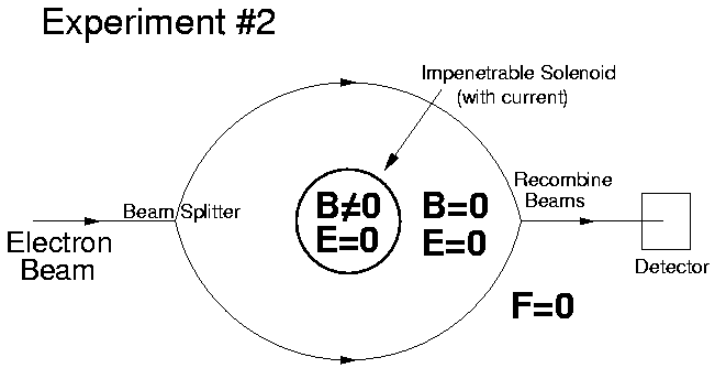

Figure 2.2: An idealized experiment to measure the Aharonov-Bohm effect. A beam of electrons is split in two and directed to either side of an impenetrable solenoid, then recombined and measured by a detector. In Experiment #2 there *is* a current in the solenoid, producing a magnetic field inside it. However, there are still no E- or B-fields in the regions where the electron beams pass through.

shown in Figure 2.2. However, there is still no electric or magnetic field outside the solenoid. Since the electrons cannot penetrate the solenoid, they never enter a region where the magnetic field is nonzero, so they continue to feel no force. Since the E- and B-fields at the location of the electrons are the same in both experiments, we would expect that the signal in the detectors should be the same in both experiments.

But this is wrong! The observed signal in Experiment #2 changes as a function of **B**, the magnetic field inside the solenoid. This surprising result, which has been confirmed experimentally, is a purely quantum mechanical effect. It is perhaps easiest to understand by using the path integral formulation of quantum mechanics, which we will briefly describe now.

2.1.2 The Path Integral Formulation of Quantum Mechanics

To understand the surprising result of Aharonov and Bohm we will use the path integral formulation of quantum mechanics, developed by Richard Feynman. This formulation is equivalent to the Schrödinger equation approach. Though it is *less* useful for actual calculations, it offers more insight into

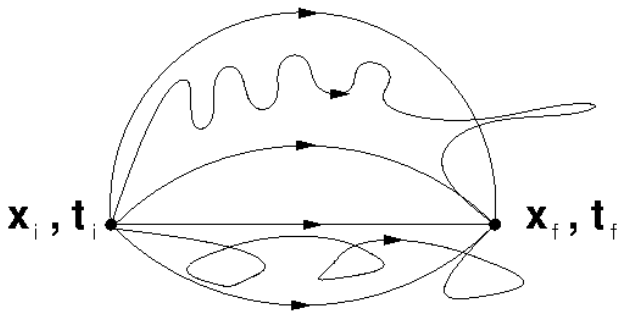

Figure 2.3: Examples of paths connecting an initial point ($\mathbf{x}_i$, t_i) to a final point ($\mathbf{x}_f$, t_f). Classically a particle will follow the path that minimizes the action, $S[\mathbf{x}(t)]$. Quantum mechanically the particle explores all the paths, where the contribution of each path to the transition amplitude is weighted by the factor $e^{\frac{i}{\hbar}S[\mathbf{x}(t)]}$.

what is really going on in quantum mechanics.

Our experiment consists of sending particles (electrons) from an initial location $\mathbf{x}_i$ at time t_i to a final location $\mathbf{x}_f$ at time t_f, and we are interested in the probability for that to occur. In our quantum mechanical Hilbert space we will let $|\mathbf{x}_i, t_i\rangle$ be the initial state vector and $|\mathbf{x}_f, t_f\rangle$ be the final state vector. Then we define the **amplitude** for the particle to move from $\mathbf{x}_i$ to $\mathbf{x}_f$ in time $t_f - t_i$ to be

$$\mathcal{A}(i \to f) = \langle \mathbf{x}_f, t_f | \mathbf{x}_i, t_i \rangle, \tag{2.11}$$

the inner product between the initial and final states. The probability $P(i \to f)$ for this transition is then given by the square of the amplitude:

$$P(i \to f) = |\mathcal{A}(i \to f)|^2 = |\langle \mathbf{x}_f, t_f | \mathbf{x}_i, t_i \rangle|^2 . \tag{2.12}$$

Since we want to know the probability, we need some way to calculate the amplitude.

In getting from point "i" to point "f" there are many paths the particle might take. A few possibilities are shown in Figure 2.3. But which path does the electron actually follow?

Classically, the particle picks the path that satisfies the **Principle of Least Action**. For every path specified by a function $\mathbf{x}(t)$ one can associate a real number called the **action** by way of the action functional, $S[x(t)]$.

A functional is a "function of a function", a map from the space of paths (continuous vector-valued functions of time) into the real numbers:

$$S : \left(\mathcal{C}^0\right)^3 \to \mathbb{R}, \tag{2.13}$$

where $\mathcal{C}^0$ stands for continuous functions of a real variable (t), and the power of three indicates the three components of a path in $\mathbb{R}^3$. (This might be non-standard notation.) For example, the action functional for a free particle with mass m traveling along a path $\mathbf{x}(t)$ is:

$$S_0[\mathbf{x}(t)] = \int_{t_i}^{t_f} dt\ \frac{1}{2}m\left(\frac{d\mathbf{x}}{dt}\right)^2 = \int_{t_i}^{t_f} dt\ \frac{1}{2}m\left[\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2\right]. \tag{2.14}$$

This is the time integral of the kinetic energy of the particle. Thus given a path $\mathbf{x}(t)$ you can compute $S_0[\mathbf{x}(t)]$ by taking derivatives and integrating. The Principle of Least Action states that the particle will follow the path $\mathbf{x}(t)$ for which the action $S[\mathbf{x}(t)]$ is a minimum (or technically, an extremum). In principle it is clear how to do this; just take all the paths, compute S, and keep the path for which S is smallest. But with an infinite number of paths connecting $\mathbf{x}_i$ to $\mathbf{x}_f$, this would be daunting in practice. The simple, systematic way to find the minimum of a functional is part of what is known as the **calculus of variations**. This is an interesting and fun subject, but we don't have time to pursue it here.

The Principle of Least Action is the classical method for finding the correct path followed by the particle. In quantum mechanics the situation is somewhat different, though there are still similarities. In the quantum mechanical description we assign to each path $\mathbf{x}(t)$ a complex number with unit modulus, $e^{\frac{i}{\hbar}S[\mathbf{x}(t)]}$ using the exact same action functional $S[\mathbf{x}(t)]$ that appears in the classical description. Then the amplitude is the *sum* of these exponential factors for all possible paths,

$$\mathcal{A}(i \to f) = \langle \mathbf{x}_f, t_f | \mathbf{x}_i, t_i \rangle = N \sum_{\text{paths } \mathbf{x}(t)} e^{\frac{i}{\hbar}S[\mathbf{x}(t)]} \tag{2.15}$$

where N is some normalization factor that we will ignore. Since there are an infinite number of paths, this expression is often written as an integral, namely the **path integral**:

$$\mathcal{A}(i \to f) = \langle \mathbf{x}_f, t_f | \mathbf{x}_i, t_i \rangle = \int_{\text{paths}} \mathcal{D}[\mathbf{x}(t)]\ e^{\frac{i}{\hbar}S[\mathbf{x}(t)]}. \tag{2.16}$$

Here $\mathcal{D}[\mathbf{x}(t)]$ stands for the "measure" over the space of paths, which we won't go into. From a mathematical perspective it is extremely annoying to write a formula without properly defining key parts. However, in practice if you do actually compute a path integral it is usually by discretizing all the paths, so the integration becomes a product of a large number of normal integrals for each of the segments of each path. Doing it this way isn't terribly revealing; it is very much like doing integrals in terms of Riemann sums.

The virtue of the path integral is that it helps build intuition about quantum mechanics. Conceptually, what you do is sum over all the paths, each with a different weighting factor $e^{\frac{i}{\hbar}S[\mathbf{x}(t)]}$ determined by the classical action.

But this weighting factor is a little weird. As systems get bigger or more complicated, we always need the quantum mechanical description to reduce to the classical description. (This is done formally by taking the limit as $\hbar \to 0$.) So one would naturally expect that the classical trajectory would contribute more than the crazy paths that we know are much less likely. However, since the weighting factors are all complex exponentials, they all have the same magnitude of 1. How do we recover the dominance of the classical path?

The dominance of the classical trajectory is related to the **method of stationary phase**, a technique used to evaluate some types of oscillatory complex integrals. The basic idea is that where the integrand is rapidly oscillating, nearby contributions are cancelling each other out and not contributing much to the value of the integral, whereas the main contributions come from regions where the integrand is not changing very quickly. Here is an example to show how this works. Suppose we want to evaluate the integral

$$\int_{-\infty}^{\infty} \cos\left[20(w^3 - w)\right]\, dw. \tag{2.17}$$

The argument of the cosine function is $\phi = 20(w^3 - w)$, the "phase". If we plot $\phi(w)$ we see that it is a cubic with a max and a min near the origin. For large values of w, ϕ is very large, and more importantly, when w changes slightly ϕ changes a lot, which means that the integrand, $\cos\phi$, moves through several full periods quite rapidly. The phase ϕ changes most slowly as a function of w near its maximum and minimum, and thus the integrand also changes most slowly there. Thus the main contributions to the integral come from the regions where ϕ is changing most slowly, namely

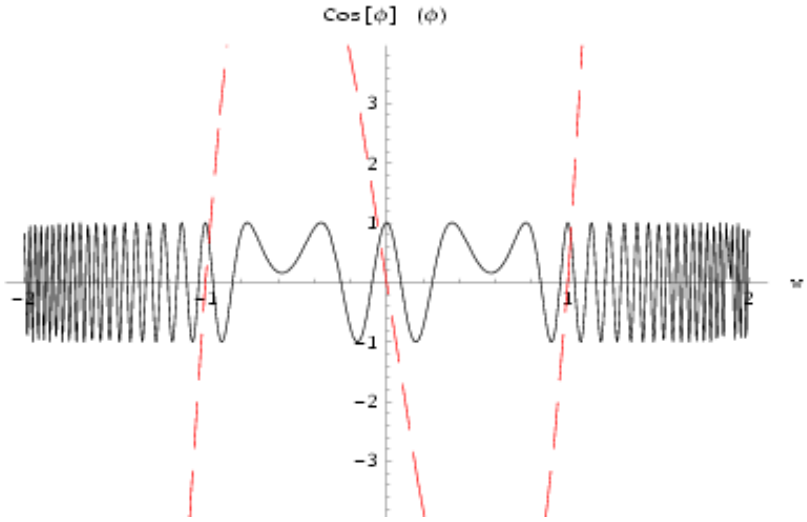

Figure 2.4: Plot of $\phi(w) = 20(w^3 - w)$ (dashed red curve) and $\cos\phi(w)$ (solid black curve). In regions where ϕ is changing quickly $\cos\phi$ oscillates rapidly, whereas $\cos\phi$ ceases to oscillate as much in the regions where ϕ is extremal.

where ϕ is "stationary", which means its derivative vanishes and ϕ is at a maximum or minimum. This is probably easier to see graphically than it is to describe in words. In Figure 2.4 the dashed red curve is a plot of ϕ as a function of w, while the solid black curve is the integrand, $\cos\phi$. Notice how $\cos\phi$ oscillates very rapidly except near the stationary points of ϕ. To actually evaluate the integral you expand the phase in a Taylor series about the point of stationary phase and do the integral in that region, ignoring the other small contributions.

