Representation Theory and Physics
SHP Fall ’19

Henry Liu

Last updated: December 14, 2019

Contents

0 Symmetries 3

1 Groups 4
1.1 Definitions and first examples 4
1.2 The dihedral group 6
1.3 The symmetric group 8
1.4 Homomorphisms 9
1.5 Operations on groups 11
1.6 Lagrange’s theorem 13
1.7 Classification 15

2 Representations of groups 16
2.1 Linear algebra 18
2.2 Definitions and first examples 22
2.3 Sums and reducibility 24
2.4 Morphisms 26
2.5 Tensors and duals 28

3 Quantum mechanics 30
3.1 Quantum states 30
3.2 Symmetries 32
3.3 Observation 34
3.4 Entanglement 36
3.5 Time evolution 37

4 Lie groups and their representations 39
4.1 Matrix Lie groups 40
4.2 Tangent spaces 41
4.3 Lie algebras 43
4.4 $\mathrm{SU}(2)$ and $\mathrm{SL}(2,\mathbb{C})$ 45
4.5 Semisimple theory 49

5 Particle physics **52**
5.1 Spacetime symmetries 52
5.2 Spin: fermions and bosons 54
5.3 Gauge theories 55
5.4 The standard model 58

6 Beyond the standard model **60**
6.1 Perturbation theory and renormalization 60
6.2 Grand unification 64
6.3 Supersymmetry 66

0 Symmetries

Many mathematical objects and physical systems possess *symmetries.* A circle stays the same no matter how it is rotated; a rotation by θ for any angle θ is therefore a symmetry of the circle. On the other hand, a square stays the same only under rotation by multiples of π. From this simple example we see that, broadly, symmetries should be separated into two types.

1. The rotation symmetry of the circle is **continuous**: one can start with the unrotated circle and apply a given rotation by θ in a continuous fashion, without affecting the circle.

2. The rotation symmetry of the square is **discrete**: one *cannot* get from an unrotated square to a square rotated by (some multiple of) π in a continuous fashion.

In real life, common continuous symmetries include translations and rotations. Discrete symmetries are less obvious, but include time reversal (flipping the arrow of time), charge conjugation (swapping what we call positive vs negative charge), and translations in lattices (like for crystals/metals). It is important to study both continuous and discrete symmetries. The study of symmetry, in mathematics, is called **representation theory**.

Once we understand the symmetries of an object, the powerful machinery of representation theory kicks in and allows us to draw marvelous conclusions about the object itself. This is especially useful in physics, where often the symmetries are more obvious/intuitive than whatever conclusions we draw from them.

Example 0.1. The three-dimensional space we live in has translation and rotation symmetries. Then Noether's theorem, which we will see later, immediately implies the conservation of momentum and energy. Together with reflection symmetries, these symmetries form what is called the "Euclidean group" of symmetries of three-dimensional space.

Example 0.2. Three-dimensional space belongs to four-dimensional *spacetime.* In spacetime, it turns out there are additional symmetries which mix space and time called "Lorentz transformations". The statement that spacetime has these extra symmetries is the *only* postulate underlying the entire theory of special relativity. Putting the Lorentz transformations together with the usual Euclidean symmetries gives the "Poincaré group" of symmetries of four-dimensional spacetime.

Note that all these symmetries we just stated are continuous symmetries. Indeed, because many fundamental objects in physics are *continuous* objects, many of the interesting applications of representation theory to physics involve continuous symmetries. However continuous symmetries are more difficult to study than discrete symmetries. Hence we will begin with discrete symmetries, which are slightly less physically relevant, in order to familiarize ourselves with the basic objects of representation theory.

1 Groups

The first step in representation theory is to understand the structure of the set of symmetries of a given object. This set, which we'll call G, has some very special structure, which we'll discuss abstractly now. First, if g_1 and g_2 are two symmetries in G, then

> applying g_1, then applying g_2, is itself a symmetry of the object.

We'll denote this composite symmetry by $g_2 g_1$. (In this notation, we apply symmetries from right to left, e.g. g_1 is applied first. This is just a notational choice.) So the **composition** $g_2 g_1$ of two symmetries is still a symmetry, and therefore still belongs to the set G. Second,

> applying a symmetry *in reverse* is still a symmetry.

In other words, if there is a symmetry g which takes the object from state A to state B, then there is an inverse symmetry which takes the object from state B back to state A. We'll denote this inverse symmetry by g^{-1}. Finally, there is always a *trivial* symmetry, obtained by doing nothing to the object. The operation of doing nothing is always a symmetry, by definition.

Most sets do not have these two interesting structures, but we see that sets of symmetries always do. So, in order to study symmetries, we give a name to sets with such structures: they are called *groups*.

1.1 Definitions and first examples

Definition 1.1. A **group** G is a set that has a **group operation** $\star$. More precisely, this means that for any two elements a and b in G, we can apply the operation $\star$ to them to obtain an element $a \star b$. This operation must satisfy some axioms:

1. there must be an **identity element** e of G such that $e \star x = x$ for all x;
2. the group operation must be **associative**, i.e. $(a \star b) \star c = a \star (b \star c)$;
3. every element x must have an **inverse**, i.e. an element x^{-1} such that $x \star x^{-1} = e$.

It is common to call the inverse x^{-1} because we often pretend the group operation is "multiplication" and refer to the group operation as a "product".

Many familiar objects that do not necessarily arise from the study of symmetries have group structures, with various group operations. It is important to note that, although the notation we use for abstract groups is "multiplicative", sometimes the group operation may be addition, or some other operation. So we often write $(G, \star)$ to mean a group G with the group operation $\star$, to make it clear what the group operation is. When it is clear from context, we sometimes just refer to the group as G.

Example 1.2. The set of integers, called $\mathbb{Z}$, forms a group using addition as the group operation. (To be precise, we should write $(\mathbb{Z}, +)$.) Clearly, given two integers x and y, their sum $x + y$ is still an integer.

1. The identity element is 0, because $0 + x = x$ for any integer x.
2. The operation of addition is associative, because $(x + y) + z = x + (y + z)$ for all integers x, y, z.
3. The inverse of an integer x is the integer $-x$ (which always exists), because $x + (-x) = 0$.

Exercise. Show that $\mathbb{Z}$ with multiplication as the group operation is *not* a group. Is it possible to "fix" $\mathbb{Z}$ so that it is?

Exercise. Let $\mathbb{Z}/n$ denote the group of **integers modulo** n, using addition modulo n as the group operation. In other words, it is the set $\{0, 1, 2, \ldots, n-2, n-1\}$ where the result of the group operation on a and b is the remainder of $a + b$ upon dividing by n. Check that $\mathbb{Z}/n$ is a group.

Example 1.3. Given an object, its **symmetry group** is the group of all symmetries of the object, using composition as the group operation. The identity element e for this operation is always the symmetry which takes the object and does nothing to it; every object clearly has such a symmetry. The inverse of a symmetry is the symmetry "in reverse".

There are many structural properties which are already illustrated by these examples. For example, groups whose elements are numbers usually have the following very special property. It is important to emphasize that most groups, particularly symmetry groups, do *not* have this property!

Definition 1.4. A group G is **abelian** if

$$x \star y = y \star x$$

for every x and y in G. We say the group operation is **commutative**.

We also want to speak about the size of groups, namely how many elements they contain. It is possible of course for a group to contain infinitely many elements, like $\mathbb{Z}$, so we usually only talk about the size of *finite* groups.

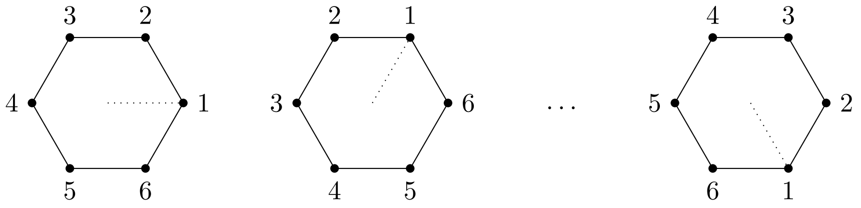
Definition 1.5. The number of elements, or **cardinality** or **order**, of a group G is written $|G|$. We say G is **finite** or **infinite** depending on its cardinality.

1.2 The dihedral group

One simple yet very interesting example of a symmetry group is the symmetry group of a regular polygon with n (equal) sides. Its symmetry group is called the **dihedral group**, and written D_n. To reduce confusion when studying D_n, it is best to label each corner of the polygon with a number, to keep track of what each symmetry does.

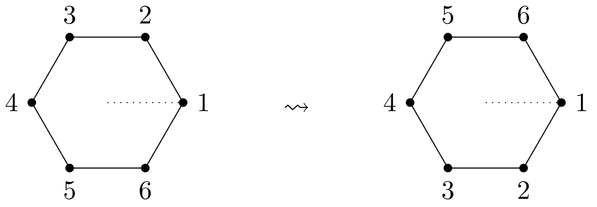
The first step in understanding D_n is to identify some of its elements, and to give names to them.

1. There are n different symmetries obtained by rotation.



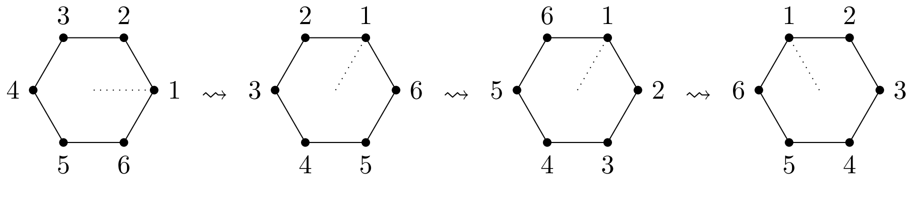
   The first one is the identity element e. If we call the second one r, note that the other rotations are just compositions of r with itself. So the rotation symmetries are
   $$e, r, r^2, r^3, \ldots, r^{n-1}.$$
   Note that $r^n = e$, which is the statement that rotating a full 360° is the same as not doing anything. From this we can tell that $r^{-1} = r^{n-1}$.

2. There is a symmetry given by flipping the polygon across a fixed axis, which we'll take to be the x-axis for simplicity.



   Call this symmetry s. Note that $s^2 = e$, since flipping twice is the same as not doing anything.

What about flips across other lines? In the same way that all rotations are obtained by compositions of r, those other flips may be obtained by an appropriate composition of r and s. For example, for the hexagon, flipping across the line between 2 and 5 is the same as the composition rsr.



Exercise. Check that rs is *not* the same symmetry as sr, and therefore conclude that the dihedral group is *not* abelian.

Exercise. Check that $rs = sr^{-1}$. Conclude that $r^k s = sr^{-k}$ for any integer k.

We can use this last exercise to obtain a full description of the dihedral group as follows. Suppose we are given a complicated composition of r and s, like

$$r^{27}srsr^8s^3.$$

Any such expression can be simplified into the form r^k or sr^k for some integer k using the following two steps.

1. Use that $r^n = e$ and $s^2 = e$ to simplify the exponents.
2. "Move" all the occurrences of s to the front using $r^k s = sr^{-k}$.

Example 1.6. Let's simplify $r^{27}srsr^8s^3$ for the hexagon. Since $r^6 = e$ and $s^2 = e$, we get

$$r^{27}srsr^8s^3 = r^3srsr^2s.$$

Then we move the first s to the front:

$$(r^3s)rsr^2s = (sr^{-3})rsr^2s = sr^{-2}sr^2s.$$

Moving the second s now gives

$$s(r^{-2}s)r^2s = s(sr^2)r^2s = r^4s.$$

Finally, moving the last s gives

$$r^4s = sr^{-4} = sr^2.$$

So even though we can write down very complicated compositions of rotations and flips, after simplifying we see that D_n actually only contains $2n$ elements:

1. n rotations $e, r, r^2, \dots, r^{n-1}$;
2. n rotations-with-a-flip $s, sr, sr^2, \dots, sr^{n-1}$.

This makes a lot of sense, because any symmetry of the n-gon must take the corner labeled 1 to some position. We can use rotations to place the 1 there. Then we are left with only two possibilities: either the numbers of corners adjacent to the 1 are already correct, in which case we have identified the symmetry as r^k for some k, or the numbers are flipped, in which case we apply an extra flip to get sr^k.

Definition 1.7. Any element in the dihedral group can be written as a composition of r and s, so we say D_n is **generated by** r and s. The rules we impose on how multiple r and s interact are called **relations**, and we identified three:

$$r^n = e, \quad s^2 = e, \quad rs = sr^{-1}.$$

A full description of D_n is given by a **presentation** using generators and relations, written

$$D_n = \langle r, s \mid r^n = e,\ s^2 = e,\ rs = sr^{-1} \rangle.$$

1.3 The symmetric group

A more complicated example of a symmetry group is the symmetry group of n indistinguishable objects, e.g. point particles. Such objects may be permuted in any order, and all permutations are symmetries. We label the objects from 1 to n, in which case permutations look like

$$\begin{matrix} \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ 1 & 2 & 3 & 4 & 5 & 6 \end{matrix} \quad \rightsquigarrow \quad \begin{matrix} \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ 3 & 2 & 6 & 4 & 1 & 5 \end{matrix}$$

This symmetry group is called the **symmetric group**, and written S_n. We can immediately note that it consists of $n!$ elements. One way to write elements is to just list the permuted labels under the original labels, like

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 2 & 6 & 4 & 1 & 5 \end{pmatrix}$$

for the above example. (We will see however that writing elements like this isn't the best way to uncover the hidden structures in S_n.)

We can ask for a generators-and-relations presentation of S_n like we did for D_n, and the first step is to identify some special kinds of elements and give names to them.

1. For any two labels i and j, we can consider the permutation which swaps i and j and leaves everything else alone. Such permutations are called **transpositions**, and are written (i, j).

2. More generally, given a sequence of labels $i_1, i_2, \dots, i_m$, we can consider the permutation which sends i_1 to i_2, and i_2 to i_3, and so on, and i_m back to i_1. Such permutations are called **cycles**, and are written $(i_1, i_2 \dots, i_m)$. The **length** of a cycle is the number of items involved in it.

Theorem 1.8. *S_n is generated by transpositions.*

Proof. Given any permutation σ in S_n, if we can sort out its items in increasing order using just transpositions (to get to the identity element e), then the inverse sequence of transpositions is equal to σ. But sorting is easy: the first transposition should swap the first element in σ with 1, the second should then swap the second element with 2, etc. □

Exercise (Hard)**.** Show that S_n is actually generated by *adjacent transpositions* $\sigma_i = (i, i+1)$ for $1 \leq i < n$, and that their compositions are governed by the relations

- $\sigma_i^2 = e$ for all i;
- $\sigma_i \sigma_j = \sigma_j \sigma_i$ when $|i - j| > 1$;
- $\sigma_i \sigma_{i+1} \sigma_i = \sigma_{i+1} \sigma_i \sigma_{i+1}$ for all i.