Exercise 24 *Use the method of stationary phase to find an a approximate value for the integral* $f(x) = \int_0^\infty \cos x(w^3 - w)\, dw$ *as* $x \to \infty$. *The reason for the limit is that the method of stationary phase gets better as* $x \to \infty$. *(Why?) [Hint: The answer is* $f(x) = \sqrt{\frac{\pi}{3xw_0}} \cos\left[x(w_0^3 - w_0) + \frac{\pi}{4}\right]$, *where* w_0 *is the location of the minimum of* ϕ. *This was computed by Stokes in 1883.*

Coming back to the path integral, the idea is the same as in the previous example. The classical trajectory is the path where the phase in the path integral, in this case the action, is at an extremum. Thus for paths near the classical trajectory, the action isn't changing very rapidly, so most of

those paths give very similar contributions to the path integral. However, for the crazy paths doing weird things, the action for two nearby paths can be quite different, resulting in contributions to the path integral that might have opposite signs and cancel each other out. Thus the path integral picture of the world is as follows: the particle explores *all possible* paths between the starting and ending point. However, only the paths where the action is nearly extremal give sizeable contributions to the amplitude because they add together "in phase", whereas other paths tend to cancel out with their neighbors.

2.1.3 The Aharonov-Bohm Effect: Explained

With our new found intuition derived from the path integral, we are now in a position to explain the Aharonov-Bohm effect resulting from the interference of electron beams.

First we need the action for a particle with charge q moving in a magnetic field. Without proof, I assert that it is given by:

$$S[\mathbf{x}(t)] = \int_{t_i}^{t_f} dt\, \frac{1}{2} m \left(\frac{d\mathbf{x}}{dt}\right)^2 + q\frac{d\mathbf{x}}{dt}\cdot\mathbf{A} = S_0[\mathbf{x}(t)] + \int_{t_i}^{t_f} q\frac{\mathbf{x}}{dt}\cdot\mathbf{A}\, dt. \quad (2.18)$$

Note that it is the vector potential $\mathbf{A}$ and *not* the magnetic field $\mathbf{B}$ that appears in this expression. This is crucial, because even though $\mathbf{B} = 0$ outside the solenoid, $\mathbf{A} \neq 0$ in that region.

Exercise 25 *Working in cylindrical coordinates, show that if* $\mathbf{B} = B\hat{\mathbf{z}}$ *is a uniform magnetic field in the region* $r < R$ *and* $\mathbf{B} = 0$ *for* $r > R$*, then the magnetic vector potential is given by* $\mathbf{A} = \frac{1}{2}Br\hat{\boldsymbol{\phi}}$ *for* $r < R$ *and* $\mathbf{A} = \frac{BR^2}{2r}\hat{\boldsymbol{\phi}}$ *for* $r > R$*. Why don't we just take* $\mathbf{A} = 0$ *for* $r > R$*?*

The time integral of $\mathbf{A}$ can be converted to a line integral,

$$q\int_{t_i}^{t_f} \frac{\mathbf{x}}{dt}\cdot\mathbf{A}\, dt = q\int_{\mathbf{x}_i}^{\mathbf{x}_f} \mathbf{A}\cdot d\mathbf{x}, \quad (2.19)$$

which allows us to write the action as

$$S[\mathbf{x}(t)] = S_0 + q\int \mathbf{A}\cdot d\mathbf{x}. \quad (2.20)$$

So the amplitude for the electron to travel from the initial point to the detector is given by

$$\mathcal{A}(i \to f) = \langle \mathbf{x}_f, t_f | \mathbf{x}_i, t_i \rangle = \int_{\text{paths}} \mathcal{D}[\mathbf{x}(t)]\ e^{\frac{i}{\hbar} S[\mathbf{x}(t)]} \tag{2.21}$$

$$= \int_{\text{paths}} \mathcal{D}[\mathbf{x}(t)]\ e^{\frac{i}{\hbar} S_0 + q \int \mathbf{A} \cdot d\mathbf{x}}. \tag{2.22}$$

We can separate this path integral into two pieces, one that sums up paths that travel "above" the solenoid (as seen in the diagram) and another piece summing paths that go "below" the solenoid,

$$\mathcal{A}(i \to f) = \int_{\text{above}} \mathcal{D}[\mathbf{x}(t)]\ e^{\left[\frac{i}{\hbar} S_0 + \int \mathbf{A} \cdot d\mathbf{x}\right]_{\text{above}}} + \int_{\text{below}} \mathcal{D}[\mathbf{x}(t)]\ e^{\left[\frac{i}{\hbar} S_0 + \int \mathbf{A} \cdot d\mathbf{x}\right]_{\text{below}}}$$

Each path integral has an exponential factor that depends on the magnetic field through the vector potential, $e^{\frac{iq}{\hbar} \int \mathbf{A} \cdot d\mathbf{x}}$. However, because $\mathbf{B} = \nabla \times \mathbf{A} = 0$ in the region outside the solenoid where the paths are located, the line integral of $\mathbf{A}$ depends only on the endpoints $\mathbf{x}_i$ and $\mathbf{x}_f$ and *not* on the specific path between them. Therefore the exponential pieces containing $\mathbf{A}$ are independent of the paths that are being integrated over, so they can be pulled outside of the integral, yielding

$$\mathcal{A}(i \to f) = \left[e^{\frac{iq}{\hbar} \int \mathbf{A} \cdot d\mathbf{x}}\right]_{\text{above}} \int_{\text{above}} \mathcal{D}[\mathbf{x}(t)]\ e^{\left[\frac{i}{\hbar} S_0\right]_{\text{above}}} \tag{2.23}$$

$$+ \left[e^{\frac{iq}{\hbar} \int \mathbf{A} \cdot d\mathbf{x}}\right]_{\text{below}} \int_{\text{below}} \mathcal{D}[\mathbf{x}(t)]\ e^{\left[\frac{i}{\hbar} S_0\right]_{\text{below}}}. \tag{2.24}$$

The probability of arriving at the detector is given by the absolute square of the amplitude. Using abbreviated notation, this is

$$P(i \to f) = |\mathcal{A}(i \to f)|^2 = |\langle \mathbf{x}_f, t_f | \mathbf{x}_i, t_i \rangle|^2 \tag{2.25}$$

$$= \left| \int_{\text{above}} \right|^2 + \left| \int_{\text{below}} \right|^2 \tag{2.26}$$

$$+2\ \text{Re} \left(e^{-\frac{iq}{\hbar} \int \mathbf{A} \cdot d\mathbf{x} \big|_{\text{above}}} e^{\frac{iq}{\hbar} \int \mathbf{A} \cdot d\mathbf{x} \big|_{\text{below}}} \int_{\text{above}}^{*} \int_{\text{below}} \right) \tag{2.27}$$

$$= \left| \int_{\text{above}} \right|^2 + \left| \int_{\text{below}} \right|^2 + 2\ \text{Re} \left(e^{\frac{iq}{\hbar} \oint \mathbf{A} \cdot d\mathbf{x}} \int_{\text{above}}^{*} \int_{\text{below}} \right) \tag{2.28}$$

The last line combines a path below the solenoid with the negative of a path above to form a closed line integrals of the vector potential around the solenoid. Evaluating this using vector calculus theorems yields $\oint \mathbf{A} \cdot d\mathbf{x} = \int_{\text{surface}} (\nabla \times \mathbf{A}) \cdot d\mathbf{a} = \int_{\text{surf}} \mathbf{B} \cdot d\mathbf{a} = \Phi$, where Φ is the magnetic flux passing through the closed loop, which is exactly the flux inside the solenoid. Thus the magnetic field dependence of the probability for electrons to reach the detector will have a contribution from the cross-terms that is proportional to a sine and/or cosine of $q\Phi/\hbar$. This is how the magnetic field dependence arises in our Aharonov-Bohm experiment, even though $\mathbf{B} = 0$ in all regions where the electrons travel. In short, the probability has a term dependent on the flux enclosed between two different paths taken by the electrons.

The crucial point here is that the setup with a nonzero B-field in the solenoid means there are paths in the plane that encircle the magnetic flux and others that don't. One can also imagine paths that encircle the flux several times. The number of times a closed path encircles the solenoid is called the **winding number** and is a topological property of the space in which the electrons are traveling. Thus the Aharonov-Bohm effect is a topological effect in quantum mechanics.

One might have few objections at this point. First, doesn't this prove that the magnetic vector potential $\mathbf{A}$ is physical? It doesn't, because the final result only depends on $\Phi = \int \mathbf{B} \cdot d\mathbf{a}$, not $\mathbf{A}$! $\mathbf{A}$ appears in intermediate steps, so you might want to think about whether this result could be formulated in a way that makes no reference to the vector potential.

A second protest is that this is all well and good for the simple, idealized two-dimensional example presented here, but in the real, three-dimensional world isn't some magnetic field leaking out of the solenoid, which isn't really infinite anyway? Are the electron beams *really* probing the topology of the space? The answer is a resounding "yes". Many experiments to test the Aharonov-Bohm effect have been done, but despite their solid results skeptics always come up with protests about leakage fields and the like. However, in 1985 the definitive experiment was done by Tonomura and collaborators.[1] They fabricated a toroidal ferromagnet, covered it with a superconducting layer to keep the magnetic field from leaking out, and covered that with a copper conducting layer to prevent the electron wavefunctions from leaking in. Then they looked at the interference pattern produced by illuminating the torus with an electron beam, as shown in Figure 2.5. In the first figure,

[1] Tonomura, et al., Phys. Rev. Lett. **56** p. 792, 1986.

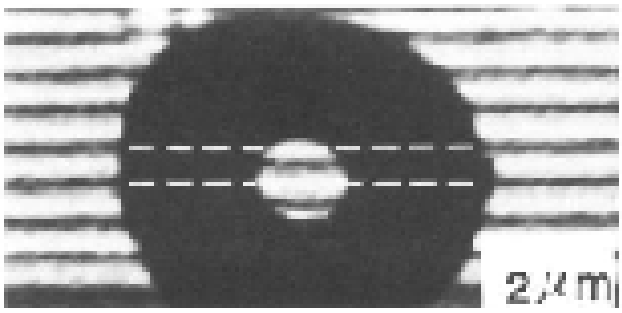

Figure 2.5: Interference fringes for electron beams passing near a toroidal magnet from the experiment by Tonomura and collaborators in 1985. The electron beam passing through the center of the torus acquires an additional phase, resulting in fringes that are shifted with respect to those outside the torus, demonstrating the Aharonov-Bohm effect. For details see the original paper from which this image was borrowed: Figure 6 in Tonomura et al., Phys. Rev. Lett. **56**, p.792, 1986.

the part of the electron beam that passes outside the torus interfere with a reference beam to produce the horizontal interference fringes. The part of the beam that passes through the hole in the torus also interferes with the reference beam, but the fringes are obviously shifted, indicating a different phase ($e^{\frac{iq}{\hbar}\int \mathbf{A}\cdot d\mathbf{x}}$) for the electron beam passing on the other side of the magnetic flux. The other two figures show various checks that can be done to prove that everything is working as intended. Tonomura and company conclude their paper with the following wonderful paragraph:

> The most controversial point in the dispute over experimental evidence for the AB effect has been whether or not the phase shift would be observed when both electron intensity and magnetic field were extremely small in the region of overlap. Since experimental realization of absolutely zero field is impossible, the continuity of physical phenomena in the transition from negligibly small field to zero field should be accepted instead of perpetual demands for the ideal; if a discontinuity there is asserted, only a futile agnosticism results.

Why does this experimental setup with the torus reflect the same situation we analyzed? The toroidal configuration also has the property that

closed loops surround the magnetic field an integer number of times. In mathematical language this means that the plane with a solenoid removed ($\mathbb{R}^2 - 0$) and three-space with a toroid removed ($\mathbb{R}^3 - T^2$) has the same first homotopy group, π_1. In the next section we will develop the idea of homotopy groups in more detail and rigor.

2.2 Homotopy Groups

This introduction to homotopy groups will closely follow the notation and presentation in Chapter 4 of Nakahara. We start by defining a **path** $\alpha : [0,1] \to X$ as a continuous map from the unit interval into a space X. The initial point is $\alpha(0) = x_0$ and the final point is $\alpha(1) = x_1$.

A **loop** is a path where the initial and final points coincide, $\alpha(0) = \alpha(1) = x_0$, and x_0 is called the **base point**.

As an example of a path, $\alpha : [0,1] \to \mathbb{R}^2$ with $\alpha(s) = (\cos \pi s, \sin \pi s)$ is a path in the plane that traces out the upper half of the unit circle from $(1,0)$ to $(-1,0)$.

Another example is the constant path, $c_x : [0,1] \to X$ with $c_x(s) = x \ \forall s$. Of course, the constant path is also a constant loop since its initial and final points coincide.

We can define a product of paths as follows. Let $\alpha, \beta : [0,1] \to X$ be two paths such that the final point of α is the same as the initial point of β, namely $\alpha(1) = \beta(0)$. Then we define the product $\alpha * \beta : [0,1] \to X$ to be the path in X where

$$\alpha * \beta = \begin{cases} \alpha(2s), & s \in [0, \frac{1}{2}] \\ \beta(2s-1), & s \in [\frac{1}{2}, 1] \end{cases} . \tag{2.29}$$

This definition is fairly intuitive, since you first travel along α and then travel along β, as shown in Figure 2.6. The complication comes from redefining the parameter so that the product path $\alpha * \beta$ also has a domain of $[0,1]$.

We can also define the inverse of a path, $\alpha^{-1} : [0,1] \to X$, by $\alpha^{-1}(s) \equiv \alpha(1-s)$. This is simply the path α traveled in reverse, starting at $\alpha(1)$ and ending at $\alpha(0)$.

Now that we've defined a product, $*$, and and inverse, α^{-1}, we can ask whether the set of paths in X forms a group. The zeroth order requirement for a group is that you have a well defined multiplication operation. However, if we have two paths where the endpoint of the first is not the same as the

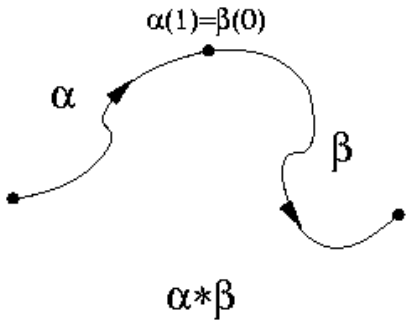

Figure 2.6: The product of two paths α and β is written $\alpha*\beta$ and corresponds to first traversing α and then traversing β.

starting point of the second, we don't know how to multiply them. We might try to remedy this by considering only paths that have a given initial point x_0 and a given final point x_f, but we wouldn't be able to multiply them because they all have the same initial point. Further, the inverse path wouldn't live in this set. To make our multiplication well defined and allow the inverse path, we apparently need to restrict to loops with a fixed base point, x_0.

Now that we have a well-defined multiplication on the set of loops with a fixed base point, does that set form a group? First we check associativity, where we need $\alpha * (\beta * \gamma) = (\alpha * \beta) * \gamma$. Looking at the sketch on the left side of Figure 2.7 suggests that this should hold, since both products have us tracing out the loops α then β and then γ. But closer inspection of the definitions for multiplication show that there is a subtle difference in the parameterizations. Consider $\alpha * (\beta * \gamma)(s)$. The rule for multiplication says that during the first half of the interval, $s \in [0, \frac{1}{2}]$, we follow α, while in the second half, $s \in [\frac{1}{2}, 1]$ we follow $\beta * \gamma$. But the since this second half is composed of the product of β and γ, that means that for the third quarter of the interval, $s \in [\frac{1}{2}, \frac{3}{4}]$ we follow β and in the last quarter we follow γ. Compare this with $(\alpha * \beta) * \gamma$, where the first half of the unit interval is devoted to α and β, whereas the whole second half of the interval is given to γ. So even though the points in X visited by these two paths are the same, the parameterizations are different, which means that the paths are, technically, different. This is easy to see graphically by labeling the portions of the unit interval that correspond to traversal of each loop, as shown in the right half of Figure 2.7. One can also write out the product loops explicitly

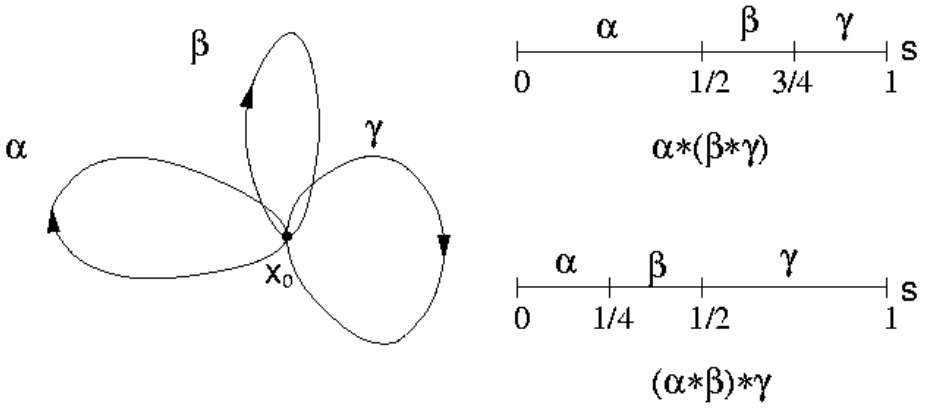

Figure 2.7: Graphically $\alpha * \beta * \gamma$ looks associative, because all three loops α, β, and γ are traversed in order, however closer inspection of the explicit functions reveals that the two groupings give different parameterizations of the combined paths.

to verify that they are indeed distinct.

$$\alpha * (\beta * \gamma)(s) = \begin{cases} \alpha(2s), & s \in [0, \frac{1}{2}] \\ \beta(4s-2), & s \in [\frac{1}{2}, \frac{3}{4}] \\ \gamma(4s-3), & s \in [\frac{3}{4}, 1] \end{cases} \tag{2.30}$$

versus

$$(\alpha * \beta) * \gamma(s) = \begin{cases} \alpha 4s, & s \in [0, \frac{1}{4}] \\ \beta(4s-1), & s \in [\frac{1}{4}, \frac{1}{2}] \\ \gamma(2s-1), & s \in [\frac{1}{2}, 1] \end{cases} . \tag{2.31}$$

So associativity does not hold immediately, though we would certainly like to come up with a clean way for two loops that differ only in their parameterization to be considered equal. Never fear, we will do this shortly.

The second property of groups is the presence of an identity element. The natural choice is certainly the constant loop, c_{x_0}, which stays put at the base point. But do we have $\alpha * c_{x_0} = \alpha = c_{x_0} * \alpha$? Again, we have some problems with different parameterizations, which is clearly seen by looking at the unit intervals labeled with the loops that are traversed during each segment, shown in Figure 2.8. We see that $\alpha * c_{x_0}$ follows α for half the time and then sits at the base point for the rest of the time, whereas $c_{x_0} * \alpha$ does the reverse, sitting at the base point for the first half and tracing out α

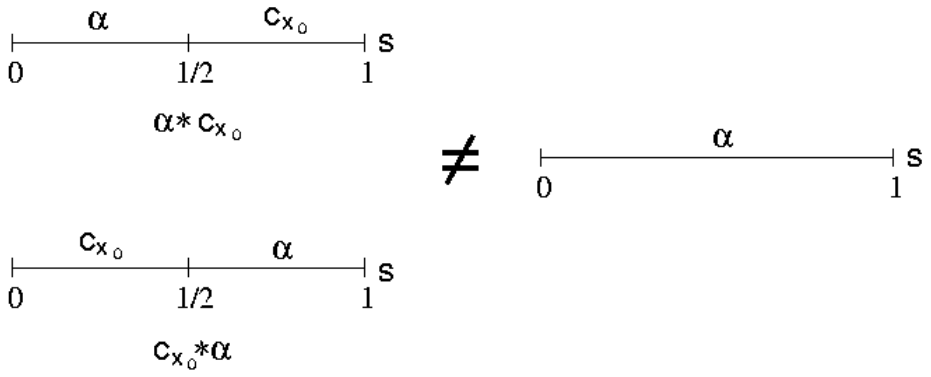

Figure 2.8: Products of α with the constant loop c_{x_0} yield loops that are technically different from α by itself. Again, the difference is only in parameterization, since the same points of X are visited by the various paths.

during the second half. And neither of these are identical to α, which doesn't do any waiting at the base point. Hopefully any fix of the parameterization ambiguities will solve this problem as well.