Because of the theorem, it is useful to write elements of S_n as compositions of transpositions. But this can often become cumbersome to write. Instead, we write them as compositions of *cycles*, due to the following exercise.

Exercise. Show that cycles are just shorthand for compositions of transpositions, because

$$(i_1, i_2, \ldots, i_m) = (i_1, i_2)(i_2, i_3) \cdots (i_{m-1}, i_m).$$

To decompose a given permutation σ into a product of cycles, it is easiest to start with the label 1 and write down the sequence $1, \sigma(1), \sigma(\sigma(1)), \ldots$ until we return to 1; this forms a cycle. Then take the next smallest label not included in this cycle, and form a new cycle starting with it, and so on. Note that sometimes there will be cycles of length 1, which we *omit* writing.

Example 1.9. Consider the permutation $\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 3 & 4 & 1 & 6 & 2 & 5 & 7 \end{pmatrix}$ in S_7.

1. There is a cycle $1 \to 3 \to 1$. This is written $(1, 3)$.
2. The next smallest number not involved in a cycle so far is 2. There is a cycle $2 \to 4 \to 6 \to 5 \to 2$. This is written $(2, 4, 6, 5)$.
3. The next smallest number not involved in a cycle so far is 7. There is a cycle $7 \to 7$. This is a length-1 cycle and we do not write it.
4. There are no more labels not involved in a cycle, so we are done.

Hence $\sigma = (1, 3)(2, 4, 6, 5)$.

Note that it does not matter which order we compose *disjoint* cycles, i.e. cycles that involve no common labels. Disjoint cycles commute with each other.

1.4 Homomorphisms

Now we return to discussing groups more abstractly. Given a group G, it is conceptually helpful to consider its "multiplication" table, where we write down all products of elements in the group. The convention we will use is to multiply the row element by the column element, not vice versa.

Example 1.10. The symmetric group S_2 (of two objects) has two elements, with the following multiplication table.

	e	$(1,2)$
e	e	$(1,2)$
$(1,2)$	$(1,2)$	e

Example 1.11. The group $\mathbb{Z}/2$ (of integers mod 2) also has two elements, with the following multiplication table.

	0	1
0	0	1
1	1	0

Note that, in some sense, we've written the same multiplication table twice but with elements renamed. The way to translate between S_2 and $\mathbb{Z}/2$ while preserving the multiplication table is

$$e \leftrightarrow 0, \quad (1,2) \leftrightarrow 1.$$

Using this dictionary, the two groups are actually *equivalent.* This notion of equivalence is expressed mathematically as follows.

Definition 1.12. Let G and H be two groups, with group operations $\star_G$ and $\star_H$. We say G and H are **isomorphic**, written

$$G \cong H,$$

if there exists a function $f\colon G \to H$ which:

1. is a **bijection**, i.e. a one-to-one correspondence between the elements of the two sets;

2. is a **homomorphism**, meaning that

$$f(a \star_G b) = f(a) \star_H f(b).$$

If we view f as a "dictionary" between elements of G and H, being a homomorphism means that the dictionary is compatible with the group operations in G and H, and being an isomorphism means the dictionary covers all elements of G and H.

Exercise. Show that D_3 is isomorphic to S_3.

Exercise. Show that D_n cannot be isomorphic to S_n for $n > 3$, using cardinality.

Importantly, it is possible for $f\colon G \to H$ to be a homomorphism without being an isomorphism. One trivial way is to send everything in G to the identity element e_H in H. Then clearly

$$f(a \star_G b) = e_H = f(a) \star_H f(b).$$

Example 1.13. Consider the map $f\colon \mathbb{Z}/2 \to D_3$ given by

$$0 \mapsto e, \quad 1 \mapsto s.$$

This is not an isomorphism because $\mathbb{Z}/2$ is much smaller than D_3. But it *is* a homomorphism. The most important check is

$$f(1+1) = e = s^2 = f(1)f(1).$$

The existence of this homomorphism means that there is a copy of $\mathbb{Z}/2$ hiding inside D_3.

Definition 1.14. A subset $H \subset G$ which itself is a group is called a **subgroup** of G. This is written $H \leq G$.

Exercise. Show that, in D_3, the elements e, r, r^2 form a subgroup isomorphic to $\mathbb{Z}/3$. On the other hand, show that e, r, sr^2 does not form a subgroup. Are there any other subgroups of D_3 that we haven't found yet?

Suppose we want to specify a homomorphism $f\colon G \to H$, and G has generators a, b, and c. Then it is actually enough to specify what $f(a)$, $f(b)$, and $f(c)$ are. This is because any element in G can be written as some product of a, b, and c, and therefore e.g.

$$f(a^7 b^{11} c^{-3}) = f(a)^7 f(b)^{11} f(c)^{-3}.$$

So a homomorphism is fully specified by what it does to generators.

Example 1.15. A homomorphism $\phi\colon \mathbb{Z} \to \mathbb{Z}$ is completely determined by the integer $\phi(1)$. This is because

$$\phi(n) = \phi(\underbrace{1 + \cdots + 1}_{n \text{ times}} = \underbrace{\phi(1) + \cdots + \phi(1)}_{n \text{ times}} = n\phi(1).$$

We also speak about generators of a *subgroup*. For example, the set of even integers forms a subgroup of $\mathbb{Z}$. It is often written $2\mathbb{Z}$, because it is generated by the element 2.

1.5 Operations on groups

Whenever we define a type of mathematical object (e.g. a group) along with some notion of equivalence (e.g. isomorphism of groups), we can start asking about *classification*. Namely,

> can we classify all the different objects of this type?

If the answer turns out to be yes, then usually the result is that every such object is built from a small collection of basic building blocks. In our case, this means we require a way to build a bigger group using two smaller ones.

Definition 1.16. Given two groups G and H, their **product** $G \times H$ is the group whose elements are pairs (g, h) with $g \in G$ and $h \in H$, and group operation given by

$$(g_1, h_1) \star (g_2, h_2) = (g_1 \star_G g_2, h_1 \star_H h_2).$$

Example 1.17. The group $\mathbb{Z}/2 \times \mathbb{Z}/2$ has elements

$$\{(0,0), (0,1), (1,0), (1,1)\}$$

and multiplication table

	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$(0,0)$	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$(0,1)$	$(0,1)$	$(0,0)$	$(1,1)$	$(1,0)$
$(1,0)$	$(1,0)$	$(1,1)$	$(0,0)$	$(0,1)$
$(1,1)$	$(1,1)$	$(1,0)$	$(0,1)$	$(0,0)$

.

Note that even though it has order 4, it is *not* isomorphic to $\mathbb{Z}/4$. One way to see this is that every element in $\mathbb{Z}/2 \times \mathbb{Z}/2$ becomes zero when added to itself, but this is not true for every element of $\mathbb{Z}/4$.

Exercise. Show that $\mathbb{Z}/2 \times \mathbb{Z}/3$ is isomorphic to $\mathbb{Z}/6$.

Exercise. Show that $\mathbb{Z}/n \times \mathbb{Z}/m$ is isomorphic to $\mathbb{Z}/nm$ whenever $\gcd(n,m) = 1$. Hint: construct an isomorphism

$$\phi\colon \mathbb{Z}/nm \to \mathbb{Z}/n \times \mathbb{Z}/m$$

by picking wisely what $\phi(1)$ should be.

A related operation that will also be relevant is the "inverse" operation to taking the product: the *quotient*. In some sense, this operation takes a group H and a group that looks like $G \times H$ and produces just G. If G is a group and $H \leq G$ is a subgroup, we can construct the **quotient** G/H. The idea is to forcibly make everything in H equal to the identity element e, thereby "getting rid of" H in G.

Example 1.18. Take the subgroup $2\mathbb{Z} \leq \mathbb{Z}$. In the quotient $\mathbb{Z}/2\mathbb{Z}$, all even integers are equivalent to 0. This is exactly the same group as what we have been calling $\mathbb{Z}/2$, which we now recognize is shorthand for $\mathbb{Z}/2\mathbb{Z}$.

In the quotient $\mathbb{Z}/2\mathbb{Z}$, the two elements are 0 and 1. In $\mathbb{Z}$, they correspond to the subsets of even and odd integers respectively. In the context of quotients, such subsets are called **cosets**, and are written like $g \star H$. Since the group operation is often multiplication, this is usually abbreviated as gH.

Example 1.19. For $2\mathbb{Z} \leq \mathbb{Z}$, the cosets of even and odd integers are

$$0 + 2\mathbb{Z} = \{0 + 2n \mid n \in \mathbb{Z}\}, \quad 1 + 2\mathbb{Z} = \{1 + 2n \mid n \in \mathbb{Z}\}.$$

Note that $0 + 2\mathbb{Z} = 2 + 2\mathbb{Z} = 4 + 2\mathbb{Z} = \cdots$. Addition can be done on cosets directly, e.g.

$$(1 + 2\mathbb{Z}) + (1 + 2\mathbb{Z}) = (1 + 1) + 2\mathbb{Z} = 2 + 2\mathbb{Z} = 0 + 2\mathbb{Z}.$$

This is the same as writing $1 + 1 = 0$ in $\mathbb{Z}/2\mathbb{Z}$.

Abstractly, we can treat elements of G/H as cosets, and the group operation in the quotient G/H is *defined* to be the group operation on cosets: to define $x \star y$ in G/H is the same as defining $xH \star yH$ in G. This is slightly problematic, because in general there is no good reason for $xH \star yH$ to be a coset at all! We would like to define

$$xH \star yH \overset{?}{=} (x \star y)H$$

but in general this is not true. The problem would be solved if $Hy = yH$, so that

$$xHyH = xyHH = xyH.$$

Here we used that H is a subgroup, so the product of any two elements of H remains in H, and so $H \star H = H$.

Definition 1.20. A subgroup $H \leq G$ is **normal** if $gH = Hg$ for any element $g \in G$. We write $H \triangleleft G$ to mean H is a normal subgroup of G.

Definition 1.21. Given a *normal* subgroup $H \triangleleft G$, the **quotient** G/H has a well-defined group operation, given by

$$g_1 H \star_{G/H} g_2 H = (g_1 \star_G g_2)H.$$

Example 1.22. Take $G = S_3$ and the subgroup $H = \{e, (1,2)\}$. The (distinct) cosets of H in G are

$$\begin{gathered} eH = (1,2)H = \{e, (1,2)\} \\ (1,3)H = (1,2,3)H = \{(1,3), (1,2,3)\} \\ (2,3)H = (1,3,2)H = \{(2,3), (1,3,2)\}. \end{gathered}$$

However H is *not* normal, because

$$H(1,3) = \{(1,3), (1,3,2)\} \neq (1,3)H.$$

This makes G/H fail to be a group. For example, we can check that

$$(1,2,3)H(1,2,3)H = \{e, (1,2), (2,3), (1,3,2)\},$$

which is not a coset at all.

Exercise. Show that in S_3, the subgroup H generated by $(1,2,3)$ is a normal subgroup. What is the quotient S_3/H?

1.6 Lagrange's theorem

Understanding cosets is useful for more than studying quotients. In this section we'll see an application, in the context of *finite* groups. In finite groups we can count things like number of elements.

Suppose we have *any* subgroup $H \leq G$. Then all the cosets of H have the same number of elements. For example, elements of eH and gH are related to each other by multiplying by g. So there are a total of $|G|/|H|$ cosets, each of size $|H|$. This immediately implies the following.

Theorem 1.23 (Lagrange). *For any subgroup $H \leq G$, the order of H divides the order of G.*

Definition 1.24. The **index** of the subgroup $H \leq G$ is the number

$$[G : H] = |G|/|H|.$$

So $|G| = [G : H]|H|$.

In particular, we can apply Lagrange's theorem to subgroups generated by a *single* element. Given $g \in G$, the subgroup it generates is usually denoted $\langle g \rangle$. Studying properties $\langle g \rangle$ is helpful for studying properties of g itself.

Definition 1.25. A (sub)group G is **cyclic** if it is generated by a single element. The **order** of an element g of a group is the size of the cyclic subgroup it generates, i.e. the smallest integer $n \geq 1$ such that

$$g^n = e.$$

Corollary 1.26. *The order of an element $g \in G$ divides $|G|$.*

Proof. Let $H = \langle g \rangle$ be the subgroup generated by g. Then the order of g is the order of H, by definition. But Lagrange's theorem says $|H|$ divides $|G|$. □

Lagrange's theorem can be viewed as a vast generalization of Fermat's little theorem from number theory, which says that for any integer a and any prime p,

$$a^p \equiv a \bmod p.$$

This is actually Lagrange's theorem applied to the group $(\mathbb{Z}/p)^\times$, defined as follows.

Definition 1.27. Given the group $(\mathbb{Z}/n, +)$, we can form its associated **group of units**, denoted $(\mathbb{Z}/n)^\times$.

- Its elements are the integers in $\mathbb{Z}/n$ that have *multiplicative* inverses, i.e. those x such that there exist y with $xy = 1$.
- Its group operation is multiplication.

Exercise. Check that if p is a prime, $(\mathbb{Z}/p)^\times$ consists of *all* the elements of $\mathbb{Z}/p$ except 0.

Exercise (Hard)**.** How many elements are in $(\mathbb{Z}/n)^\times$ for an arbitrary integer $n \geq 2$?

We can use Lagrange's theorem to prove Fermat's little theorem as follows. In $(\mathbb{Z}/p)^\times$, every non-zero element must generate the whole group, because the order of every non-trivial subgroup divides the prime p. So the order of every non-zero element is $|(\mathbb{Z}/p)^\times| = p - 1$. This means that every non-zero element a satisfies

$$a^{p-1} = e,$$

which is the same thing as saying $a^{p-1} \equiv 1 \bmod p$. Multiplying both sides by a gives Fermat's little theorem.

1.7 Classification

Now we can return to the problem of classifying different types of groups. The simplest type we can start thinking about are the *finite* and *abelian* ones. It is clear that $\mathbb{Z}/n$ and products of $\mathbb{Z}/n$'s are finite abelian groups, while $\mathbb{Z}$ (infinite) and S_3 (non-abelian) are not. If we start listing the non-isomorphic finite abelian groups of small order, it turns out we get

cardinality	non-isomorphic groups
1	1
2	$\mathbb{Z}/2$
3	$\mathbb{Z}/3$
4	$\mathbb{Z}/2 \times \mathbb{Z}/2$ and $\mathbb{Z}/4$
5	$\mathbb{Z}/5$
6	$\mathbb{Z}/2 \times \mathbb{Z}/3$
7	$\mathbb{Z}/7$
8	$(\mathbb{Z}/2)^3$ and $\mathbb{Z}/2 \times \mathbb{Z}/4$ and $\mathbb{Z}/8$
$\vdots$	$\vdots$

.

It is not obvious why they are all products of $\mathbb{Z}/n$'s. In fact this empirical observation is true in general.

Theorem 1.28 (Classification of finite abelian groups)**.** *Any finite abelian group is isomorphic to*

$$\mathbb{Z}/n_1 \times \cdots \times \mathbb{Z}/n_k$$

for some integers $n_1, \ldots, n_k \geq 2$ which are all prime powers.