Finally, the last property needed to have a group is the existence of an inverse. The product of α and α^{-1} is given by

$$\alpha * \alpha^{-1}(s) = \begin{cases} \alpha(2s), & s \in [0, \frac{1}{2}] \\ \alpha^{-1}(2s-1) = \alpha(2-2s), & s \in [\frac{1}{2}, 1] \end{cases} \tag{2.32}$$

which corresponds to traveling along *alpha* and then backtracking and traversing alpha in the opposite direction. Even though you haven't really gone anywhere, this loop is definitely different from the constant loop, $c_{x_0}(s) = x_0$ for $s \in [0, 1]$, where you just sit at the base point the whole time. This is shown in Figure 2.9. Here the problem is worse than in the previous two cases, because it isn't just a matter of parameterization. $\alpha * \alpha^{-1}$ reaches points in X that are not reached by c_{x_0}.

Is there some way to fix the situation so that we can define a group structure on the set of loops with a fixed base point? Since the word "group" is in the title of this section, it seems very likely that this is the case. The fix is called **homotopy**, which is where we will turn now.

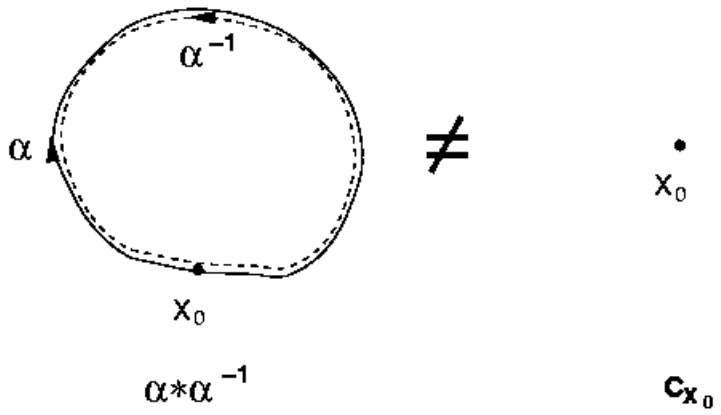

Figure 2.9: The product of α with α^{-1} yields a loop that is traversed first in one direction and then the other. This is very different from the constant loop c_{x_0} where only one point x_0 in X is ever visited.

2.2.1 Homotopy

Without further ado, we define homotopy. Let $\alpha, \beta : [0,1] \to X$ be two loops with the same base point x_0. Then α and β are **homotopic**, written $\alpha \sim \beta$, if there exists a continuous map $F : [0,1] \times [0,1] \to X$ such that

$$F(s,0) = \alpha(s) \qquad F(s,1) = \beta(s) \qquad \forall s \in [0,1] \tag{2.33}$$

$$F(0,t) = F(1,t) = x_0 \qquad \forall t \in [0,1]. \tag{2.34}$$

F is called a homotopy between α and β.

Let's unpack this formal definition. F is a function of two variables, s and t. The first variable, s is the parameter that traces out the trajectory of a loop, so the first condition says that for $t = 0$ F is the loop α, whereas for $t = 1$ F is the loop β. The second variable, t, gives a continuous transition from $\alpha(s)$ to $\beta(s)$. The second condition tells us that every t corresponds to a loop in X, since the initial and final points are fixed to be the base point x_0. Thus the homotopy $F(s,t)$ is a specific prescription for turning one loop into another in a continuous way such that you always have a loop in X. This can be shown graphically by looking at the domain, which is a square in the s-t plane, and the image in X, shown in Figure 2.10.

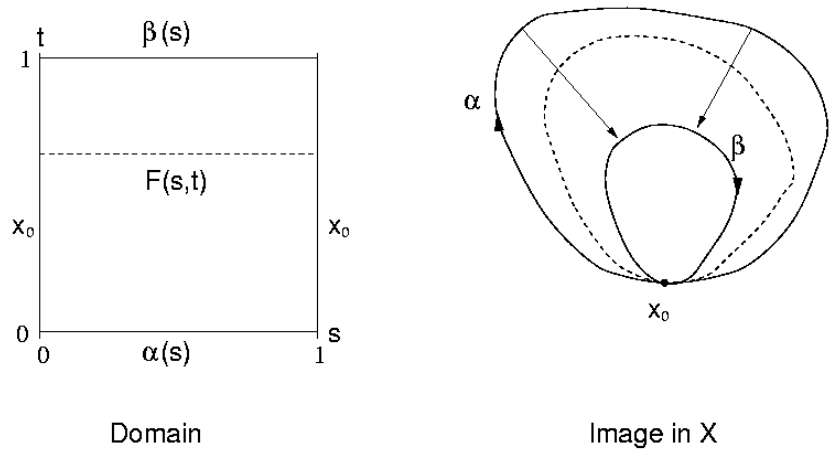

Figure 2.10: Graphical representation of a homotopy between loops α and β, showing both the domain and the image in X.

Exercise 26 *Prove that homotopy, $\alpha \sim \beta$, is an **equivalence relation**. Namely, that (i) $\alpha \sim \alpha$, (ii) $\alpha \sim \beta$ implies $\beta \sim \alpha$, and (iii) $\alpha \sim \beta$ and $\beta \sim \gamma$ implies $\alpha \sim \gamma$.*

With the concept of homotopy, we can now turn our set of loops into a group. The **homotopy class** $[\alpha]$ of a loop α consists of all loops that are homotopic to α. The multiplication $*$ between paths naturally gives a multiplication rule for homotopy classes: $[\alpha] * [\beta] \equiv [\alpha * \beta]$. We need to show that this multiplication is well defined, which means that it doesn't depend on the representative loop in the homotopy class. In other words, if $\alpha \sim \alpha'$ and $\beta \sim \beta'$, we need to show $\alpha * \beta \sim \alpha' * \beta'$.

First, assume $F(s,t)$ is a homotopy between α and α', and $G(s,t)$ is a homotopy between β and β'. We need to find $H(s,t)$ which gives a homotopy between $\alpha * \beta$ and $\alpha' * \beta'$. It is easiest to see how to proceed by manipulating the graphical representations of the domains, as shown in Figure 2.11. By putting the domains of F and G side by side and redefining the parameter s to keep it in $[0,1]$ we achieve our objective. Namely,

$$H(s,t) = \begin{cases} F(2s,t), & s \in [0, \frac{1}{2}] \\ G(2s-1,t), & s \in [\frac{1}{2}, 1] \end{cases} . \tag{2.35}$$

With this well defined multiplication, we can finally define our main object of interest.

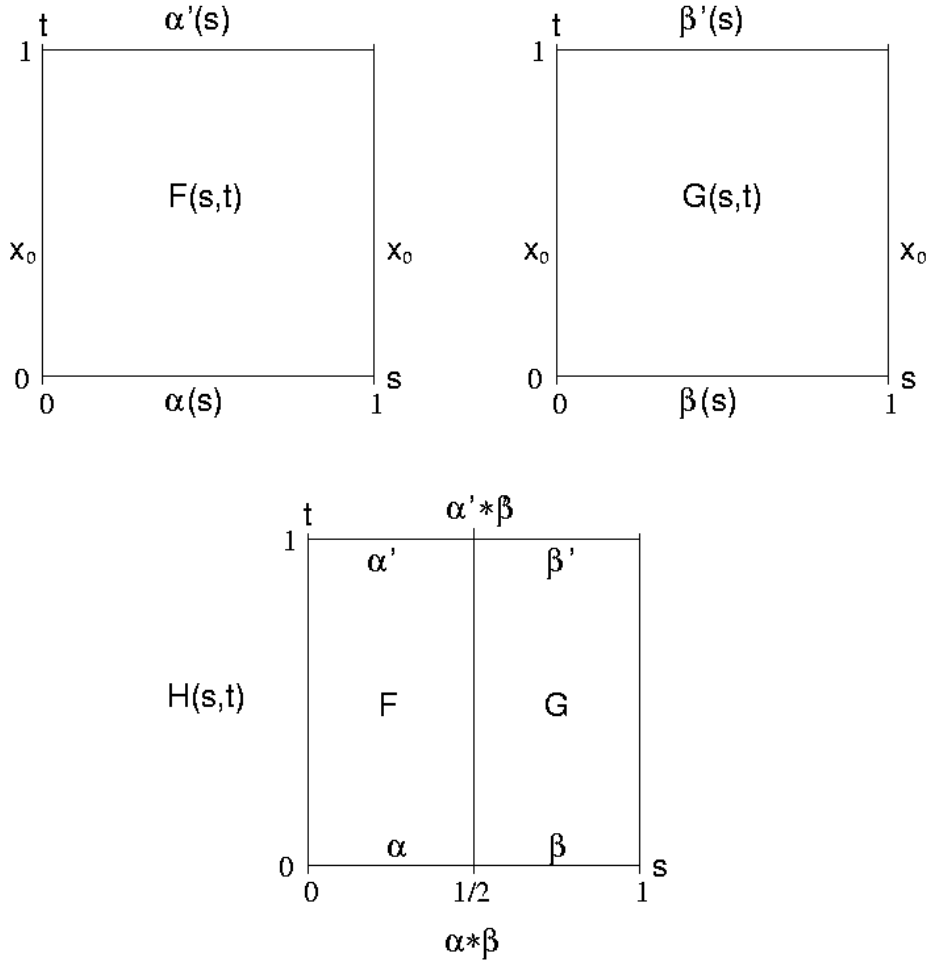

Figure 2.11: Given homotopies F between α and α' and G between β and β', we can put them "side-by-side" to produce the homotopy H between $\alpha * \beta$ and $\alpha' * \beta'$.

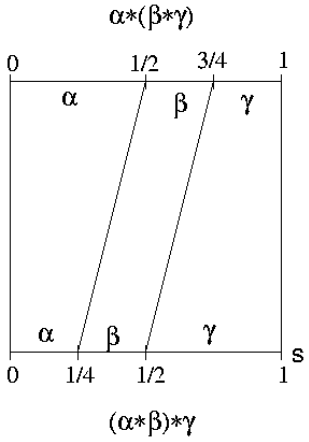

Figure 2.12: Graphical representation of the homotopy between $(\alpha * \beta) * \gamma$ and $\alpha * (\beta * \gamma)$.

The **Fundamental Group** or **First Homotopy Group** $\pi_1(X, x_0)$ is the set of homotopy classes $[\alpha]$ of loops in X with base point x_0 with multiplication defined by $*$: $[\alpha] * [\beta] = [\alpha * \beta]$.

Now we will prove that this is indeed a group. By restricting to loops with a fixed base point and further to their homotopy classes, we have arranged it so that the multiplication is well defined, as we just proved above. So the first step is to show associativity: $[\alpha] * ([\beta] * [\gamma]) = ([\alpha] * [\beta]) * [\gamma]$. In other words, we need to find a homotopy between $\alpha * (\beta * \gamma)$ and $(\alpha * \beta) * \gamma$. It is easiest to find our way graphically, as in Figure 2.12.

Exercise 27 *Write out the explicit form of $F(s,t)$ that yields the homotopy shown in the figure.*

The next step is to prove that c_{x_0} is the identity, $[c_{x_0}]*[\alpha] = [\alpha] = [\alpha]*[c_{x_0}]$. In other words, we need homotopies $c_{x_0} * \alpha \sim \alpha$ and $\alpha * c_{x_0} \sim \alpha$. Again, proceeding graphically is the best starting point, as shown in Figure 2.13.

Exercise 28 *Write out the homotopies G and G' explicitly in terms of s and t.*

Finally we need to prove the existence of inverses, namely $[\alpha]^{-1} = [\alpha^{-1}]$. We want $[\alpha] * [\alpha^{-1}] = [\alpha * \alpha^{-1}] = [c_{x_0}]$, which means we need to show

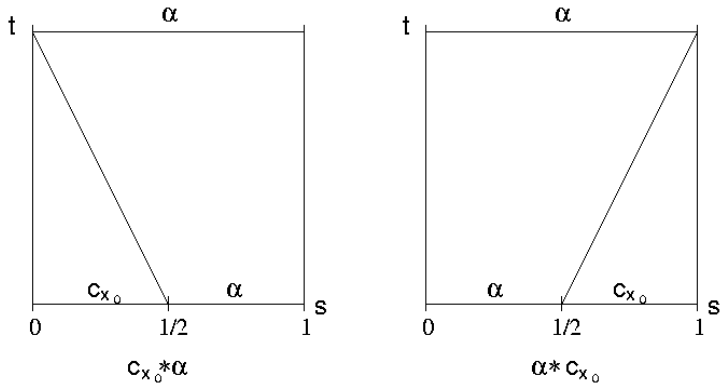

Figure 2.13: Graphical representation of the homotopies $c_{x_0} * \alpha \sim \alpha$ and $\alpha * c_{x_0} \sim \alpha$.

$\alpha * \alpha^{-1} \sim c_{x_0}$. Recalling that $\alpha^{-1}(s) = \alpha(1-s)$, we have

$$\alpha * \alpha^{-1}(s) = \begin{cases} \alpha(2s), & s \in [0, \frac{1}{2}] \\ \alpha^{-1}(2s-1) = \alpha(2-2s), & s \in [\frac{1}{2}, 1] \end{cases} , \tag{2.36}$$

whereas $c_{x_0}(s) = x_0 = \alpha(0)\ \forall s$. Thus we can define $H(s,t)$ as

$$H(s,t) = \begin{cases} \alpha(2s(1-t)), & s \in [0, \frac{1}{2}] \\ \alpha((2-2s)(1-t)), & s \in [\frac{1}{2}, 1] \end{cases} \tag{2.37}$$

to give the required homotopy.

Exercise 29 *Sketch a picture of the homotopy $H(s,t)$.*

Having proven all three properties, we conclude that $\pi_1(X, x_0)$ is a group!

Now we consider some examples of fundamental groups. First, take the setup for the Aharonov-Bohm effect, which was essentially the plane $\mathbb{R}^2$ with something special happening at the origin. Topologically, this is represented by the plane minus the origin, or $X = \mathbb{R}^2 - \{\mathbf{0}\}$. Clearly any loop that does not encircle the origin will be homotopic to the constant loop, in other words, contractible to a point. But a loop that goes around the origin cannot be contracted to a point because it can never cross over the origin, since the origin is not in the space X under consideration. If we take a loop α that encircles the origin once and multiply it by itself, we get a new loop $\alpha * \alpha$ that

encircles the origin twice. There is no way to deform $\alpha * \alpha$ to get back to α. Proceeding this way, we can form loops that encircle the origin any number of times. Winding a negative number of times is achieved by considering α^{-1}, which encircles the origin in the opposite direction. Thus we conclude that $\pi_1(\mathbb{R}^2 - \{\mathbf{0}\}) = \mathbb{Z}$. The integer corresponding to the number of times the loop wraps about the origin is called the **winding number**.

The next example is the circle, S^1. Considering loops that lie on the circle, we quickly see that the situation is nearly identical to the previous example, so conclude that $\pi_1(S^1) = \mathbb{Z}$.

What about $\mathbb{R}^3 - \{\mathbf{0}\}$? In this case, any loop in the xy-plane that encircles the origin can be lifted up in the z-direction and there contracted to a point. Thus $\pi_1(\mathbb{R}^3 - \{\mathbf{0}\}) = 0$, the trivial group. A space where the fundamental group is trivial said to be **simply connected**.

Now consider the sphere, S^2. If you imagine tying a string around an orange, you rapidly find that the string always falls off, which is directly related to the fact that the 2-sphere is simply connected, $\pi_1(S^2) = 0$. We can build on this example by considering $\mathbb{R}P^2$, the real projective plane, which is defined as the sphere S^2 with antipodal points identified. This is *not* a simply connected space, because a path β starting at the north pole and ending at the south pole is a closed loop, since the north and south poles are identified, so really the same point. But such a path cannot be continuously deformed into a constant path. (Try!) However, $\beta * \beta$ *is* homotopic to the constant path. Thus $\pi_1(\mathbb{R}P^2) = \mathbb{Z}_2$.

Exercise 30 *Show graphically that $\beta * \beta$ defined above is indeed homotopic to the constant loop in $\mathbb{R}P^2$.*

Finally, the previous two examples can be ratcheted up a dimension, and we can consider the 3-sphere S^3 and the real projective space $\mathbb{R}P^3$, defined as S^3 with antipodal points identified. You should convince yourself that the situation is completely analogous to the two-dimensional case.

Exercise 31 *We talked about $SU(2)$ and $SO(3)$ a lot in the first part of this tutorial. What are their fundamental groups?*

Exercise 32 *What is the fundamental group of the Möbius strip, the surface you get by taking a strip of paper, twisting it by half, and then gluing the ends together?*

Exercise 33 *Find $\pi_1(T^2)$, the two-dimensional torus with one hole. Is it Abelian? What is $\pi_1(T_2)$, the two-dimensional torus with two holes? Is it Abelian?*

In all of these examples we haven't mentioned the base point at all. For a space that is all one piece, it is pretty clear that it doesn't matter which base point you pick, the fundamental group will be the same. To be precise about being "all one piece" we need the concept of arcwise connectedness. A space X is arcwise connected if for all x_0 and x_1 in X there exists a path η such that $\eta(0) = x_0$ and $\eta(1) = x_1$. Then the statement about fundamental groups is the following: If X is arcwise connected, the $\pi_1(X, x_0)$ is isomorphic to $\pi_1(X, x_1)$.

To prove this statement, we consider a loop α with base point x_0. Then $\eta^{-1} * \alpha * \eta$ is a loop with base point x_1. We define the map $P_\eta : \pi_1(X, x_0) \to \pi_1(X, x_1)$ that takes $[\alpha] \mapsto [\eta^{-1} * \alpha * \eta]$.

Exercise 34 *Prove the map P_η is a group isomorphism.*

Thus if X is arcwise connected, we don't need to specify the base point.

2.2.2 Higher Homotopy Groups

The notation $\pi_1(X)$ suggests that perhaps $\pi_2(X)$ and $\pi_n(X)$ might exists, and indeed they do. We will generalize $\pi_1(X)$ to get the **higher homotopy groups** $\pi_n(X)$ for $n > 2$.

Since a closed loop α maps the interval $[0, 1]$ into X with $\alpha(0) = alpha(1) = x_0$, we can equally well consider the domain to be the circle, S^1, so $\alpha : S^1 \to X$. So the obvious generalization is to consider maps $\beta : S^2 \to X$ that map the sphere into X. More formally we will define "2-loops" β as

$$\beta : [0, 1] \times [0, 1] \to X \tag{2.38}$$

where $\beta(s_1, 0) = \beta(s_1, 1) = x_0$ and $\beta(0, s_2) = \beta(1, s_2) = x_0$. Now the domain for β is a unit square, where the entire boundary is mapped to a single point in X. This is shown in Figure 2.14. To make a group structure we first need to define multiplication of 2-loops, which is done just as for 1-loops, using the first variable. If α and β are 2-loops, then we define $\alpha * \beta$ by

$$\alpha * \beta = \begin{cases} \alpha(2s_1, s_2), & s_1 \in [0, \frac{1}{2}] \\ \beta(2s_1 - 1, s_2), & s_1 \in [\frac{1}{2}, 1] \end{cases} . \tag{2.39}$$

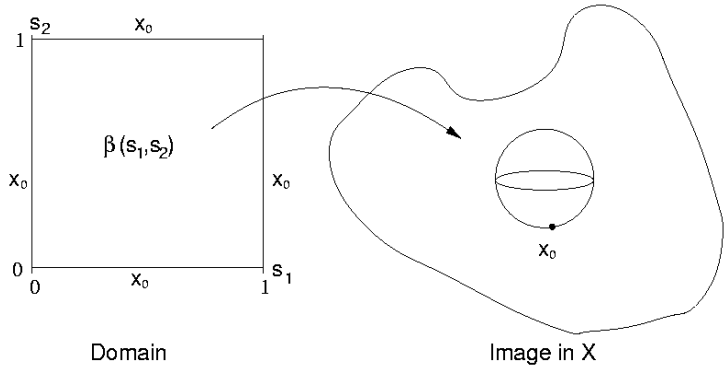

Figure 2.14: A 2-loop is a map of the unit square into X where the boundary of the square is mapped to the base point x_0.

Graphically, this amounts to squishing the domains of α and β together along the s_1-axis. This is shown in Figure 2.15.

For this multiplication to lead to a group structure we still need the concept of homotopy, which can be readily generalized from the one-dimensional case. If α and β are 2-loops, then we say they are homotopic, $\alpha \sim \beta$, if there exists a continuous map $F : [0,1] \times [0,1] \times [0,1] \to X$ such that $F(s_1, s_2, 1) = \beta(s_1, s_2)$, $F(s_1, s_2, 0) = \alpha(s_1, s_2)$, and $F(s_1, s_2, t) = x_0$ for all t and all (s_1, s_2) on the boundary of $[0,1] \times [0,1]$. Graphically this can be represented by a cube where the bottom face is the domain of the 2-loop α and the top face is the domain of β, and the interior represents the continuous deformation of α into β. This is shown in Figure 2.16.