Even though this only classifies a very special type of group, the proof of this theorem is already somewhat intricate, and we will skip it. The complexity of group theory is evident even in this special case.

We can go further and now ask for a classification of *all* finite groups, regardless of whether they are abelian. In the non-abelian case it turns out there are different ways to "take the product" of two groups, called *semidirect products*. So trying to decompose a group G into a product is not the best approach. Instead, we can write a **composition series** for G. This is a sequence

$$1 = H_0 \lhd H_1 \lhd H_2 \lhd \cdots \lhd H_n = G$$

of normal subgroups such that H_i is a *largest possible* normal subgroup of H_{i+1}. Equivalently, H_{i+1}/H_i is a simple group.

Definition 1.29. A group G with no normal subgroups aside from 1 and itself is called **simple**.

Finite simple groups are the building blocks for finite (non-abelian) groups. Unlike the abelian case, where the building blocks have a nice classification, the classification of finite simple groups involves 18 infinite families and 26 sporadic groups. The classification was a major mathematical milestone, "completed" in February 1981 (with some minor holes that were patched by 2004). The complete proof of the classification spans over 10,000 pages and is spread out across 500 or so papers. There is a current ongoing project to simplify and coalesce the proof into a 12-volume series, expected in 2023.

2 Representations of groups

The classification of even just the finite simple groups should indicate that groups in general are extremely complicated objects. The deep insight of representation theory is that

> a rich and fruitful way to study groups is by interpreting them as symmetries of some object, i.e. to examine their *actions* on objects.

While we can study an object by understanding its symmetry group, this insight says that, conversely, we can understand groups in general by making them act by symmetries on objects.

One immediately runs into a problem with this train of thought: the objects being acted on may be *more* complicated than the groups! It turns out, for various reasons, that we should only consider group actions on *linear* objects. This is like how we often use a linear approximation to a function instead of the function itself, because the linear approximation is much simpler. For example, in physics, we often use the *small-angle approximation* $\sin(x) \approx x$ for small angles x. The mathematical formalism for the "linear object" we want is a *vector space.*

Definition 2.1. A **vector space** is a group $(V, +)$ (always using addition as the group operation) and a **scalar multiplication** operation called $\cdot$.

- Elements in V are called **vectors**, usually called $\mathbf{v}$ or $\mathbf{w}$.
- Scalar multiplication defines how to multiply a vector by a real (or complex) number, called a **scalar**. We call V a vector space "over the real numbers" (or "over the complex numbers").
- Scalar multiplication must satisfy some axioms:

$$\begin{aligned}
a \cdot (b \cdot \mathbf{v}) &= (ab) \cdot \mathbf{v} \\
1 \cdot \mathbf{v} &= \mathbf{v} \\
a \cdot (\mathbf{v} + \mathbf{w}) &= a \cdot \mathbf{v} + a \cdot \mathbf{w} \\
(a + b) \cdot \mathbf{v} &= a \cdot \mathbf{v} + b \cdot \mathbf{v}.
\end{aligned}$$

Example 2.2. Let $\mathbb{R}^n$ denote the set of all points in n-dimensional space, i.e. all points $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ for arbitrary real numbers $x_1, \ldots, x_n$. Given a point, we can interpret it as the vector which starts at the origin $\mathbf{0}$ and ends at $\mathbf{x}$. Vectors in $\mathbb{R}^n$ are usually written like

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

1. Two vectors $\mathbf{x}$ and $\mathbf{y}$ can be added entry-wise. For example, in $\mathbb{R}^3$,

$$\begin{pmatrix} 3 \\ -1 \\ 0 \end{pmatrix} + \begin{pmatrix} 1 \\ 5 \\ 3/2 \end{pmatrix} = \begin{pmatrix} 4 \\ 4 \\ 3/2 \end{pmatrix}.$$

2. A scalar is a real number. We multiply a vector by a scalar by multiplying each entry by the scalar. For example, in $\mathbb{R}^3$,

$$3 \cdot \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 3 \\ 6 \\ 9 \end{pmatrix}.$$

Then one can verify that $\mathbb{R}^n$ is a vector space over the real numbers.

Exercise. Define $\mathbb{C}^n$ as all n-dimensional vectors with *complex numbers* as entries. Show that $\mathbb{C}^n$ is a vector space over the complex numbers.

For the purposes of this course, it doesn't hurt to pretend that any vector space is $\mathbb{R}^n$ or $\mathbb{C}^n$ for some n. Most of the time whether we use the real numbers $\mathbb{R}$ or the complex numbers $\mathbb{C}$ makes no difference.

If we want to make a group G act by symmetries on a vector space V, we better first understand symmetries of V. Such symmetries had better preserve the vector space structure of V, just like how symmetries of a square shouldn't "break apart" the structure of the square. Namely, if we view a symmetry as a function $\phi\colon V \to V$, it better be that

$$\begin{aligned} \phi(\mathbf{v} + \mathbf{w}) &= \phi(\mathbf{v}) + \phi(\mathbf{w}) \\ \phi(a \cdot \mathbf{v}) &= a \cdot \phi(\mathbf{v}) \end{aligned} \tag{1}$$

for any vectors $\mathbf{v}, \mathbf{w} \in V$ and scalar a.

Definition 2.3. A function $\phi\colon V \to W$ satisfying the conditions (1) is called an **homomorphism** of vector spaces. (Note that a vector space homomorphism is a group homomorphism which preserves scalar multiplication.) If $W = V$, it is an **endomorphism**.

Also, a symmetry must be *reversible*, i.e. the function ϕ must have an inverse, called ϕ^{-1}, such that

$$\phi(\phi^{-1}(\mathbf{v})) = \mathbf{v} = \phi^{-1}(\phi(\mathbf{v})). \tag{2}$$

Definition 2.4. A vector space homomorphism $\phi\colon V \to W$ is an **isomorphism** if there exists a function $\phi^{-1}\colon W \to V$ satisfying (2). If $W = V$, it is an **automorphism**.

In other words, to make G act on V is the same as making elements of G correspond to *automorphisms* of V. To do so, we must develop some tools and notation to work with automorphisms of V, and more generally homomorphisms of vector spaces. This is the purpose of *linear algebra*.

2.1 Linear algebra

Given a vector space V, one can pick a *basis* for it. A basis is, in some sense, a choice of what "coordinate system" to use for V. For example, moving one unit north and then three units east is the same as moving two units northeast and one unit southeast, but in the former we used $\{\text{north}, \text{east}\}$ as the coordinate system and in the latter we used $\{\text{northeast}, \text{southeast}\}$.

Definition 2.5. A **basis** of a vector space V is a set $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ of vectors in V such that:

1. (linearly independent) there is no way to write $\mathbf{v}_k$ as some combination

   $$a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_{k-1}\mathbf{v}_{k-1}$$

   for any scalars $a_1, \ldots, a_{k-1}$;

2. (spanning) every vector $\mathbf{v} \in V$ can be written in terms of vectors in the basis, in the form

   $$\mathbf{v} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n$$

   for some scalars $a_1, \ldots, a_n$.

Example 2.6. There is a *standard basis* for $\mathbb{R}^n$, given by $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ where

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \quad \ldots, \quad \mathbf{e}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}.$$

Then we can rewrite any vector as

$$\begin{pmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{pmatrix} = a_1\mathbf{e}_1 + a_2\mathbf{e}_2 + \cdots + a_n\mathbf{e}_n.$$

This is the true meaning of the vector notation. By default, the entries of a vector tell us what combination of *standard* basis vectors to take. However, there are many different choices for a basis of V in general. A vector $\mathbf{v} \in V$ may "look" different in different bases.

Example 2.7. Consider the vector $\mathbf{v} = \begin{pmatrix}1\\3\end{pmatrix} \in \mathbb{R}^2$. It is, by default, written in the standard basis $\{\mathbf{e}_1, \mathbf{e}_2\}$.

- In the basis $\{\mathbf{v}_1, \mathbf{v}_2\}$ where $\mathbf{v}_1 = \begin{pmatrix}1\\1\end{pmatrix}$ and $\mathbf{v}_2 = \begin{pmatrix}1\\-1\end{pmatrix}$, it is
$$\begin{pmatrix}2\\-1\end{pmatrix} = 2\mathbf{v}_1 - 1\mathbf{v}_2.$$

- In the basis $\{\mathbf{w}_1, \mathbf{w}_2\}$ where $\mathbf{w}_1 = \begin{pmatrix}2\\3\end{pmatrix}$ and $\mathbf{w}_2 = \begin{pmatrix}-1/2\\0\end{pmatrix}$, it is
$$\begin{pmatrix}1\\2\end{pmatrix} = \mathbf{v}_1 + 2\mathbf{v}_2.$$

In this way, after choosing a basis we can encode any vector in an n-dimensional vector space using n numbers, called the **coordinates** of the vector. (Choosing a basis is the same as choosing an isomorphism $V \cong \mathbb{R}^n$.) Similarly, we can encode an endomorphism ϕ of an n-dimensional vector space using n^2 numbers as follows. If the basis is $\{\mathbf{v}_i\}$, then ϕ is completely specified (using linearity) by the n vectors $\phi(\mathbf{v}_1), \ldots, \phi(\mathbf{v}_n)$. We put the coordinates of these vectors as columns in a **matrix**:

$$\boldsymbol{\phi} = \begin{pmatrix} | & | & & | \\ \phi(\mathbf{v}_1) & \phi(\mathbf{v}_2) & \cdots & \phi(\mathbf{v}_n) \\ | & | & & | \end{pmatrix}.$$

Definition 2.8. The entry on the i-th row and j-th column of a matrix $\mathbf{M}$ is denoted M_{ij}.

Given an arbitrary vector $\mathbf{v} = a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n$, linearity says

$$\phi(\mathbf{v}) = a_1\phi(\mathbf{v}_1) + \cdots + a_n\phi(\mathbf{v}_n).$$

We can express this resulting vector using just the matrix for ϕ, whose entries we'll denote by ϕ_{ij}:

$$\phi(\mathbf{v}) = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1n} \\ \phi_{21} & \phi_{22} & \cdots & \phi_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{n1} & \phi_{n2} & \cdots & \phi_{nn} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix},$$

where the i-th entry of the resulting vector is

$$b_i = \phi_{i1}a_1 + \phi_{i2}a_2 + \cdots + \phi_{in}a_n.$$

This is called **matrix-vector multiplication**.

Exercise. Formulate the analogous rule for how to multiply two matrices, in order to compute the *composition* of two endomorphisms of V.

Remember that we want groups to act by *automorphisms* of V, not just by *endomorphisms*. In other words, we need a tool to determine whether the endomorphism specified by a matrix $\mathbf{M}$ is *invertible* or not. The idea is to view $\mathbf{M}$ as the endomorphism sending the original basis vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ to the vectors given by its columns $\mathbf{w}_1, \ldots, \mathbf{w}_n$; as long as $\{\mathbf{w}_i\}$ are linearly independent, the inverse map is

$$\mathbf{w}_1 \mapsto \mathbf{v}_1, \quad \ldots, \quad \mathbf{w}_n \mapsto \mathbf{v}_1.$$

One way to systematically check for linear independence is to use the following numerical invariant of a matrix.

Definition 2.9. The **determinant** of a matrix $\mathbf{M}$, written $\det(\mathbf{M})$, is essentially the volume of the parallelepiped formed by the column vectors of $\mathbf{M}$.

1. If $\mathbf{M}$ is **upper triangular**, i.e. $M_{ij} = 0$ if $i > j$, then

$$\det(\mathbf{M}) = M_{11} M_{22} \cdots M_{nn}.$$

2. If $\mathbf{M}$ is not upper triangular, **row-reduce** it until it is in upper triangular form, and then use (1). Row-reducing $\mathbf{M}$ means we can apply any of the following operations:
   - add a multiple of one row to a different row, which leaves $\det(\mathbf{M})$ unchanged;
   - swap two rows, which multiplies $\det(\mathbf{M})$ by -1;
   - multiply a single row by a scalar c, which also multiplies $\det(\mathbf{M})$ by c.

Theorem 2.10. *The determinant* $\det(\mathbf{M})$ *is non-zero if and only if* $\mathbf{M}$ *is invertible.*

Example 2.11. We can compute a formula for the determinant of an *arbitrary* 2×2 matrix, since they are small enough. Let

$$\mathbf{M} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

This is not in row-reduced form, so let's row-reduce it.

1. Add $-c/a$ times the first row to the second row, to get

$$\mathbf{N} = \begin{pmatrix} a & b \\ 0 & d - bc/a. \end{pmatrix}.$$

   This doesn't change the determinant, i.e. $\det(\mathbf{M}) = \det(\mathbf{N})$.

2. Now $\mathbf{N}$ is upper triangular, so

$$\det(\mathbf{N}) = a \cdot (d - bc/a) = ad - bc.$$

So $\det(\mathbf{M}) = ad - bc$.

Exercise. The $-c/a$ in the above example means we must assume $a \neq 0$. How should we fix the example so it actually works for *all* possible matrices, not just those with $a = 0$?

The determinant is useful for much more than just measuring linear independence. It turns out to be an *invariant* of matrices. This means that if two matrices $\mathbf{M}$ and $\mathbf{N}$ actually represent the *same* endomorphism but in different bases, we will still have $\det(\mathbf{M}) = \det(\mathbf{N})$, even though the matrices $\mathbf{M}$ and $\mathbf{N}$ may look completely different.

Example 2.12. Take the endomorphism $\phi\colon \mathbb{R}^2 \to \mathbb{R}^2$ which takes a vector and rotates it counterclockwise by $\pi/2$.

- In the standard basis $\mathbf{e}_1, \mathbf{e}_2$, it is represented by the matrix

  $$\mathbf{M} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

  whose columns are $\phi(\mathbf{e}_1)$ and $\phi(\mathbf{e}_2)$.

- In the basis $\mathbf{v}_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ and $\mathbf{v}_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, it is represented by the matrix

  $$\mathbf{N} = \begin{pmatrix} 4/3 & 5/3 \\ -5/3 & -4/3 \end{pmatrix}.$$

  This is because

  $$\begin{aligned} \phi(\mathbf{v}_1) &= \frac{4}{3}\mathbf{v}_1 - \frac{5}{3}\mathbf{v}_2 \\ \phi(\mathbf{v}_2) &= \frac{5}{3}\mathbf{v}_1 - \frac{4}{3}\mathbf{v}_2. \end{aligned}$$

Note that $\mathbf{M} \neq \mathbf{N}$, but because they both come from the same endomorphism,

$$\det(\mathbf{M}) = 0 \cdot 0 - (-1) \cdot 1 = (4/3) \cdot (-4/3) - (-5/3) \cdot (5/3) = \det(\mathbf{N}).$$

It is not easy to find other numerical invariants of matrices. Suppose we have a function f which takes a matrix $\mathbf{M}$ and gives a number. Then f being an invariant means

$$f(\mathbf{M}) = f(\mathbf{P}^{-1}\mathbf{M}\mathbf{P})$$

for any invertible matrix $\mathbf{P}$. We call the operation

$$\mathbf{M} \mapsto \mathbf{P}^{-1}\mathbf{M}\mathbf{P}$$

conjugation by $\mathbf{P}$. The idea is that $\mathbf{P}$ is a *change of basis* matrix, namely a "dictionary" to translate from one basis to another. To apply $\mathbf{M}$ in a different basis, we first translate to that basis, apply $\mathbf{M}$, and then translate back. So $\mathbf{M}$ and $\mathbf{P}^{-1}\mathbf{M}\mathbf{P}$ both express the same endomorphism, but in different bases.

Exercise. A change of basis matrix $\mathbf{P}$ from the standard basis to a new basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is given by writing $\mathbf{v}_1, \ldots, \mathbf{v}_n$ as the columns of $\mathbf{P}$. Check in the previous example that

$$\mathbf{N} = \mathbf{P}^{-1}\mathbf{M}\mathbf{P}.$$

For an $n \times n$ matrix, it turns out there are exactly n different invariants. (They are different in the sense that they are linearly independent.) Aside from the determinant, the most useful one is also the simplest one.

Definition 2.13. The **trace** of a matrix $\mathbf{M}$ is the quantity

$$\operatorname{tr}(\mathbf{M}) = M_{11} + M_{22} + \cdots + M_{nn}.$$

Exercise. Show that $\operatorname{tr}(\mathbf{MN}) = \operatorname{tr}(\mathbf{NM})$ for any two matrices $\mathbf{M}$ and $\mathbf{N}$. This immediately shows tr is an invariant, because

$$\operatorname{tr}(\mathbf{P}^{-1}\mathbf{M}\mathbf{P}) = \operatorname{tr}(\mathbf{M}\mathbf{P}\mathbf{P}^{-1}) = \operatorname{tr}(\mathbf{M}).$$

2.2 Definitions and first examples

Now we return to making groups act on vector spaces. Recall that they should act by automorphisms. The collection of *all* automorphisms of V forms a group, using composition as the group operation. The identity matrix

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

is the identity element.

Definition 2.14. Let V be an n-dimensional vector space. The group of automorphisms of V is called $\mathrm{GL}(V)$, or $\mathrm{GL}(n)$. Equivalently,

$$\mathrm{GL}(n) = \{\text{invertible } n \times n \text{ matrices}\}.$$

A **representation** of a group G is a group homomorphism

$$\rho \colon G \to \mathrm{GL}(n)$$

for some n. The **dimension** of the representation is n.

One way to think about a representation is that we "represented" each element in G with a matrix, in such a way that multiplying the matrices gives exactly the same result as multiplying the elements in G. The matrix corresponding to an element $g \in G$ encodes exactly how g is supposed to act on vectors.

Example 2.15. There is a 2-dimensional representation $S_2 \to \mathrm{GL}(2)$ given by

$$e \mapsto \mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad (1,2) \mapsto \mathbf{S} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

One can check that this is a homomorphism, basically because $\mathbf{S}^2 = \mathbf{I}$. As an action on vectors, we therefore have

$$\begin{aligned} e \cdot \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} &= \mathbf{I} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \\ (1,2) \cdot \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} &= \mathbf{S} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} a_2 \\ a_1 \end{pmatrix}. \end{aligned}$$

In other words, in this representation, e is the symmetry which does nothing (as expected) and $(1,2)$ is the symmetry which swaps the two entries of a given vector. This is exactly the original symmetry we used to define S_n, and for this reason this representation is called the **permutation representation**.

Exercise. To define the (n-dimensional) permutation representation for S_n in general, it suffices to say what the generators (i,j) do in terms of matrices. Describe the matrix which takes a vector and swaps only its i-th and j-th entries.

The key distinction between this permutation representation and the original definition of S_n is that the original definition acted on just n indistinguishable objects. The set of all possible configurations of those objects has *no linearity* properties; it makes no sense to "add" two configurations. Now we have made S_2 act on *vectors*, where it makes sense to take two vectors and add them.

Example 2.16. For any group G and any $\mathrm{GL}(n)$, there is always the n-dimensional **trivial representation** given by

$$g \mapsto \mathbf{I},$$

the identity matrix.

Example 2.17. Define a 2-dimensional representation $\rho\colon D_3 \to \mathrm{GL}(2)$ as follows.

- The rotation r acts on vectors in $\mathbb{R}^2$ by rotating them by $2\pi/3$. As a matrix,

  $$\rho(r) = \begin{pmatrix} -1/2 & -\sqrt{3}/2 \\ \sqrt{3}/2 & -1/2 \end{pmatrix},$$

  because rotation sends

  $$\mathbf{e}_1 \mapsto \begin{pmatrix} -1/2 \\ \sqrt{3}/2 \end{pmatrix}, \quad \mathbf{e}_2 \mapsto \begin{pmatrix} -\sqrt{3}/2 \\ -1/2 \end{pmatrix}.$$

- The flip s acts on vectors in $\mathbb{R}^2$ by flipping them across the x-axis. As a matrix,

  $$\rho(s) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Since $D_3 \cong S_3$ via

$$r \mapsto (1,2,3), \quad s \mapsto (1,2),$$

this also defines a 2-dimensional representation of S_3.

Given a representation $G \to \mathrm{GL}(V)$, it is common to work with formulas which express the action of $g \in G$ on vectors $v \in V$ as $g \cdot v$. For example, the previous example has

$$r \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -1/2 \\ -\sqrt{3}/2 \end{pmatrix}.$$

We will constantly switch between thinking of $G \to \mathrm{GL}(V)$ as an assignment of matrices to group elements, and as a way to equip vectors in V with an action by elements in G. In the latter way of thought, we often just say "V is a representation", since V and the G-action it carries is the most important piece of data.

2.3 Sums and reducibility

To begin understanding representations, we should have a notion of building up new representations from simpler ones, like we did for groups.

Definition 2.18. Let $\phi\colon G \to \mathrm{GL}(V)$ and $\rho\colon G \to \mathrm{GL}(W)$ be two representations of G. Define the **direct sum** $V \oplus W$ to be the vector space arising from the product of the two groups $(V,+)$ and $(W,+)$. In other words, its elements are pairs $(\mathbf{v}, \mathbf{w})$, with element-wise addition and scalar multiplication. Then $V \oplus W$ is also a representation of G, because we can define the G-action as

$$g \cdot (\mathbf{v}, \mathbf{w}) = (g \cdot \mathbf{v}, g \cdot \mathbf{w}).$$

Formally, the resulting representation is

$$\phi \oplus \rho\colon G \to \mathrm{GL}(V \oplus W).$$

Example 2.19. Let $\phi\colon S_3 \to \mathrm{GL}(3)$ be the permutation representation, and $\rho\colon S_3 \to \mathrm{GL}(2)$ be the trivial representation. The direct sum of these two representations is a new representation

$$\phi \oplus \rho\colon S_3 \to \mathrm{GL}(5),$$

where, for example,

$$(\phi \oplus \rho)((1,2,3)) = \begin{pmatrix} 0 & 0 & 1 & & \\ 1 & 0 & 0 & & \\ 0 & 1 & 0 & & \\ & & & 1 & 0 \\ & & & 0 & 1 \end{pmatrix}.$$

This is because

$$\phi((1,2,3)) = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad \rho((1,2,3)) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

In general, one way to think about direct sum is that we take the two original matrices $\mathbf{A}$ and $\mathbf{B}$ and form a new *block-diagonal* matrix, where the blocks are $\mathbf{A}$ and $\mathbf{B}$ and everything else is zero. In this way, it is very easy to tell when a given representation can be written as a direct sum: it can be if and only if, in some basis, all the matrices associated to group elements split as block-diagonal matrices.

Definition 2.20. Let V be a representation. A **sub-representation** is a subspace $W \subset V$ (like a sub*group*, but with vector spaces) which is itself a representation of G.

In other words, given $w \in W$, every action by elements in G must *remain* in W. This is very restrictive. For example, consider the representation of D_3 on $\mathbb{R}^2$. The subspace $W = \{\begin{pmatrix} x \\ 0 \end{pmatrix}\}$, i.e. everything on the x-axis, is indeed a subspace, but

$$r \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -1/2 \\ -\sqrt{3}/2 \end{pmatrix} \notin W.$$

So W does *not* give a sub-representation. On the other hand, given a representation $V = W_1 \oplus W_2$, each W_i is clearly always a sub-representation.

Definition 2.21. A representation is called **irreducible** if it has no non-trivial sub-representations. We often abbreviate "irreducible representation" as "**irrep**".

If W is a sub-representation of V, we would like to conclude that V decomposes as $V = W \oplus W'$ where W' is some other sub-representation. Then it would be true that any representation decomposes as a direct sum of irreps, and we could study only the irreps as building blocks of all representations. Unfortunately this is not true for all groups.

Example 2.22. Consider the representation $\phi \colon \mathbb{R} \to \mathrm{GL}(2)$ given by

$$x \mapsto \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix}.$$

Then the x-axis is a sub-representation, but there is no complementary subspace which is also a sub-representation. So ϕ has non-trivial sub-representations but cannot be written as $\rho_1 \oplus \rho_2$.

Theorem 2.23 (Maschke's theorem)**.** *If G is a finite group, then every representation of G decomposes into irreducible representations.*

Groups with this property are called **semisimple**. Later when we study the representation theory of Lie groups, which are infinite groups, we'll see that semisimple Lie groups are particularly nice. For now, since finite groups are all semisimple by the theorem, we can think of all representations as built from irreps via direct sum.

2.4 Morphisms

In the same way that a group homomorphism captures the appropriate notion of a "function preserving group structure", a morphism of representations captures the notion of a "function preserving G-action". More rigorously, given two representations V and W, we would like a function $\mathbf{T}\colon V \to W$ to preserve:

- the linear vector space structure of V and W, i.e. ρ should be a *linear* map;
- (**new**) the G-action on V vs. the G-action on W, since both vector spaces are representations of G.

Definition 2.24. A **morphism of representations** (or an **intertwiner**) between two representations $\phi_1\colon G \to \mathrm{GL}(V)$ and $\phi_2\colon G \to \mathrm{GL}(W)$, is a linear transformation

$$\mathbf{T}\colon V \to W$$

such that

$$\mathbf{T}\phi_1(g) = \phi_2(g)\mathbf{T} \quad \text{for every } g \in G.$$

One way intertwiners arise is as follows. Suppose we defined a representation $\phi_1\colon G \to \mathrm{GL}(V)$ by picking some matrices corresponding to generators, but then we decided to do a *change of basis* on V. Then the matrices defining ϕ_1 will changed as well, to give a new representation $\phi_2\colon G \to \mathrm{GL}(V)$ on the same vector space. If the change of basis matrix is $\mathbf{P}$, then this means

$$\mathbf{P}^{-1}\phi_2(g)\mathbf{P} = \phi_1(g).$$

Rearranging, we see that $\mathbf{P}$ is an intertwiner between ϕ_1 and ϕ_2, by definition.

Definition 2.25. Two representations are **equivalent** if there exists an invertible intertwiner between them.

Using this notion of equivalence on irreps gives the following fundamental result in representation theory. It essentially says that there's no way to intertwine between two truly different irreps. The caveat is that we must work with the *complex* numbers $\mathbb{C}$, and vector spaces with *complex* scalars. This is so that we can guarantee every polynomial has a solution.

Theorem 2.26 (Schur's lemma). *Let V and W be two irreps of G. If $\phi\colon V \to W$ is an intertwiner, then:*

1. *either ϕ is an isomorphism, or $\phi = 0$;*
2. *if $V = W$, then $\phi = \lambda\mathbf{I}$ for some constant $\lambda \in \mathbb{C}$.*

The proof is straightforward, but requires slightly more linear algebra than we have covered. We will present it as a series of exercises.

Exercise. Prove the first part of Schur's lemma as follows.

1. Show that the kernel of ϕ is a sub-representation of V. Since V is irreducible, conclude that ϕ is either injective or zero.

2. Show that the image of ϕ is a sub-representation of W. Since W is irreducible, conclude that ϕ is either surjective or zero.

3. Conclude that either ϕ is an isomorphism, or $\phi = 0$.

Exercise. Prove the second part of Schur's lemma as follows.

1. Explain why ϕ must have at least one eigenvalue $\lambda \in \mathbb{C}$, with some eigenvector $\mathbf{v}$.

2. Explain why $\phi - \lambda\mathbf{I}$ is still an intertwiner, and why it cannot be an isomorphism.

3. Apply the first part of Schur's lemma to $\phi - \lambda\mathbf{I}$ to conclude that $\phi = \lambda\mathbf{I}$.

We will see later why Schur's lemma is of *crucial* importance to quantum physics. For now, we should think of it as follows. Suppose we took a representation $\phi\colon G \to \mathrm{GL}(V)$ and broke it up into irreps

$$\phi(g) = \begin{pmatrix} \phi_1(g) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \phi_2(g) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \phi_k(g) \end{pmatrix}$$

and all the irreps are *inequivalent* to each other. Then the only possible things an intertwiner $\mathbf{T}\colon V \to V$ can do are:

- multiply each block $\phi_i(g)$ by some scalar λ_i;
- shuffle around the order of the blocks. (This can essentially be ignored, since it is equivalent to just re-ordering the vectors in the basis.)

Corollary 2.27. *Any irrep of an abelian group is one-dimensional.*

Proof. Let $\phi\colon G \to \mathrm{GL}(V)$ be an irrep of an abelian group G. Since G is abelian, all the operators $\phi(g)$ for $g \in G$ commute with each other. By definition, this means $\phi(g)\colon V \to V$ is an intertwiner. Schur's lemma then tells us $\phi(g) = \lambda\mathbf{I}$ for some constant λ. Since this is true for all $g \in G$, the entire group G must act by only scalar multiplication. The only way for such a representation to have no subreps (i.e. to be irreducible) is if it is one-dimensional. □

2.5 Tensors and duals

There are two more important operation on representations that we need to introduce. If the direct sum $\oplus$ is thought of as “addition” of reps, then one new operation $\otimes$ should be thought of as “multiplication” of reps. (We will try to make this analogy more precise later.) As with $\oplus$, first we need to define what $V \oplus W$ is as a vector space, and then specify how G acts on it.