Exercise 35 *Define the 2-loop $\alpha^{-1}(s_1, s_2)$ and show that it behaves as it should.*

Putting these pieces together, the homotopy classes of 2-loops in X define the group $\pi_2(X, x_0)$.

Now for some examples. Sidney Coleman gives a nice argument that $\pi_2(S^2) \neq 0$: "You cannot peel an orange without breaking the skin."[2] The point is that the flesh of the orange is the space $X = S^2$ and the skin represents a map of $S^2 \to X$. If this map were homotopic to the identity, you could push the skin around and deform it into a single point and then just

[2]Coleman, S. *Aspects of Symmetry*, Cambridge University Press, 1995, p. 208.

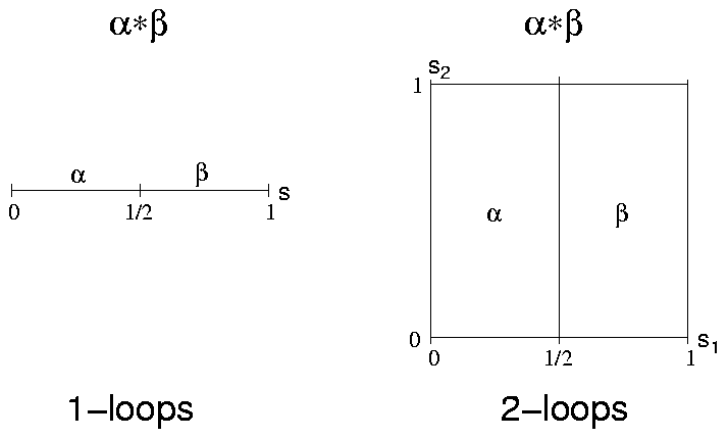

Figure 2.15: The product of 1-loops and 2-loops both reparametrize the domain of one variable to first traverse one loop and then the other.

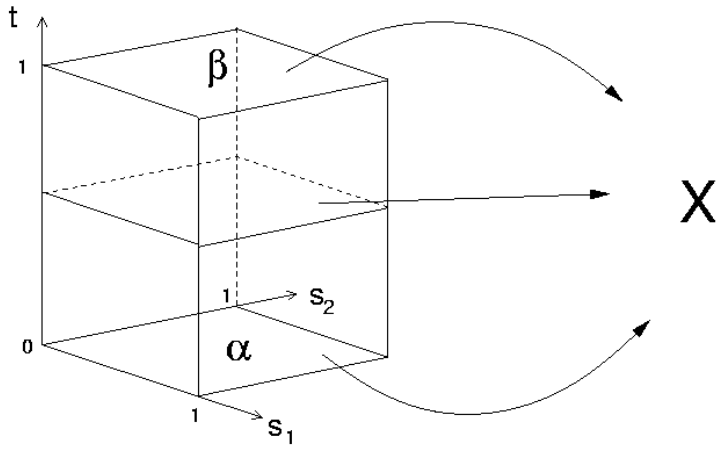

Figure 2.16: The homotopy between the 2-loops α and β can be represented by a cube where each horizontal slice represents a map of S^2 into X, where the bottom slice is α and the top slice is β.

lift it off. Since this is clearly impossible, there is at least one nontrivial homotopy class in $\pi_2(S^2)$. In fact, there is also a "winding number" associated with maps of the sphere into itself, much as there was for maps of the circle into itself. This leads to the result that $\pi_2(S^2) = \mathbb{Z}$.

The higher homotopy groups are, not surprisingly, more difficult to visualize. However, you should be able to convince yourself that $\pi_2(S^1) = 0$. Considering lower dimensional analogies, you can also show that $\pi_2(S^3) = 0$.

This construction can be easily generalized to n-loops, $\alpha : S^n \to X$, and their homotopy classes define the group $\pi_n(X)$. I will leave this as an exercise.

Exercise 36 *Formally define $\pi_n(X)$. That is, define n-loops, their multiplication and homotopy, and show that in an arcwise connected space $\pi_n(X)$ is independent of the base point.*

Exercise 37 *Prove that for 2-loops α and β, $[\alpha] * [\beta] \sim [\beta] * [\alpha]$. That is, prove that $\pi_2(X)$ is Abelian. Can you generalize this result to $\pi_n(X)$?*

Exercise 38 *How would you define $\pi_0(X)$?*

Exercise 39 *Determine $\pi_n(S^m)$ for as many combinations of n and m as you can.*

So how are these homotopy groups useful in physics? One place they appear is when considering gauge field configurations in quantum field theory, which is where we turn now.

2.3 Quantum Field Theory

What is **quantum field theory** (QFT)? There are many ways to answer that question, so here I give two of them.

- The unique framework that reconciles quantum mechanics with special relativity.
- A framework for computing the abundance of various products produced by collisions between elementary particles.

Notice that I call QFT a *framework* rather than a theory. Quantum electrodynamics (QED) is a specific theory that describes the interactions of electrons and photons using the *language* of QFT. Similarly, quantum chromodynamics (QCD) is the theory of quarks and gluons (the constituents of protons and neutrons) using QFT. QED and QCD are theories that describe specific systems or aspects of reality, whereas quantum field theory is the language or framework that they use. Thus just like classical mechanics or quantum mechanics, QFT is a formalism that can be applied to lots of different systems. It isn't only high energy physics that uses QFT; condensed matter systems, like superconductors, are also described by QFT.

So what does QFT look like mathematically and how does it work conceptually? Quantum field theory is essentially quantum mechanics on steroids. This is easiest to see when using the path integral formulation of both frameworks.

Recall that the path integral formulation of quantum mechanics gives the amplitude for a transition from an initial state $\mathbf{x}_i$ to a final state $\mathbf{x}_f$ in terms of a sum over all paths connecting those two states:

$$\mathcal{A}(i \to f) = \langle \mathbf{x}_f, t_f | \mathbf{x}_i, t_i \rangle = \int \mathcal{D}[\mathbf{x}(t)] \; e^{\frac{i}{\hbar} S[\mathbf{x}(t)]}. \tag{2.40}$$

This is a path integral with three coordinates, x, y, and z, representing the position of the particle we're studying. We can generalize this slightly by adding another particle and another three coordinates. The resulting path integral is

$$\langle \mathbf{x}_{1f}, \mathbf{x}_{2f}, t_f | \mathbf{x}_{1i}, \mathbf{x}_{2i}, t_i \rangle = \int \mathcal{D}[\mathbf{x}_1(t)] \mathcal{D}[\mathbf{x}_2(t)] \; e^{\frac{i}{\hbar} S[\mathbf{x}_1(t), \mathbf{x}_2(t)]}, \tag{2.41}$$

which depends on a new action $S[\mathbf{x}_1(t), \mathbf{x}_2(t)]$ which is a functional of both paths $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$.

To go to QFT we replace the quantum mechanical coordinates $\mathbf{x}(t)$ and with **fields** which are functions defined over all space, $\phi(\mathbf{x}, t)$. Now the location $\mathbf{x}$ is no longer a coordinate itself, but rather a label for the infinite number of coordinates $\phi(\mathbf{x})$. In QFT the "coordinates" in the path integral are the values of the field at each point in space, so $\phi(\mathbf{x}_0)$, $\phi(\mathbf{x}_1)$, $\phi(\mathbf{x}_2)$ are three of the infinite number of coordinates that will be summed over in our path integral.

Very roughly you can think of $\phi(\mathbf{x})$ as the wavefunction for a ϕ particles, so $|\phi(\mathbf{x}_0)|^2$ would be the probability of finding the ϕ particle at position $\mathbf{x}_0$.

However, the field $\phi(\mathbf{x})$ is not restricted to having total probability equal to one. As a field it is allowed to represent the presence of 0, 1, 2 or many ϕ particles at different places in the universe. This is one of the important features of QFT, namely the ability to deal with systems where the total number of particles is not conserved. That is something difficult to handle in regular quantum mechanics.

In the quantum mechanical path integral we summed over all paths $\mathbf{x}(t)$, which meant summing over all values for our coordinates $\mathbf{x}$. In QFT we do the same thing, summing over all values of the coordinates. This means summing over all "field configurations" between some initial configuration, $\phi_i(\mathbf{x}, t_i)$ and final configuration $\phi_f(\mathbf{x}, t_f)$. The corresponding path integral is written

$$\langle \phi_f(\mathbf{x}, t_f) | \phi_i(\mathbf{x}, t_i) \rangle = \int \mathcal{D}[\phi] \; e^{\frac{i}{\hbar} S[\phi]}. \tag{2.42}$$

Here $S[\phi]$ is still called the action and is a functional that takes a "path" in field configuration space and returns a real number. Recall that in quantum mechanics the action for a free particle was

$$S_0^{\mathrm{QM}}[\mathbf{x}(t)] = \int_{t_i}^{t_f} dt \; \frac{1}{2} m \left(\frac{d\mathbf{x}}{dt} \right)^2 = \int_{t_i}^{t_f} dt \; \frac{1}{2} m \left[\left(\frac{dx}{dt} \right)^2 + \left(\frac{dy}{dt} \right)^2 + \left(\frac{dz}{dt} \right)^2 \right]. \tag{2.43}$$

Notice that the action sums up contributions from all three coordinates x, y, and z. In QFT we have a similar action for a single, spinless point particle,

$$\begin{aligned} S_0^{\mathrm{QFT}}[\phi] &= \int_{t_i}^{t_f} dt \int d^3\mathbf{x} \left[\left| \frac{\partial \phi}{\partial x^\mu} \right|^2 - m^2 |\phi|^2 \right] && (2.44) \\ &= \int_{t_i}^{t_f} dt \int d^3\mathbf{x} \left[\left| \frac{\partial \phi}{\partial t} \right|^2 - \left| \frac{\partial \phi}{\partial x} \right|^2 - \left| \frac{\partial \phi}{\partial y} \right|^2 - \left| \frac{\partial \phi}{\partial z} \right|^2 - m^2 |\phi|^2 \right] && (2.45) \end{aligned}$$

where the second line expands the shorthand notation of $\partial/\partial x^\mu$.[3] So knowing $\phi(\mathbf{x}, t)$ you can take derivatives and do the integrals and arrive at a real number $S[\phi]$. In practice this isn't done explicitly, but it is useful to know that it wouldn't be too hard to do.