Definition 2.28. Given two vector spaces V and W, their **tensor product** $V \otimes W$ is a new vector space consisting of elements called

$$\mathbf{v} \otimes \mathbf{w} \quad \text{for } \mathbf{v} \in V,\ \mathbf{w} \in W.$$

These elements can be:

- added one coordinate at a time, i.e.
$$\begin{aligned} \mathbf{v}_1 \otimes \mathbf{w} + \mathbf{v}_2 \otimes \mathbf{w} &= (\mathbf{v}_1 + \mathbf{v}_2) \otimes \mathbf{w} \\ \mathbf{v} \otimes \mathbf{w}_1 + \mathbf{v} \otimes \mathbf{w}_2 &= \mathbf{v} \otimes (\mathbf{w}_1 + \mathbf{w}_2); \end{aligned} \tag{3}$$

- multiplied by a scalar in *either* coordinate, i.e.
$$c(\mathbf{v} \otimes \mathbf{w}) = (c\mathbf{v}) \otimes \mathbf{w} = \mathbf{v} \otimes (c\mathbf{w})$$
for any scalar c.

The elements $\mathbf{v} \otimes \mathbf{w}$ of a tensor product are called **tensors**. Tensors should be thought of as a generalization of vectors. In particular, while vectors are *linear* objects, tensors are *multi-linear* objects. This means that they are linear in “one coordinate at a time”, like in (3), but not in all coordinates simultaneously:

$$\mathbf{v}_1 \otimes \mathbf{w}_1 + \mathbf{v}_2 \otimes \mathbf{w}_2 \neq (\mathbf{v}_1 + \mathbf{v}_2) \otimes (\mathbf{w}_1 + \mathbf{w}_2).$$

(This should be compared to how addition works in $V \oplus W$.) Indeed, the rhs should actually be expanded as

$$\begin{aligned} (\mathbf{v}_1 + \mathbf{v}_2) \otimes (\mathbf{w}_1 + \mathbf{w}_2) &= \mathbf{v}_1 \otimes (\mathbf{w}_1 + \mathbf{w}_2) + \mathbf{v}_2 \otimes (\mathbf{w}_1 + \mathbf{w}_2) \\ &= (\mathbf{v}_1 \otimes \mathbf{w}_1 + \mathbf{v}_1 \otimes \mathbf{w}_2) + (\mathbf{v}_2 \otimes \mathbf{w}_1 + \mathbf{v}_2 \otimes \mathbf{w}_2)\,. \end{aligned}$$

Exercise. Show, from the definition, that if V has basis $\{\mathbf{v}_i \mid i = 1, 2, \ldots, m\}$ and W has basis $\{\mathbf{w}_j \mid j = 1, 2, \ldots, n\}$, then $V \otimes W$ has basis

$$\{\mathbf{v}_i \otimes \mathbf{w}_j \mid i = 1, \ldots, m,\ j = 1, \ldots, n\}.$$

So $\dim(V \otimes W) = \dim(V)\dim(W)$.

Tensors and tensor calculus famously form the foundations of Einstein’s general relativity, in which objects like the curvature of spacetime (due to gravity) are represented by tensors.

Definition 2.29. If V and W are representations of G, then so is $V \otimes W$ via the action

$$g \cdot (v \otimes w) = (g \cdot v) \otimes (g \cdot w).$$

We call $V \otimes W$ the **tensor product** of representations.

The final operation we need to introduce is, at the level of matrices, essentially the *transpose* operation. At the more abstract level of vector spaces and linear transformations between them, it is called taking the *dual.* The idea is to view a transposed vector

$$\mathbf{v}^T = \begin{pmatrix} v_1 & v_2 & \cdots & v_n \end{pmatrix}$$

not as a vector, but rather as a *linear function on vectors.* In other words, $\mathbf{v}^T$ is actually something which takes a vector $\mathbf{w}$ and produces a scalar $\mathbf{v}^T\mathbf{w}$.

Definition 2.30. The **dual** of a vector space V is called $V^\vee$. Its elements are *linear functions* $f\colon V \to \mathbb{R}$.

- (Addition) Given two functions $f, g \in V^\vee$, their sum $f + g$ is the function such that
$$(f + g)(\mathbf{v}) = f(\mathbf{v}) + g(\mathbf{v}).$$
- (Scalar multiplication) Given a function $f \in V^\vee$ and a scalar c,
$$(cf)(\mathbf{v}) = cf(\mathbf{v}).$$

If V is a representation of G, then so is $V^\vee$ via the action

$$(g \cdot f)(\mathbf{v}) = f(g^{-1} \cdot \mathbf{v}). \tag{4}$$

Exercise (Technical)**.** Verify that we really need g^{-1} in (4) in order to have

$$(g_1 g_2) \cdot f = g_1 \cdot (g_2 \cdot f).$$

If we put g instead, we would have

$$(g_1 g_2) \cdot f = g_2 \cdot (g_1 \cdot f),$$

which is not the same and does *not* make $V^\vee$ a valid representation.

Given a vector space V with basis $\{\mathbf{v}_i\}$, the dual vector space $V^\vee$ has the same dimension and has a **dual basis** $\{f_i\}$ such that

$$f_i(\mathbf{v}_j) = \begin{cases} 1 & i = j \\ 0 & \text{otherwise.} \end{cases}$$

Using this dual basis, we can see that linear transformations $\mathbf{A}\colon V \to W$ are secretly tensors of a special kind. Namely, if $A(\mathbf{v}_i) = \mathbf{w}_i$ for basis elements $\mathbf{v}_i$, then

$$A = f_1 \otimes \mathbf{w}_1 + f_2 \otimes \mathbf{w}_2 + \cdots + f_n \otimes \mathbf{w}_n.$$

This is because

$$\begin{aligned}
A(\mathbf{v}_i) &= f_1(\mathbf{v}_i) \otimes \mathbf{w}_1 + \cdots + f_n(\mathbf{v}_i) \otimes \mathbf{w}_n \\
&= 0 \otimes \mathbf{w}_1 + \cdots + 1 \otimes \mathbf{w}_i + \cdots + 0 \otimes \mathbf{w}_n \\
&= 1 \otimes \mathbf{w}_i.
\end{aligned}$$

Here we are identifying $1 \otimes W \subset V \otimes W$ with the space W itself. In this manner, the study of linear transformations becomes the study of tensors in $V^\vee \otimes W$.

3 Quantum mechanics

Now we have enough machinery to be able to appreciate how representation theory and quantum physics interact. Every type of object we have discussed so far (groups, vector spaces, representations, etc.) all have specific interpretations in the context of quantum physics. Of course, they have meaning in *classical* physics as well, but the biggest application of representation theory lies solidly in the quantum world. In fact one can say that

> quantum mechanics is essentially linear algebra, and quantum mechanics *in the presence of symmetries* is essentially representation theory.

3.1 Quantum states

What is a quantum state? First we must understand classical states. Suppose we have a point particle moving around on a unit circle. Then its *classical* state is completely specified by a unit vector $\mathbf{x}$ and its velocity $\mathbf{v}$ (which must be perpendicular to $\mathbf{x}$). We don't need to specify its acceleration, jerk, etc. because Newton's second law $\mathbf{F} = m\mathbf{a}$ computes its acceleration from the information of what forces are applied to it. So the classical state of the particle is the pair $(\mathbf{x}, \mathbf{v})$. The set of such pairs of vectors, where $\mathbf{x}$ is a unit vector and $\mathbf{v}$ is perpendicular to $\mathbf{x}$, is called the **state space**.

The *double slit experiment* showed us that quantum states are very different from classical states. In essence, the experiment showed that it is possible for a particle to "interfere" with itself, in the same way that waves can. The only way this can happen is if the state of the particle were actually composed of several pieces. If we write $|x\rangle$ to denote the state where the particle is at position x, then we can consider states like

$$\frac{1}{2}|x\rangle + \frac{1}{2}|y\rangle, \tag{5}$$

which means that the particle has a 50% chance of being at x and a 50% chance of being at y. (In fact, to make the particle behave like a wave, we should assign it some probability $f(x)$ of being in *any* state $|x\rangle$, for all real numbers x. This function f is the **wave function**.)

The takeaway from the double slit experiment is that quantum states, unlike classical states, can be *added* together (with some probabilities) to create a **superposition** of states. For example, the state (5) is very different from the state $|x+y\rangle$, whatever $x+y$ means. This means that whatever the quantum state space is, it has a notion of addition, and also scalar multiplication.

Definition 3.1. The state space of a quantum system is sometimes called the **Hilbert space** of the system, and denoted $\mathcal{H}$. It is an axiom of quantum mechanics that $\mathcal{H}$ must be a vector space. A **quantum state** is a vector in $\mathcal{H}$.

There are two technicalities arising from how to interpret the scalar coefficients as probabilities.

1. For mathematical simplicity, we allow scalars to be *complex* numbers. A coefficient of $z = x + iy \in \mathbb{C}$ is actually interpreted as the probability $|z|^2 = x^2 - y^2$, not just z. This is called **Born's rule**. We call the original z the **amplitude**. Born's rule is usually taken as an axiom of quantum mechanics; I don't know any satisfying explanation for why the quadratic function $|z|^2$ should be the probability as opposed to e.g. some linear function like $x + y$, other than that in the wavefunction language it is what agrees with experiment.

2. How do we interpret a state like $|x\rangle + |y\rangle$, where the total probability is 2? The answer is to ignore normalization, in the sense that we look at the coefficient of each term, and divide it by the *total* of all coefficients. So for all intents and purposes, $|x\rangle + |y\rangle$ is exactly the same state as (5), which should be written as

$$\frac{1}{\sqrt{2}}\,|x\rangle + \frac{1}{\sqrt{2}}\,|y\rangle\,.$$

   Mathematically, the operation of declaring vectors in a vector space equivalent up to multiplying by an overall scalar is called taking the **projectivization** of the vector space $\mathcal{H}$, denoted $\mathbb{P}\mathcal{H}$. Technically the quantum state space is $\mathbb{P}\mathcal{H}$, not $\mathcal{H}$.

Example 3.2. Consider a quantum particle on the circle S^1. Let ψ be its wavefunction: given a point $z \in S^1$, the quantity $\psi(z)$ is the probability that the particle is at the point z. Note that whatever function ψ is, the *total* probability

$$\int_{S^1} |\psi(z)|^2\, dz$$

should be a finite number (so we can normalize by it). This quantity is called the L^2 **norm** of ψ and written $\|\psi\|_{L^2}$. The Hilbert space in this case is therefore

$$\mathcal{H} = L^2(S^1) = \left\{\psi\colon S^1 \to \mathbb{C} \;\middle|\; \|\psi\|_{L^2} < \infty\right\}.$$

In general, a particle in a space X should have Hilbert space $L^2(X)$, whose elements are complex-valued functions on X with finite L^2 norm.

3.2 Symmetries

Many physical systems have symmetries. Suppose there is a symmetry group G of a given physical system. Then, in particular, its Hilbert space $\mathcal{H}$ must also be acted on by G. This means $\mathcal{H}$ is a representation of G, and decomposes into irreps:

$$\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2 \oplus \cdots .$$

Like with summation notation $\sum$, this is sometimes written

$$\mathcal{H} = \bigoplus_n \mathcal{H}_n.$$

Irreps are physically very meaningful, because for a given system, different types of particles belong to different irreps.

Example 3.3. The circle S^1 has a rotational symmetry, forming a symmetry group called $U(1)$. This is the group consisting of elements $e^{i\theta}$ for $0 \leq \theta < 2\pi$, with (complex) multiplication as the group operation. Then $\mathcal{H} = L^2(S^1)$ becomes a representation of $U(1)$, where

$$e^{i\theta} \cdot \psi(z) = \psi(e^{-i\theta} z).$$

How does $L^2(S^1)$ decompose as a $U(1)$-representation? Since $U(1)$ is abelian, Schur's lemma says all irreps must be one-dimensional. So an irrep is spanned by a single function $\psi(z)$. This means that to find an irrep, we must look for a function $\psi(z)$ such that

$$\psi(e^{-i\theta} x) = \alpha \psi(x) \quad \text{for all } \theta, \tag{6}$$

for some constant α (which may depend on θ). For clarity, note that any point $z \in S^1$ is of the form $e^{i\phi}$, so we may as well write $\psi(\phi)$ instead of $\psi(z)$. In this notation, the equation is

$$\psi(\phi - \theta) = \alpha \psi(\phi) \quad \text{for all } \theta.$$

Exercise. Show that, for any given integer n, the function $\psi(e^{i\phi}) = e^{in\phi}$ satisfies the property (6).

It follows that irreps of $L^2(S^1)$ are the 1-dimensional subspaces spanned by the functions $\psi(e^{i\phi}) = e^{in\phi}$. We denote these subspaces by $\mathbb{C}e^{in\phi}$. Putting everything together yields the following.

Proposition 3.4 (Fourier decomposition)**.** *As $U(1)$-representations,*

$$L^2(S^1) = \widehat{\bigoplus}_{n\in\mathbb{Z}} \mathbb{C}e^{in\phi}.$$

The hat on top of $\oplus$ is a technicality arising from $L^2(S^1)$ being infinite-dimensional; since we mostly work with finite-dimensional reps, we won't comment on what the hat means.

Proposition 3.4 is quite powerful. Note that a function $\psi(\phi)$ on the circle S^1 can be interpreted as a function $f(x)$ which is 2π-*periodic*, i.e. one which satisfies

$$f(x + 2\pi) = f(x).$$

Then the proposition says any 2π-periodic function can be decomposed as a sum of the form

$$f(x) = \sum_{n \in \mathbb{Z}} \alpha_n e^{inx}$$

for some coefficients α_n. This is known as **Fourier decomposition**, and the α_n are called **Fourier coefficients**. The decomposition of $L^2(S^1)$ into irreps is essentially the theory of Fourier series.

Exercise. The sphere is denoted S^2, and its group of (rotational) symmetries is called SO(3). (Another name for $U(1)$ is SO(2), and in general SO($n+1$) acts on the n-dimensional sphere S^n.) We can consider the decomposition of $L^2(S^2)$ into irreps, as SO(3) representations. Unlike irreps in $L^2(S^1)$, the functions which form irreps in $L^2(S^2)$ are not simple. They are called **spherical harmonics**.

1. Read a little about spherical harmonics to convince yourself that they aren't simple functions.

2. Explain why rep theory tells us we shouldn't expect spherical harmonics to be nice simple functions. (Hint 1: Schur's lemma. Hint 2: there is a crucial difference between $U(1)$ and SO(3) as groups.)

Note that for $L^2(S^1)$, each irrep is classified by an integer n. For $L^2(S^2)$, it turns out we need two integers m, ℓ to classify vectors in irreps; this is why spherical harmonics are functions called Y_ℓ^m. This is because the irrep labeled by ℓ is actually $(2\ell + 1)$-dimensional, so $m = 1, 2, \ldots, 2\ell + 1$ labels which basis vector we pick in the irrep. the In general, in the context of quantum physics, the data needed to specify which irrep to consider are called **quantum numbers**.

Example 3.5 (Atomic orbitals)**.** Orbital states of electrons in atoms are described by spherical harmonics and the rep theory of SO(3) acting on $L^2(S^2)$, because electrons orbit in spherical shells around the nucleus. In addition to the quantum numbers ℓ and m, electrons have an additional quantum number n, called the *energy* . Just like how $1 \leq m \leq 2\ell + 1$, we have $0 \leq \ell < n$. In physics/chemistry we call these irreps **subshells** and give them specific names.

1. The $1s$ subshell corresponds to the irrep labeled by $n = 1$ and $\ell = 0$, which has dimension $2\ell + 1 = 1$. Therefore there is only one state called $1s$ inside the $1s$ orbital. In general this is true for the $2s, 3s, \ldots$ orbitals too.

2. The $2p$ subshell corresponds to the irrep labeled by $n = 2$ and $\ell = 1$, which has dimension $2\ell + 1 = 3$. Hence we have states called $2p^1, 2p^2, 2p^3$.

3. The $3d$ subshell corresponds to the irrep labeled by $n = 3$ and $\ell = 2$, which has dimension $2\ell + 1 = 5$. Hence we have states called $3d^1, 3d^2, 3d^3, 3d^4, 3d^5$.

Electrons prefer to be in the lowest-energy unoccupied state. The energy of a state in a subshell is primarily determined by n, but also depends on ℓ due to **screening** effects. The ordering in which states are actually filled up is

$$1s,\ 2s,\ 2p,\ 3s,\ 3p,\ 4s,\ 3d,\ 4p,\ 5s,\ \ldots.$$

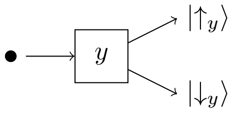
3.3 Observation

Part of what makes quantum mechanics unintuitive is that superpositions of states exist. The other part is the idea that

> measuring something about the state of a system *changes* the state of the system.

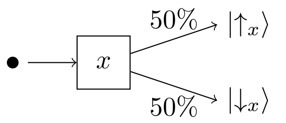
This is known as **wavefunction collapse**. Its most famous experimental evidence is in the double slit experiment, where it was found that if sensors were installed at the slits to record which slit the photon actually went through, the interference effect disappears. In layman's terms, "observing" the photon caused it to "collapse" from being a wave to being a particle. However, there is a more paradigmatic experiment which better illustrates the concept.

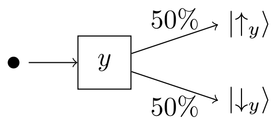
Stern–Gerlach experiment. In 1922 Stern and Gerlach shot silver atoms through a strong (vertical) magnetic field. If the atoms were classical particles, modeled as (spinning) magnetic dipoles, their interaction with the magnetic field means they would be deflected away from their original path by a (vertical) distance which depends on the orientation of the dipole. In particular, one expects a continuous distribution in the angle of deflection. The actual result is that all atoms were deflected by exactly either half a unit up or down, with no other possibilities.



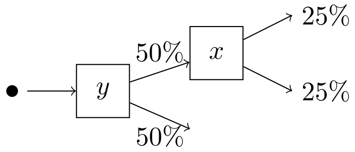
Schematically we represent this setup as a "Stern–Gerlach box" in the y direction. It turns out many elementary particles, including the silver atoms, have a quantum number called **spin** whose value is $1/2$, with the corresponding irrep being two-dimensional. The two states $|\uparrow_y\rangle$ and $|\downarrow_y\rangle$, called "spin up" and "spin down" in y, form a basis of this irrep. Later we'll see how spin arises as the quantum number for a more complicated relative of the SO(3) symmetry of space, called SU(2).

We can also imagine a Stern–Gerlach box in the x direction, i.e. so that the magnetic field is *horizontal*. As with the y box, half the incoming particles will be spin up in x, denoted $|\uparrow_x\rangle$, and the other half will be spin down in x, denoted $|\downarrow_x\rangle$. Schematically we draw this as follows.

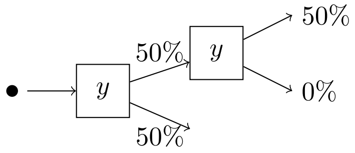




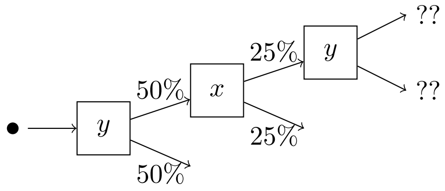
A more interesting configuration is to shoot only $|{\uparrow_y}\rangle$ states into a x box. If we think classically and imagine the particles as magnetic dipoles again, just knowing that the y component of the dipole moment points up tells us nothing about the x component. So we expect a half/half outcome, which is what we get.



Just as a sanity check (which will be important later), we can shoot only $|{\uparrow_y}\rangle$ states into another y box.



So far there is no quantum weirdness, because we have chosen to do fairly simple sequences of measurements. The simplest configuration where something strange happens is the following one.



One expects 25% and 0%, because we *have already measured* y and kept only the $|{\uparrow_y}\rangle$ particles. But in reality, we get 12.5% and 12.5%. It is as if the x measurement "destroyed" the state of being spin up in y.

The only consistent way to deal with such phenomena is to accept, as one of the axioms of quantum mechanics, that if a measurement tells us a system is in a state $|\mathbf{v}\rangle$, then even if the system were in a superposition before the measurement, after the measurement it is firmly, 100% in the state $|\mathbf{v}\rangle$. We call these **pure states**, as opposed to **mixed states** like

$$a\,|\mathbf{v}\rangle + b\,|\mathbf{w}\rangle\,.$$

The probability amplitude that such a mixed state becomes a pure state $|\mathbf{u}\rangle$ upon measurement is given by the **inner product**

$$a\langle\mathbf{u}|\mathbf{v}\rangle + b\langle\mathbf{u}|\mathbf{w}\rangle.$$

This can be thought of as the "dot product" of vectors, for finite-dimensional Hilbert spaces.

In the context of the Stern–Gerlach experiment, recall that we said $|\uparrow_y\rangle$ and $|\downarrow_y\rangle$ form a basis for the two-dimensional irrep containing the state. It turns out that

$$\begin{aligned} |\uparrow_x\rangle &= \frac{1}{\sqrt{2}}|\uparrow_y\rangle + \frac{1}{\sqrt{2}}|\downarrow_y\rangle \\ |\downarrow_x\rangle &= \frac{1}{\sqrt{2}}|\uparrow_y\rangle - \frac{1}{\sqrt{2}}|\downarrow_y\rangle . \end{aligned}$$

Measurement of spin in y is *incompatible* with measurement of spin in x, because these are two different bases!

Exercise. The mathematical way of expressing how measurements work encodes a measurement as a linear transformation $\mathbf{A} \colon \mathcal{H} \to \mathcal{H}$. If $\mathcal{H}$ is finite-dimensional, $\mathbf{A}$ can be thought of as a matrix. Then possible values of the measurement, along with the resulting collapsed state, are encoded as pairs of **eigenvalues** and **eigenvectors** of $\mathbf{A}$. Read a little about how this works, and write down (in the standard basis) the matrices corresponding to the Stern–Gerlach y and x boxes.

3.4 Entanglement

If the Hilbert space of a one-particle system is $\mathcal{H}$, what happens if we consider a system with two such (identical) particles? In particular, what is the Hilbert space of the two-particle system? The naive guess is $\mathcal{H} \oplus \mathcal{H}$, the direct sum, which turns out to be wrong. The right answer is the *tensor product* $\mathcal{H} \otimes \mathcal{H}$.

Example 3.6. Consider a particle on the real line $\mathbb{R}$. Its Hilbert space is

$$\mathcal{H}_{\text{1-particle}} = L^2(\mathbb{R}),$$

consisting of states $|\psi\rangle$ where $\psi(x)$ is a wavefunction. If we have *two* particles on the real line, the composite state should be some wavefunction $\psi(x, y)$ which depends on both the position x of the first particle and the position y of the second particle. So we should have

$$\mathcal{H}_{\text{2-particle}} = L^2(\mathbb{R}^2).$$

One can show mathematically that

$$L^2(\mathbb{R}^2) = L^2(\mathbb{R}) \otimes L^2(\mathbb{R}).$$

(This is true more generally for $L^2(X \times Y)$, not just $\mathbb{R}^2$)

The structure of the tensor product $\mathcal{H} \otimes \mathcal{H}$ has some interesting consequences when we think about measurement. Consider the states in the Stern–Gerlach experiment, namely the two-dimensional $\mathcal{H}$ for a spin-1/2 particle. For simplicity fix either the

x basis or the y basis, and write $|\uparrow\rangle$ and $|\downarrow\rangle$ as a basis for $\mathcal{H}$. Then a basis for the two-particle Hilbert space is

$$|\uparrow\uparrow\rangle\,, |\uparrow\downarrow\rangle\,, |\downarrow\uparrow\rangle\,, |\downarrow\downarrow\rangle\,.$$

Here $|\uparrow\uparrow\rangle$ means $|\uparrow\rangle \otimes |\uparrow\rangle$. Now imagine constructing a superposition

$$\frac{1}{\sqrt{2}}\,|\uparrow\uparrow\rangle + \frac{1}{\sqrt{2}}\,|\downarrow\downarrow\rangle\,.$$

This is called a **Bell state**, which is where the two particles are "maximally entangled". **Quantum entanglement** is the general term used when the state of one part of the system is *not independent* of the state of another part.

To understand what this means, imagine creating such an entangled state in the lab and then separating the two particles by a very large distance, e.g. by putting the other particle in a different galaxy. When we measure the particle in the lab, as usual we have a 50% chance of getting spin up or down. But regardless of what result we get, the moment we perform the measurement we know the other particle *must* be in the same state, because of wavefunction collapse. This collapse of the state of the other particle happens *instantaneously*, without reference to the speed of light. Such a phenomenon was first considered by Einstein, Podolsky, and Rosen in 1935, and was known as the **EPR paradox** because it seems to involve *faster than light* effects, contradicting the theory of relativity.

Exercise. Convince yourself that the EPR paradox is not actually a paradox, because no *information* is being transmitted faster than light. (Hint: imagine trying to use an entangled pair to communicate; how do you control what information is being received on the other side?)

3.5 Time evolution

Given a state $|\mathbf{v}\rangle \in \mathcal{H}$, how do we evolve it forward in time? Suppose we want to move forward in time by t units. Then there must be some operator $\mathbf{U}_t \colon \mathcal{H} \to \mathcal{H}$ which takes a state and produces the new state after time t.

- Since $\mathcal{H}$ is a vector space, $\mathbf{U}_t$ should preserve this structure. So $\mathbf{U}_t$ is a *linear* operator.
- Since quantum states are supposed to be normalized to 1, i.e. have total probability 1, the time evolution operator $\mathbf{U}_t$ should not change the norms of vectors.
- We should have $\mathbf{U}_t\mathbf{U}_s = \mathbf{U}_{t+s}$.

Definition 3.7. Working with vector spaces and norms over $\mathbb{C}$, a linear operator which preserves norms of vectors is called **unitary**. The space of all unitary operators on a vector space V is denoted $U(V)$, or $U(n)$ if $\dim V = n < \infty$.

Hence the time evolution operators $\mathbf{U}_t$ are a collection of unitary operators satisfying $\mathbf{U}_t\mathbf{U}_s = \mathbf{U}_{t+s}$ for all $t, s \in \mathbb{R}$. Such a collection of operators forms a subgroup of the group of *all* unitary operators. This subgroup contains elements which are specified by a single parameter t. In general, such subgroups are called **one-parameter subgroups**.

Proposition 3.8. *All one-parameter subgroups of $U(V)$ are of the form*

$$\mathbf{U}_t = \exp(i\mathbf{H}t)$$

for some linear (but not unitary) operator $\mathbf{H}\colon V \to V$.

Here when we write the exponential of a *matrix*, we mean to take the *matrix exponential.* This just means to write $\exp(x)$ as the series $1 + x + x^2/2 + \cdots$, and to plug in the matrix into the series.

Definition 3.9. The matrix $\mathbf{H}$ is called the **Hamiltonian** of the quantum system, and governs its time-evolution.

The way to summarize all this is to view the state $|\mathbf{v}\rangle$ of the system as a function of time, i.e. $|\mathbf{v}(t)\rangle$, and to take as an axiom of quantum mechanics that

$$i\frac{d}{dt}|\mathbf{v}(t)\rangle = \mathbf{H}|\mathbf{v}(t)\rangle,$$

called the **Schrödinger equation**. This is a *differential equation* for the unknown $\mathbf{v}(t)$, and has the unique solution

$$|\mathbf{v}(t)\rangle = e^{i\mathbf{H}t}|\mathbf{v}(0)\rangle.$$

In our notation, this is exactly saying the time-evolution operator is $U_t = e^{i\mathbf{H}t}$.

Suppose now that the quantum system has a G-symmetry, so $\mathcal{H}$ is a representation of G. Then the time evolution operator $\mathbf{U}_t$ should preserve the symmetry, in the sense that

$$\mathbf{U}_t(g \cdot |\mathbf{v}\rangle) = g \cdot (\mathbf{U}_t|\mathbf{v}\rangle).$$

By definition, this means U_t is an *intertwiner.* One way to think about this is to remember the action of g is by linear operators. If $\phi\colon G \to \mathrm{GL}(\mathcal{H})$ is the rep, then $\mathbf{U}_t$ being an intertwiner means it commutes with all the operators $\phi(g)$.

Schur's lemma says intertwiners $\mathcal{H} \to \mathcal{H}$ cannot do very much beyond multiplying irreps by scalars. Since we normalize states, multiplication by an overall scalar does not change states. Hence if we decompose

$$\mathcal{H} = \bigoplus_n \mathcal{H}_n$$

into irreps, then $\mathbf{U}_t$ preserves each $\mathcal{H}_n$. Namely, the type of a state, where "type" here means which irrep of this decomposition it lives in, is unchanged by time evolution.

This argument can be repeated for *any* linear operator $\mathbf{A}\colon \mathcal{H} \to \mathcal{H}$ which commutes with the Hamiltonian $\mathbf{H}$ (and therefore with $\mathbf{U}_t$). Quantities and measurements arising from such operators $\mathbf{A}$ are therefore *unchanged* over time, and give rise to conservation laws. For example, the eigenspaces of **H** have eigenvalues which are the **energies** of those states. That **H** commutes with itself says that energy is conserved over time.

The only way to change the "type" of a state is via an operator $\mathcal{H} \to \mathcal{H}$ which is *not* an intertwiner. In other words, only operators which *do not* commute with **H** can cause energy levels, or other conserved quantities, to change. In the symmetry group $\mathrm{SO}(3, 1)$ of spacetime, it turns out Lorentz boosts do not commute with **H** in general. Hence the apparent energy of a system changes depending on which relativistic reference frame we use. This forms the basis for mass-energy equivalence in the theory of relativity.

4 Lie groups and their representations

The symmetry groups we encountered so far in our discussion of physics were all *infinite* groups. This is typical in physics because symmetries in the real world are usually *continuous*. In fact, these infinite symmetry groups are usually have more structure than just the group structure: they are also *geometric spaces*. For example, we saw that $U(1)$, the group of rotational symmetries of the circle S^1, is essentially S^1 itself with a group operation. So first, we need to clarify what is meant by "geometric space".