[3]The index μ runs over 0, 1, 2, 3 in order to label the components of four-vectors that are used in special relativity. For instance, $x^\mu = (x^0, x^1, x^2, x^3) = (ct, x, y, z)$ where c is the speed of light.

The path integral in QFT tells you the amplitude for getting from one specified inital field configuration to a final field configuration. In quantum mechanics the path $x(t)$ was a function that could take on any values between its fixed endpoints, and shown in Figure 2.3. In QFT there is a similar "path" at each point in space. For fixed $\mathbf{x}_0$, $\phi(\mathbf{x}_0, t)$ is a function of time which can take on any values between fixed endpoints. The QFT path integral sums over the infinity of adjacent paths $\phi(\mathbf{x}_0, t)$, $\phi(\mathbf{x}_1, t)$, etc. Another way to visualize this sum over field configurations is by thinking of $\phi(\mathbf{x}, t_i)$ for fixed t_i as a surface lying over the domain $\mathbf{x}$. As time evolves, that surface can change in many ways until it reaches a specified final surface $\phi_f(\mathbf{x}, t_f)$. The QFT path integral sums over all different sets of surfaces that connect the initial and final surfaces, or field configurations.

Unfortunately, the discussion at this level is as much metaphysics as it is math or physics. In order to learn the details of how to extract meaningful results from path integrals you will need to either dive into a textbook or take a course on the subject.

2.3.1 Symmetries in Quantum Field Theory

The reason for this qualitative discussion of the path integral in quantum field theory is that it gives the easiest way to see how symmetries manifest themselves in QFT. Since physical results are computed from the path integral shown in Eqn. (2.42), any transformations on the fields ϕ that leave the path integral unchanged will not change those results. For the path integral to remain unchanged any transformation must not change the action, $S[\phi]$, or the measure, $\mathcal{D}[\phi]$. We will be concerned with unitary transformations of the fields ϕ which leave the measure invariant. There are important cases where only the measure is changed by an apparent symmetry, but we will not pursue that direction. Thus we will assume that the measure does not change and look for symmetries of the action.

We start by considering the action of a single, complex, scalar field $\phi(x)$ as shown in Eqn. (2.44). We recall that in quantum mechanics the overall phase of the wavefunction didn't have any physical meaning. The same thing is true here, where we say that the $S[\phi]$ has a "global $U(1)$ invariance". $U(1)$ is the set of 1×1 unitary matrices, namely complex numbers with magnitude 1 which form the unit circle in the complex plane. The $U(1)$ invariance means that if we take $\phi(x) \to \phi'(x) = e^{i\alpha}\phi(x)$ the action is unchanged. You can readily check that $S[\phi'] = S[\phi]$ because the $e^{i\alpha}$ and its complex conjugate

both appear in each term of the action. Thus our single, complex scalar field exhibits a $U(1)$ symmetry.

Now, earlier I said that QFT reconciles special relativity and quantum mechanics. One of the things that special relativity says is that no information can travel faster than light. Thus we might be a little suspicious about a transformation $\phi(x) \to e^{i\alpha}\phi(x)$ that changes the phase of $\phi(x)$ *everywhere* by exactly the same amount at exactly the same time. So perhaps we should consider changing the phase of ϕ in a way that can be different at different places. This is called a **local symmetry**, and is achieved by taking $\phi(x) \to \phi'(x) = e^{i\alpha(x)}\phi(x)$, where the phase change $\alpha(x)$ is itself a function of space. But is such a transformation a symmetry of the action?

The second term, $m^2\phi^\dagger\phi$ is clearly unaffected by the local change in phase since $m^2 |\phi'|^2 = m^2\phi'^\dagger\phi' = m^2 e^{-i\alpha(x)}\phi^\dagger e^{i\alpha(x)}\phi = m^2 |\phi|^2$. However, the fist term is problematic:

$$\frac{\partial}{\partial x^\mu} e^{i\alpha(x)}\phi(x) = i\frac{\partial\alpha(x)}{\partial x^\mu} e^{i\alpha(x)}\phi(x) + e^{i\alpha(x)}\frac{\partial\phi(x)}{\partial x^\mu}. \tag{2.46}$$

Thus we see that there is an extra term coming from the derivative acting on $\alpha(x)$, so when the above result is multiplied by its Hermitian conjugate, we will definitely have extra contributions that weren't present in the action initially.

Thus as it stands this local transformation is simply not a symmetry of the action. We could stop there and go home, but it proves to be useful to find a way to patch things up and modify the action so that it is invariant under the local $U(1)$ transformation. The way to do that is to add a new vector field $A_\mu(x)$ that transforms in a different way when the phase of ϕ is changed, namely

$$A_\mu(x) \to A'_\mu(x) = A_\mu(x) - \frac{1}{e}\frac{\partial\alpha(x)}{\partial x^\mu}, \tag{2.47}$$

where e is a real number called the **coupling constant** that represents the strength of the interaction between the fields $A(x)$ and $\phi(x)$. Now we write a new action

$$S[\phi, A] = \int dt \int d^3\mathbf{x} \left| \left(\frac{\partial}{\partial x^\mu} + ieA_\mu \right) \phi \right|^2 - m^2 |\phi|^2 \tag{2.48}$$

Now the first term in the action is invariant under the local $U(1)$ transfor-

mation, since

$$\left(\frac{\partial}{\partial x^\mu} + ieA'_\mu\right)\phi' = \left(\frac{\partial}{\partial x^\mu} + ie\left(A_\mu - \frac{1}{e}\frac{\partial\alpha}{\partial x^\mu}\right)\right)e^{i\alpha(x)}\phi \quad (2.49)$$

$$= e^{i\alpha(x)}\left(\frac{\partial}{\partial x^\mu} + ieA_\mu\right)\phi \quad (2.50)$$

so when taking the magnitude-square of this quantity the phase $e^{i\alpha(x)}$ drops out.

So what just happened? By requiring that our action be invariant under the local $U(1)$ transformation, we were forced to introduce a new vector field $A_\mu(x)$, called a **gauge field**, that transforms as in Eqn. (2.47). This field is a four-vector, $A^\mu = (A_0, A_x, A_y, A_z) = (\varphi/c, \mathbf{A})$, where φ (not to be confused with our field ϕ) is the electric potential and $\mathbf{A}$ is the magnetic vector potential that we encountered earlier when discussing electromagnetism. Notice that the transformation in Eqn. (2.47) gives $\mathbf{A} \to \mathbf{A'} = \mathbf{A} - \frac{1}{e}\nabla\alpha(x)$, which is exactly the gauge transformation for $\mathbf{A}$ discussed earlier. Thus the local $U(1)$ invariance is really a restatement of the gauge symmetry of electromagnetism! The electromagnetic potential $A_\mu = (\varphi/c, \mathbf{A})$ appears as a gauge field, whose small fluctuations, or "quanta", are interpreted as photons. There is a deep geometric meaning to the gauge fields (as connections on fiber bundles) but unfortunately we don't have time to get into that.

Since we have uncovered electromagnetism, you might wonder where $\mathbf{E}$ and $\mathbf{B}$ come in. They enter the action as a new term that depends only on A_μ and represents the "kinetic energy" of the electric and magnetic fields, which now we would call the gauge fields. This is written as an antisymmetric matrix F defined by

$$F_{\mu\nu} = \frac{\partial A_\nu}{\partial x^\mu} - \frac{\partial A_\mu}{\partial x^\nu} = \begin{pmatrix} 0 & E_x/c & E_y/c & E_z/c \\ -E_x/c & 0 & -B_z & B_y \\ -E_y/c & B_z & 0 & -B_x \\ -E_z/c & -B_y & B_x & 0 \end{pmatrix}. \quad (2.51)$$

The new term in the action is

$$S_0[A] = \int dt \int d^3\mathbf{x}\left(-\frac{1}{4}F_{\mu\nu}F^{\mu\nu}\right) = \int dt \int d^3\mathbf{x}\left(\frac{1}{2c^2}\mathbf{E}^2 - \frac{1}{2}\mathbf{B}^2\right) \quad (2.52)$$

Exercise 40 *Verify that $F_{\mu\nu}$ is gauge invariant, namely that it doesn't change under the transformation in Eqn. (2.47). (You can do this even without understanding the 4-vector notation.)*

2.3.2 General Gauge Theory

Now comes the time for us to connect some of what we learned early on about Lie groups and Lie algebras to quantum field theory actions. A general, non-Abelian gauge theory is specified by a Lie group G, (such as $SO(3)$, $SU(2)$, or $SU(n)$) and a set of fields $\vec{\phi} = (\phi_1, \phi_2, \ldots, \phi_n)$ which form an n-component vector. We start with an action that is similar to Eqn. (2.44), except now there is a standard dot product between the vector $\vec{\phi}$ and it's Hermitian conjugate, $\vec{\phi}^\dagger$:

$$S[\vec{\phi}] = \int dt \int d^3\mathbf{x} \left[\left| \frac{\partial \vec{\phi}}{\partial x^\mu} \right|^2 - m^2 \left| \vec{\phi} \right|^2 \right] \tag{2.53}$$

We assume that the vector space of the fields $\vec{\phi}$ forms an n-dimensional, unitary representation of G. That means that there is a map $T : G \to T(G)$ that take $g \mapsto T(g)$ where $T(g)$ is a unitary matrix acting on the vectors space of $\vec{\phi}$. Thus an element $g \in G$ changes the fields ϕ by

$$\vec{\phi}(x) \to T(g)\vec{\phi}(x) \qquad \text{i.e.} \qquad \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_n \end{pmatrix} \to \begin{pmatrix} \\ \quad T(g) \quad \\ \\ \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_n \end{pmatrix} \tag{2.54}$$

If $T(g)$ is independent of the position in space, then the terms in Eqn. (2.53) is each invariant. For the second term we have

$$\left|\vec{\phi}'\right|^2 = \vec{\phi}'^\dagger \vec{\phi}' = \left(T(g)\vec{\phi}\right)^\dagger \left(T(g)\vec{\phi}\right) = \vec{\phi}^\dagger T^\dagger(g) T(g) \vec{\phi} = \vec{\phi}^\dagger \vec{\phi} = \left|\vec{\phi}\right|^2 . \tag{2.55}$$

Exercise 41 *Show that the first term in Eqn. (2.53) is also invariant under the transformation $\vec{\phi} \to T(g)\vec{\phi}$.*

However, just as we did a local phase transformation $\phi \to e^{i\alpha(x)}\phi$ where the element in $U(1)$ was a function of space, we can also consider a map $g(x) : \mathbb{R}^3 \to G$ which assigns an element $g(x) \in G$ to each point x of space.[4] Then $\vec{\phi}(x) \to \vec{\phi}'(x) = T(g(x))\vec{\phi}(x)$ is a local or gauge transformation of $\vec{\phi}$.