Definition 4.1. A real (or complex) **(smooth) manifold** M is a space which is *locally isomorphic* to $\mathbb{R}^n$ (or $\mathbb{C}^n$). Here, "locally" means if we zoom in far enough around any point in M. The **dimension** of the manifold is the integer n.

Example 4.2. To get a feel for what it means to be a smooth manifold, here are some two-dimensional smooth manifolds (also called **surfaces**):

1. (real) the sphere S^2;
2. (real) the torus, i.e. donut, T^2;
3. (real) the real plane $\mathbb{R}^2$;
4. (complex) the complex plane $\mathbb{C}^2$;
5. (complex) the complex version of the sphere, given by $x^2 + y^2 + z^2 = 1$ in $\mathbb{C}^3$.

Note that since $\mathbb{C} \cong \mathbb{R}^2$, a complex manifold of dimension n is also a real manifold of dimension $2n$.

Example 4.3. Here are some (real) spaces which are *not* smooth surfaces:

1. (wrong dimension) $\mathbb{R}^3$, since it has the wrong dimension;
2. (singularity) the quadric cone $x^2 + y^2 = z$ in $\mathbb{R}^3$, since no region around the cone point $(0, 0, 0)$ looks anything like $\mathbb{R}^2$;
3. (boundary) the square $[0, 1] \times [0, 1]$, because zooming in on points in the boundary cannot yield anything like $\mathbb{R}^2$, which has no boundary.

Definition 4.4. A **Lie group** is a group which is also a smooth manifold.

To understand more deeply the symmetries of nature, we must therefore understand the representation theory of certain Lie groups. The general theory for arbitrary Lie groups is fairly hard. First we'll restrict our attention to certain types of Lie groups. Then we'll see that studying their reps is almost the same as studying the reps of an associated object called the *Lie algebra*.

4.1 Matrix Lie groups

Most Lie groups arising in nature come from matrices. More precisely, most of them arise as subgroups of $\mathrm{GL}(n)$, the group of all invertible $n \times n$ matrices. Such Lie groups are called **matrix Lie groups**. (Since they are all Lie groups on their own, they are all *Lie subgroups* of $\mathrm{GL}(n)$, not just subgroups.)

Example 4.5. Let $\mathrm{GL}(n, \mathbb{R})$ be the group of invertible *real* $n \times n$ matrices.

- The **orthogonal group** $O(n)$ is the Lie subgroup of all matrices preserving the norm $\|\mathbf{v}\|^2 = v_1^2 + \cdots + v_n^2$ of vectors. Equivalently, it is all matrices $\mathbf{A}$ such that $\mathbf{A}^T\mathbf{A} = \mathbf{I}$. These matrices are called **orthogonal**.

- The **special linear group** $\mathrm{SL}(n, \mathbb{R})$ (or $\mathrm{SL}(n)$ when it is clear from context what scalars we use) is the Lie subgroup of all matrices with $\det \mathbf{A} = 1$.

- The **special orthogonal group** $\mathrm{SO}(n)$ is the Lie subgroup which is the intersection $O(n) \cap \mathrm{SL}(n)$, i.e. all orthogonal matrices with determinant 1. We saw it before as the rotational symmetry group of the sphere S^{n-1}.

Example 4.6. Let $\mathrm{GL}(n, \mathbb{C})$ be the group of invertible *complex* $n \times n$ matrices.

- The **unitary group** $U(n)$ is the Lie subgroup of all matrices preserving the norm $\|\mathbf{v}\|^2 = |v_1|^2 + \cdots + |v_n|^2$, where $|z|^2 = z\bar{z}$ for complex numbers. Equivalently, it is all matrices $\mathbf{A}$ such that $\mathbf{A}^\dagger\mathbf{A} = \mathbf{I}$.

- The definitions of $\mathrm{SL}(n, \mathbb{C})$ and $\mathrm{SU}(n)$ are analogous to the real case.

To see why these matrix groups are actually Lie groups, i.e. why they are smooth manifolds, we start with $\mathrm{GL}(n, \mathbb{R})$. By taking an $n \times n$ matrix and "flattening" it into an $n^2 \times 1$ vector, we can interpret the set of invertible matrices as a subset of $\mathbb{R}^{n^2}$, which is clearly a smooth manifold. Only very special matrices are not invertible, namely those with $\det = 0$, and removing them from $\mathbb{R}^{n^2}$ is like removing a curve from $\mathbb{R}^2$: the resulting space is still a smooth manifold, of the same dimension. Of course, the same argument works with $\mathrm{GL}(n, \mathbb{C})$.

For other matrix Lie groups, we view them as being "cut out by equations" in the smooth manifold $\mathrm{GL}(n)$. For example, $\mathrm{SL}(n) \subset \mathrm{GL}(n)$ is the subset defined by the equation $\det \mathbf{A} = 1$, just like $S^1 \subset \mathbb{R}^2$ is defined by $x^2 + y^2 = 1$.

Proposition 4.7 (Regular value theorem)**.** *Let M be a smooth manifold of dimension n. If a function $f\colon M \to \mathbb{R}$ has non-zero gradient ∇f at all points where $f = 0$, then the subset of all such points is a smooth submanifold of dimension $n - 1$.*

Applying the regular value theorem to the function $f(\mathbf{A}) = \det \mathbf{A} - 1$ in $\mathrm{GL}(n)$ immediately shows that $\mathrm{SL}(n)$ is also a manifold, but of dimension

$$\dim_{\mathbb{R}} \mathrm{SL}(n) = \dim_{\mathbb{R}}(\mathrm{GL}(n)) - 1 = n^2 - 1.$$

We write $\dim_{\mathbb{R}}$ to clarify that we mean *real* dimension.

Exercise. Check that the equation $\mathbf{A}^T\mathbf{A} = \mathbf{I}$ defining $O(n)$ is actually a *system* of $n(n+1)/2$ distinct equations. Conclude that

$$\dim_{\mathbb{R}} O(n) = \frac{n(n-1)}{2}.$$

Exercise. Explain why

$$\dim_{\mathbb{R}} \mathrm{SL}(n, \mathbb{R}) = \dim_{\mathbb{R}} \mathrm{GL}(n, \mathbb{R}) - 1$$

but $\dim_{\mathbb{R}} \mathrm{SO}(n) = \dim_{\mathbb{R}} O(n)$.

Exercise. Show that $U(1) \cong \mathrm{SO}(2)$, as Lie groups. Compute that $\dim_{\mathbb{R}} U(n) = n^2$ and show that $U(n)$ and $\mathrm{SO}(n+1)$ are *not* isomorphic for $n > 1$.

4.2 Tangent spaces

Studying representations of Lie groups is much harder than studying representations of finite (or discrete) groups. This is because now there are infinitely many elements in the group, which interact with each other in some highly non-linear way. For example, SO(2) is the same as a circle, geometrically. But Lie groups are very special geometric spaces. The key observation is the following.

> Lie groups are *homogeneous* spaces: no matter where in the space you are, the surrounding region looks exactly the same.

In technical terms, this is because given two points $g_1, g_2 \in G$, there is a *symmetry* of G itself given by multiplying (in G) by $g_2 g_1^{-1}$. This multiplication transports everything around the point g_1 to stuff around the point g_2.

It should not be too surprising, then, that to study a Lie group G is nearly the same as studying some region U around the identity element $e \in G$. In fact, by the same argument, we can make U as small as we want. By shrinking U further and further, at some point it becomes basically a bent piece of the ordinary flat space $\mathbb{R}^N$. By that point, U is indistinguishable from the "linear approximation" to G at e, which is an example of a *tangent space*.

Definition 4.8. Given a manifold M, the **tangent space** at $p \in M$ is the space of all tangent vectors at p. This tangent space is written T_pM.

Example 4.9. Consider the circle S^1, thought of as the unit circle in the plane. The tangent space at the point $p = (1, 0)$ is a vertical line. We say

$$T_{(1,0)}S^1 = \mathbb{R}^1.$$

Note that $S^1 \cong \mathrm{SO}(2)$ is also a Lie group, and therefore homogeneous. This is reflected in how actually $T_pS^1 = \mathbb{R}^1$ for *any* point $p \in S^1$.

The tangent space T_pM is the *linear approximation* to the manifold M at the point p. It generalizes the idea of the tangent line to a curve, from calculus, to higher dimensions. The following observation is very important.

Proposition 4.10. *T_pM is a vector space of dimension* $\dim M$*.*

One way to compute T_pM is to find all the possible tangent vectors at p. If we consider a curve $\mathbf{x}(t)$ such that $\mathbf{x}(0) = p$, i.e. passing through p at time $t = 0$, then $\nabla\mathbf{x}(0)$ is a tangent vector at p.

Example 4.11. For the circle S^1, since all points are of the form $e^{i\theta}$, curves passing through $p = (1, 0)$ are of the form

$$z(t) = e^{if(t)}$$

for *any* function $f : \mathbb{R} \to \mathbb{R}$ such that $f(0) = 0$. Hence

$$z'(t) = e^{if(t)}f'(t)$$

and plugging in $t = 0$ gives $z'(0) = f'(0)$. Since f can be any function, $f'(0)$ can be any real number. So the space of all possible tangent vectors is $\mathbb{R}$, exactly as we saw earlier.

One can perform this exact same procedure for Lie groups. We'll do it for matrix Lie groups, where we can start by finding $T_e\,\mathrm{GL}(n)$ and use it to describe T_eG for other matrix Lie groups $G \subset \mathrm{GL}(n)$. Note that for matrix Lie groups, the identity element is the identity matrix $\mathbf{I}$.

Proposition 4.12. *The tangent space $T_e\,\mathrm{GL}(n)$ is the vector space of all $n \times n$ matrices, called* Mat_n*.*

Proof. We saw earlier that $\dim \mathrm{GL}(n) = n^2$. It is not hard to see that $\mathrm{Mat}_n \cong \mathbb{R}^{n^2}$ also has dimension n^2. Hence if we can show that any $n \times n$ matrix can arise as a tangent vector in $T_e\,\mathrm{GL}(n)$, then we have shown $T_e\,\mathrm{GL}(n) = \mathrm{Mat}_n$.

Let X be an $n \times n$ matrix. Consider the matrices $\mathbf{A}(t) = \mathbf{I} + t\mathbf{X}$. For small enough t, these matrices are all invertible. So $\mathbf{A}(t)$ is a curve in $\mathrm{GL}(n)$ passing through $\mathbf{A}(0) = \mathbf{I}$, which is the identity element of the group. Since

$$\mathbf{A}'(0) = \mathbf{X},$$

it follows that $\mathbf{X}$ is a tangent vector. □

Now suppose $G \subset \mathrm{GL}(n)$ is a matrix Lie group. It will consist of invertible $n \times n$ matrices satisfying some conditions, expressed as certain equations, e.g. $\mathbf{A}^T\mathbf{A} = \mathbf{I}$ for the orthogonal group $O(n)$, or $\det(\mathbf{A}) = 1$ for $\mathrm{SL}(n)$. Each such constraint on the *group element* $\mathbf{A}$ puts a corresponding constraint on the *tangent vector* $\mathbf{X} = \mathbf{A}'(0)$, given precisely by the *derivative* of the equation.

Example 4.13. Consider the group $O(n)$ of orthogonal matrices. Let $\mathbf{A}(t)$ be a curve whose tangent vector at $t = 0$ is $\mathbf{X}$. Since $\mathbf{A}(t)$ is a curve in $O(n)$,

$$\mathbf{A}(t)^T\mathbf{A}(t) = \mathbf{I}$$

for all t. The derivative of this equation with respect to t, by the product rule, is

$$\mathbf{A}'(t)^T\mathbf{A}(t) + \mathbf{A}(t)^T\mathbf{A}'(t) = \mathbf{0}.$$

At $t = 0$ we have $\mathbf{A}'(0) = \mathbf{X}$ and $\mathbf{A}(0) = \mathbf{I}$. Hence plugging in $t = 0$ gives

$$\mathbf{X}^T + \mathbf{X} = \mathbf{0}.$$

It follows that

$$T_eO(n) = \{\mathbf{X} \in \mathrm{Mat}_n(\mathbb{R}) \mid \mathbf{X}^T = -\mathbf{X}\}.$$

Such matrices are called **skew-symmetric**, because symmetric matrices are those where $\mathbf{X}^T = \mathbf{X}$ (without the minus sign).

Exercise. Differentiate the constraint for $U(n)$ to show

$$T_eU(n) = \{\mathbf{X} \in \mathrm{Mat}_n(\mathbb{C}) \mid \mathbf{X}^\dagger = -\mathbf{X}\}.$$

These are **skew-Hermitian** matrices.

Exercise. It is a fact that, as polynomials in the variable t,

$$\det(\mathbf{I} + t\mathbf{X}) = 1 + \mathrm{tr}(\mathbf{X})t + \cdots$$

where the dots are higher-degree terms in t. Use this to show

$$T_e\,\mathrm{SL}(n) = \{\mathbf{X} \in \mathrm{Mat}_n \mid \mathrm{tr}(\mathbf{X}) = 0\}.$$

These are **traceless** matrices.

4.3 Lie algebras

To better study the tangent spaces T_eG, we need to be more careful about which curves $\mathbf{x}(t)$ we use. It turns out that, for theoretical purposes, there are better curves to use than things like $\mathbf{I} + t\mathbf{X}$. Ideally we want some curve $\mathbf{x}(t)$ such that

1. $\mathbf{x}'(0) = \mathbf{X}$, but more importantly,
2. $\mathbf{x}(t) \in G$ for *all* $t \in \mathbb{R}$, not just for small enough t around $t = 0$.

For example, if in GL(1) we wanted a curve with tangent vector $\big(-1\big)$, it is true that for sufficiently small t the curve $1 + t \cdot (-1)$ is invertible, i.e. non-zero. But once we get to $t = 1$, suddenly the curve "falls outside" of GL(1).

Proposition 4.14. *For matrix Lie groups G, the curve*

$$\mathbf{x}(t) = \exp(t\mathbf{X})$$

has these two properties, where exp *is the matrix exponential.*

Definition 4.15. The **matrix exponential** is analogous to the exponential of a scalar, which as a series is

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots .$$

The exponential of a matrix $\mathbf{Z}$ is obtained by plugging $\mathbf{Z}$ into this series:

$$\exp(\mathbf{Z}) = \mathbf{I} + \mathbf{Z} + \frac{\mathbf{Z}^2}{2!} + \frac{\mathbf{Z}^3}{3!} + \cdots .$$

For us, it is not important how to actually compute $\exp(\mathbf{Z})$; it suffices to know its definition via series. Note that

$$\exp(t\mathbf{X}) = \mathbf{I} + t\mathbf{X} + \cdots ,$$

where $\cdots$ are higher-degree terms in t which can be thought of as "corrections" to the naive curve $\mathbf{I} + t\mathbf{X}$.

Exercise. Show that $e^{t\mathbf{A}}e^{s\mathbf{A}} = e^{(t+s)\mathbf{A}}$ by matching coefficients in t.