Clearly the derivative term in Eqn. (2.53) will no longer be invariant under this local transformation, so again we need to introduce gauge fields A_μ to

[4] More properly we should be considering a map from 4-dimensional Minkowski space-time into G, but this detail is unimportant for the general analysis under consideration.

preserve the symmetry. However, this time they need to be proportional to elements of the Lie algebra of G. So instead of $\partial\vec{\phi}/\partial x^\mu$ we use

$$\left(\frac{\partial}{\partial x^\mu} + ieA_\mu\right)\vec{\phi} \qquad \text{where} \qquad A_\mu = A^a_\mu(x)T(J^a). \tag{2.56}$$

In the above equation $A^a_\mu(x)$ are vector functions, while $T(J^a)$ form a $n \times n$ matrix representation of the generators J^a of the Lie algebra $\mathfrak{g}$.

The action for the gauge fields written as in Eqn. (2.52), but now with

$$F_{\mu\nu} = \frac{\partial A_\nu}{\partial x^\mu} - \frac{\partial A_\mu}{\partial x^\nu} + ie[A_\mu, A_\nu], \tag{2.57}$$

which is a $n \times n$ matrix and requires a trace to give a scalar. We also no longer have the interpretation in terms of $\mathbf{E}$ and $\mathbf{B}$ since that is specific to electromagnetism. So our full action is now

$$S[\phi, A] = \int dt \int d^3\mathbf{x} \left| \left(\frac{\partial}{\partial x^\mu} + ieA_\mu\right)\phi \right|^2 - m^2 |\phi|^2 - \frac{1}{4}\mathrm{Tr}(F_{\mu\nu}F^{\mu\nu}). \tag{2.58}$$

Under a gauge transformation the various fields transform as:

$$\vec{\phi}(x) \rightarrow \vec{\phi}'(x) = T(g(x))\vec{\phi}(x) \tag{2.59}$$

$$A_\mu(x) \rightarrow A'_\mu(x) = T(g(x))A_\mu(x)T^\dagger(g(x)) - \frac{i}{e}T(g(x))\frac{\partial}{\partial x^\mu}\left(T^\dagger(g(x))\right) \tag{2.60}$$

Exercise 42 *Use the second transformation and the definition of $F_{\mu\nu}$ to show that F transforms like $F_{\mu\nu}(x) \rightarrow F'_{\mu\nu}(x) = T(g(x))F_{\mu\nu}(x)T^\dagger(g(x))$.*

Exercise 43 *Show that the action in Eqn. (2.58) is invariant in under the above gauge transformations.*

This is great! We've connected the earlier ideas of symmetries, Lie groups, Lie algebras, and their representations to the path integral of quantum field theory, which is just a glorified version of quantum mechanics. Now all we need to do to "close the circle" is tie in the recent results we discussed about homotopy theory.

2.4 The Topology of Gauge Field Configurations

We will now simplify things somewhat by considering a "pure gauge theory". This is just like thinking about electromagnetic waves, which are electric and magnetic fields propagating through space without any charges or currents nearby. In the context of a general, non-Abelian gauge theory this situation is described by the action

$$S[A] = \int dt \int d^3\mathbf{x} - \frac{1}{4}\mathrm{Tr}(F_{\mu\nu}F^{\mu\nu}). \tag{2.61}$$

It turns out to be useful to study configurations of finite action, because near those configurations one can do classical approximations. The above action will be finite as long as $F_{\mu\nu}(x)$ goes to zero rapidly enough as $x \to \infty$. Clearly this will be the case if $A_\mu(x)$ falls to zero at infinity, but it will also be true if A_μ approaches a gauge transformation of zero, namely

$$\lim_{x\to\infty} A_\mu(x) = T(g(\theta))\frac{\partial}{\partial x^\mu}\left(T^\dagger(g(\theta))\right) \Rightarrow \lim_{x\to\infty} F_{\mu\nu}(x) = 0. \tag{2.62}$$

Here the function $g(\theta)$ represents a map from the boundary of the space at infinity into G, so it depends only on angles, represented schematically by θ. So to categorize the field configurations of finite action we need to categorize the maps $g(\theta)$.

But can't we just do a local gauge transformation $h(x)$ and turn all those configurations into the trivial configuration with $A_\mu = 0$ everywhere? The point is that the map $g(\theta)$ is only defined on the boundary of space, but a gauge transformation $h(x)$ must be continuous throughout the interior of the space as well. We can think of such a map as being defined on consecutive shells of smaller and smaller radii. If we choose $h(x)$ such that it "undoes" $g(\theta)$ and transforms A_μ to zero at the boundary (infinity), we still need $h(x)$ to be continuous on each of the smaller shells and eventually become a constant at the origin. This is another way of saying that $h(x)$ restricted to the surface at infinity must be homotopic to $h(x)$ at the origin, which must be the constant map. The conclusion is that the set of gauge configurations of finite action is equivalent to the homotopy classes of maps from the boundary at infinity into G.

We already saw this situation arise in our study of the Aharonov-Bohm effect. There we had a system in two dimensions and our theory was electromagnetism, which is a $U(1)$ gauge theory. The boundary at infinity in this situation is essentially a circle, S^1, so the question is whether there are nontrivial maps $g : S^1 \to U(1)$. Since $U(1)$ can be thought of the unit circle in the complex plane, it is topologically the same as S^1, so the collection of gauge configurations with finite action are classified by $\pi_1(S^1) = \mathbb{Z}$. This tells us that there are indeed nontrivial gauge configurations where $\mathbf{A}$ is a gauge transformation of zero (also called **pure gauge**) at infinity but cannot be continuously deformed to be pure gauge everywhere in the interior of the space. Exercise 25 exhibited a configuration where $\nabla \times \mathbf{A} = 0$ everywhere outside the solenoid. By writing $\mathbf{A} = \nabla V$ for $V = BR^2\phi/2$, where ϕ is the azimuthal angle in cylindrical coordinates, we see explicitly that this is a "pure gauge" configuration. The integral of ∇V around a closed path is proportional to the difference in angle ϕ between the beginning and ending of the path, which comes in integral multiples of 2π. That integer is the winding number for the gauge configuration, which labels the distinct homotopy classes.

We can now ask about a similar "Aharonov-Bohm" type effect in other, more complicated gauge theories. For example, let G be $SU(2)$ and consider a theory in four-dimensional spacetime. The surface at infinity is now a 3-sphere, S^3, so we are interested in maps $g : S^3 \to SU(2)$, i.e. $\pi_3(SU(2))$. Since we saw that $SU(2)$ can be thought of as S^3, this is the same as $\pi_3(S^3)$. Generalizing the results for spheres in fewer dimensions, one can show that $\pi_3(S^3) = \mathbb{Z}$. So there should be an "Aharonov-Bohm" effect and an associated winding number for an $SU(2)$ gauge theory in 4D spacetime. In fact, this result is more general. A theorem by Raoul Bott (who taught the topology course I took as an undergraduate) states that for any simple Lie group G, any continuous mapping of $S^3 \to G$ can be continuously deformed into a mapping into an $SU(2)$ subgroup of G. Thus $\pi_3(G) = \mathbb{Z}$ for *any* simple Lie group G.

The third homotopy group, $\pi_3(G)$, is also important in the study of gauge theory vacua. The situation appears similar to that discussed above, but is somewhat different. In this situation we want to study the vacuum of gauge theories, which is the state where all fields vanish everywhere. For gauge fields, however, the redundancy that follows from gauge transformations means that they need not vanish, only that they be gauge transformations of vanishing fields. Even after we "fix the gauge", the equivalent to

requiring $\nabla \cdot \mathbf{A} = 0$ in electromagnetism, we are left with a residual ambiguity which allows the spatial components of pure gauge configurations to depend on spatial location

$$A_i(\mathbf{x}) = T(g(\mathbf{x}))\frac{\partial}{\partial x^\mu}\left(T^\dagger(g(\mathbf{x}))\right). \tag{2.63}$$

However, in order to have a well defined non-Abelian "charge", analogous to the electric charge, the gauge transformations must approach a constant at spatial infinity,

$$\lim_{|\mathbf{x}|\to\infty} T(g(\mathbf{x})) = T(g_\infty). \tag{2.64}$$

Thus the possibilities for distinct $g(\mathbf{x})$ as a function of three spatial dimensions with the value at spatial infinity fixed is again given by $\pi_3(G) = \mathbb{Z}$ for any simple Lie group G, because a three dimensional space with the "boundary" identified gives S^3. So we know that such vacuum configurations can be labeled by the winding number n, and we will denote them as states $|n\rangle$.

However, there exist so-called **large gauge transformations** Ω that can change from a vacuum with winding number n to one with winding number $n+1$, namely $\Omega|n\rangle = |n+1\rangle$. Since we expect the "true vacuum" to be gauge invariant, in particular it should be invariant under transformations Ω. Clearly the states $|n\rangle$ are not invariant, but we can construct a superposition of those vacua that is. It is called the theta-vacuum and is defined by

$$|\theta\rangle = \sum_n e^{-in\theta}|n\rangle. \tag{2.65}$$

Operating on $|\theta\rangle$ with Ω gives

$$\Omega|\theta\rangle = \sum_n e^{-in\theta}\Omega|n\rangle = \sum_n e^{-in\theta}|n+1\rangle = e^{i\theta}\sum_{n+1} e^{-i(n+1)\theta}|n+1\rangle = e^{i\theta}|\theta\rangle, \tag{2.66}$$

demonstrating that $|\theta\rangle$ is an eigenstate of Ω. The existence of this θ-vacuum leads to "non-perturbative" effects and possible CP-violation, but that is a story for another tutorial.

Bibliography

[1] Artin, Michael, *Algebra*.

[2] Cahn, Robert, *Semi-Simple Lie Algebras and Their Representations*.

[3] Coleman, Sidney, *Aspects of Symmetry*.

[4] Fulton, William and Joe Harris, *Representation Theory: A First Course*.

[5] Georgi, Howard, *Lie Algebras in Particle Physics*.

[6] Griffiths, David J., *Introduction to Quantum Mechanics*.

[7] Munkres, James R., *Topology: A First Course*.

[8] Nakahara, Mikio, *Geometry, Topology and Physics*.

[9] Sakurai, J. J., *Modern Quantum Mechanics*.

[10] Schutz, Bernard, *Geometrical Methods of Mathematical Physics*.

[11] Shankar, R., *Principles of Quantum Mechanics*.

[12] Weinberg, S., *The Quantum Theory of Fields, Vol. I*.