Exercise. Show that $e^{\mathbf{A}}e^{\mathbf{B}} \neq e^{\mathbf{A}+\mathbf{B}}$ in general, by checking that the t^2 terms don't match. (Hint: remember that in the expansion

$$(\mathbf{A} + \mathbf{B})^2 = \mathbf{A}^2 + \mathbf{AB} + \mathbf{BA} + \mathbf{B}^2$$

it is *not true* that $\mathbf{AB} = \mathbf{BA}$.)

Now remember that Lie groups are more than just manifolds; they have a group structure. Given two elements $\mathbf{A}, \mathbf{B} \in G$ we can always multiply them to get $\mathbf{AB}$. What does this mean at the level of tangent vectors? In other words, given two curves $\mathbf{A}(t) = \exp(t\mathbf{X})$ and $\mathbf{B}(t) = \exp(t\mathbf{Y})$, we should be able to write

$$\mathbf{A}(t)\mathbf{B}(t) = \exp(t\mathbf{X})\exp(t\mathbf{Y}) = \exp(\mathbf{Z}(t))$$

for some matrix-valued function $\mathbf{Z}(t)$.

Proposition 4.16.

$$\exp(t\mathbf{X})\exp(t\mathbf{Y}) = \exp\big(t\,(\mathbf{X} + \mathbf{Y}) + t^2\,(\mathbf{XY} - \mathbf{YX}) + \cdots\big) \tag{7}$$

where $\cdots$ means higher-degree terms in t.

Definition 4.17. The **commutator** or **Lie bracket** of two matrices $\mathbf{X}$ and $\mathbf{Y}$ is

$$[\mathbf{X}, \mathbf{Y}] = \mathbf{XY} - \mathbf{YX}.$$

So the t^2 term in the formula (7) is exactly $[\mathbf{X}, \mathbf{Y}]$. If it and all other higher-degree terms were zero, then the group multiplication would exactly become addition in the tangent space. Hence the t^2 and higher-degree terms are actually measuring the *discrepancy* between G and its linear approximation $T_e G$. The main contribution to this discrepancy is measured by the commutator, and the commutator is the most important structure on $T_e G$.

Definition 4.18. A **Lie algebra** is a vector space V together with a commutator operation $[\mathbf{v}, \mathbf{w}]$ on vectors.

Example 4.19. The Lie algebra associated to a Lie group G is denoted $\mathfrak{g}$. In general, Fraktur letters are used to denote Lie algebras: the Lie algebra of $\mathrm{GL}(n)$ is $\mathfrak{gl}(n)$, the Lie algebra of $O(n)$ is $\mathfrak{o}(n)$, the Lie algebra of $\mathrm{SL}(n)$ is $\mathfrak{sl}(n)$, etc.

In other words, the tangent space $\mathfrak{g} = T_e G$ of a Lie group is not just a vector space: it is a Lie algebra. The commutator turns out to capture *everything* about group multiplication, not just the t^2 term, because of the following (fairly deep) theorem.

Theorem 4.20 (Baker–Campbell–Hausdorff formula)**.** *All higher-order terms in* (7) *can actually be written purely in terms of the commutator.*

Because of this theorem, to understand representations of the Lie group G, it (mostly) suffices to understand representations of its associated Lie algebra $\mathfrak{g}$. In some sense, a Lie algebra representation is like a linear approximation to a Lie group representation. This is why the two concepts are subtly different, as follows.

Definition 4.21. A **Lie algebra representation** is a homomorphism of *Lie algebras*

$$\rho \colon \mathfrak{g} \to \mathfrak{gl}(n) = \mathrm{Mat}_n .$$

In English, this means an assignment of an $n \times n$ matrix (not necessarily invertible!) to every element of the Lie algebra $\mathfrak{g}$, such that computing the commutator of $X, Y \in \mathfrak{g}$ is the same as computing the commutator of the matrices $\rho(X), \rho(Y)$.

4.4 SU(2) and SL(2, $\mathbb{C}$)

It is highly instructive to see all this theory in the case of the Lie group SU(2). This is because SU(2) is basically the simplest *non-abelian* Lie group (and also we'll see later that it is very important in physics). Recall that SU(2) consists of all complex 2×2 matrices $\mathbf{A}$ satisfying:

1. (special linear) $\det \mathbf{A} = 1$.
2. (unitary) $\mathbf{A}^\dagger \mathbf{A} = \mathbf{I}$;

Exercise (Lie algebra)**.** Show that the Lie algebra $\mathfrak{su}(2)$ consists of all 2×2 matrices

$$\mathbf{X} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

such that $a + d = 0$ and

$$\begin{pmatrix} \bar{a} & \bar{c} \\ \bar{b} & \bar{d} \end{pmatrix} = - \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

where $\bar{x}$ means complex conjugation.

In other words, $a = -d$ must be purely imaginary, and if $b = \beta + i\gamma$ then $c = -\beta + i\gamma$. Write $a = i\alpha$ so that $d = -i\alpha$. Putting this together, all matrices in $\mathfrak{su}(2)$ are of the form

$$\mathbf{X} = \alpha \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} + \beta \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} + \gamma \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} \tag{8}$$

for *real* scalars $\alpha, \beta, \gamma \in \mathbb{R}$. This is the proof of the following.

Proposition 4.22. *The Lie algebra* $\mathfrak{su}(2)$*, as a real vector space, has a basis*

$$\mathbf{Z} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}.$$

Exercise. Compute the commutation relations

$$[\mathbf{Z}, \mathbf{U}] = 2\mathbf{V}, \quad [\mathbf{U}, \mathbf{V}] = 2\mathbf{Z}, \quad [\mathbf{V}, \mathbf{Z}] = 2\mathbf{U}. \tag{9}$$

From the general theory we discussed, to understand the representation theory of SU(2) is (mostly) the same as understanding the representation theory of $\mathfrak{su}(2)$. From the proposition and the exercise, an n-dimensional representation $\rho\colon \mathfrak{su}(2) \to \mathrm{Mat}_n$ is just a choice of three $n \times n$ matrices

$$\rho(\mathbf{Z}), \quad \rho(\mathbf{U}), \quad \rho(\mathbf{V})$$

which must satisfy the commutation relations (9). Abstractly, we think of $\mathfrak{su}(2)$ as the vector space spanned by symbols $\mathbf{Z}, \mathbf{U}, \mathbf{V}$ satisfying the prescribed commutation relations.

Exercise. Check that the homomorphism $\rho\colon \mathfrak{su}(2) \to \mathrm{Mat}_3$ given by

$$\rho(\mathbf{Z}) = \begin{pmatrix} 0 & -2 & 0 \\ 2 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \rho(\mathbf{U}) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -2 \\ 0 & 2 & 0 \end{pmatrix}, \quad \rho(\mathbf{V}) = \begin{pmatrix} 0 & 0 & 2 \\ 0 & 0 & 0 \\ -2 & 0 & 0 \end{pmatrix}$$

is a three-dimensional representation of $\mathfrak{su}(2)$, sometimes called the **spin-**1 representation. Explain why it is an irrep.

Exercise. Explain why our construction of $\mathfrak{su}(2)$ automatically gives a two-dimensional representation, which is called the **spin-**1/2 representation. Explain why it is an irrep.

Exercise (Hard)**.** Find a four-dimensional irrep of $\mathfrak{su}(2)$.

For convenience, and also because the general theory works out easier this way, we will take the *complex* vector space with these basis elements $\mathbf{Z}, \mathbf{U}, \mathbf{V}$. This means we are allowed to take α, β, γ to be *complex* scalars in (8). Of course, this new vector space is much bigger than the previous $\mathfrak{su}(2)$. For example, it contains

$$i\mathbf{Z} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix},$$

which is *not* in $\mathfrak{su}(2)$.

Definition 4.23. If V is a real vector space with basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$, then the **complexification** of V is often called $V_{\mathbb{C}}$ and is defined as the *complex* vector space with the same basis.

Exercise (Hard)**.** Show that $V_{\mathbb{C}} = V \otimes \mathbb{C}$, viewing $\mathbb{C}$ as a two-dimensional real vector space with basis $\{1, i\}$.

The representation theory of $\mathfrak{g}$ is literally the same as the representation theory of $\mathfrak{g}_{\mathbb{C}}$, as long as in the latter case we use complex vector spaces. In the case of $\mathfrak{su}(2)$, we can pick a slightly better basis for the complexification:

$$\mathbf{H} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \mathbf{E} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

Exercise. Check that $\mathbf{H}, \mathbf{E}, \mathbf{F}$ can be written in terms of $\mathbf{Z}, \mathbf{U}, \mathbf{V}$, and vice versa, using complex scalars. Hence they form a basis of $\mathfrak{su}(2)_{\mathbb{C}}$. Then check the commutators

$$[\mathbf{H}, \mathbf{E}] = 2\mathbf{E}, \quad [\mathbf{H}, \mathbf{F}] = 2\mathbf{F}, \quad [\mathbf{E}, \mathbf{F}] = \mathbf{H}. \tag{10}$$

Exercise. Show that $\mathfrak{su}(2)_{\mathbb{C}} \cong \mathfrak{sl}(2, \mathbb{C})$.

It turns out that the commutation relations (10) for $\mathfrak{sl}(2, \mathbb{C})$ are significantly better to work with than the commutation relations (9) for $\mathfrak{su}(2)$. In some sense, $\mathbf{H}$ can be used to "label" vectors in a representation. Let V be a finite-dimensional irrep of $\mathfrak{sl}(2, \mathbb{C})$, and suppose $\mathbf{v} \in V$ is an eigenvector for $\mathbf{H}$, namely

$$\mathbf{H}\mathbf{v} = \lambda \mathbf{v}.$$

The eigenvalue λ is called the **weight** of $\mathbf{v}$. Then, because of our very special choice of basis for $\mathfrak{sl}(2, \mathbb{C})$, the key observation is we can systematically create new vectors in V with higher or lower weights.

Proposition 4.24. *If* $\mathbf{H}\mathbf{v} = \lambda \mathbf{v}$*, then the new vector* $\mathbf{w} = E\mathbf{v}$ *satisfies*

$$\mathbf{H}\mathbf{w} = (\lambda + 2)\mathbf{w}.$$

Proof. Using the commutation relation $[\mathbf{H}, \mathbf{E}] = 2\mathbf{E}$,

$$\mathbf{Hw} = \mathbf{HEv} = (2\mathbf{E} + \mathbf{EH})\mathbf{v} = 2\mathbf{w} + \mathbf{E}\lambda\mathbf{v} = (2 + \lambda)\mathbf{w}. \qquad \square$$

For this reason, $\mathbf{E}$ is often called a **raising operator**. Similarly, applying $\mathbf{F}$ *lowers* the weight. It is no coincidence that we called $\mathbf{H}$ the same name as the Hamiltonian: in physics, raising and lowering operators usually raise or lower the *energy* of a state, and the energy is just the eigenvalue of the Hamiltonian.

Theorem 4.25. *All k-dimensional irreps of $\mathfrak{sl}(2, \mathbb{C})$ have a basis of the form*

$$\mathbf{v}, \quad \mathbf{Fv}, \quad \mathbf{F}^2\mathbf{v}, \quad \dots, \quad \mathbf{F}^{k-1}\mathbf{v}$$

where $\mathbf{v}$ has weight $k - 1$. Hence they are all isomorphic. In other words, $\mathfrak{sl}(2, \mathbb{C})$ has exactly one distinct irrep for every dimension $k > 0$.

Proof sketch. Let V be a finite-dimensional irrep. It turns out that $\mathbf{H}$ is always represented by a *diagonalizable* matrix, so V splits into eigenspaces:

$$V = \bigoplus_{\lambda} V_\lambda.$$

Pick an eigenvector $\mathbf{v}$ with *highest weight* λ, so that there are no weights larger than λ in this decomposition. Then $\mathbf{Ev} = \mathbf{0}$, because otherwise it would be a vector with even higher weight. Now consider the sequence

$$\mathbf{v}, \quad \mathbf{Fv}, \quad \mathbf{F}^2\mathbf{v}, \quad \dots. \tag{11}$$

At some point, $\mathbf{F}^m\mathbf{w} = 0$. Otherwise we would keep producing distinct (linearly independent) vectors in V, and V would be infinite-dimensional.

Suppose this process produced $m + 1 < k$ vectors and we failed to generate the remaining vectors in V. Then actually the sub vector space generated by the vectors in (11) form a non-trivial sub-representation, contradicting that V is an irrep. Hence (11) must span *all* of V, and therefore gives a basis of V.

The weight of $\mathbf{v}$ is computed as follows. Suppose $\mathbf{Hv} = \lambda\mathbf{v}$. Then

$$0 = \mathbf{EF}^k\mathbf{v} = (\mathbf{H} + \mathbf{FE})\mathbf{F}^{k-1}\mathbf{v} = (\lambda - 2(k-1))\mathbf{F}^{k-1}\mathbf{v} + \mathbf{FEF}^{k-1}\mathbf{v}.$$

One can now repeat this process on $\mathbf{EF}^{k-1}\mathbf{v}$, and so on. The end result is

$$0 = ((\lambda - 2(k-1)) + (\lambda - 2(k-2)) + \dots + (\lambda - 2 \cdot 0))\,\mathbf{v}.$$

Since $\mathbf{v} \neq 0$, it must be that

$$0 = k\lambda - 2\frac{k(k-1)}{2}.$$

Solving, $\lambda = k - 1$. $\square$

This theorem describes *all* irreps of $\mathfrak{sl}(2,\mathbb{C})$. Namely, there is a k-dimensional one for every positive integer k, and we visualize it as a sequence of eigenvectors of $\mathbf{H}$ related by the raising and lowering operators $\mathbf{E}$ and $\mathbf{F}$.

$$\begin{tikzcd}
\boxed{-k+1} \arrow[r, bend right, "\mathbf{E}"'] & \boxed{-k+3} \arrow[l, bend right, "\mathbf{F}"'] \arrow[r, bend right, "\mathbf{E}"'] & \cdots \arrow[l, bend right, "\mathbf{F}"'] \arrow[r, bend right, "\mathbf{E}"'] & \boxed{k-5} \arrow[l, bend right, "\mathbf{F}"'] \arrow[r, bend right, "\mathbf{E}"'] & \boxed{k-3} \arrow[l, bend right, "\mathbf{F}"'] \arrow[r, bend right, "\mathbf{E}"'] & \boxed{k-1} \arrow[l, bend right, "\mathbf{F}"']
\end{tikzcd}$$

4.5 Semisimple theory

This beautiful picture of how $\mathfrak{sl}(2,\mathbb{C})$ looks actually generalizes to a fairly big class of Lie algebras, called the **semisimple** Lie algebras. The idea is that

> semisimple Lie algebras are built from many copies of $\mathfrak{sl}(2,\mathbb{C})$, which can interact with each other in some non-trivial but highly-restricted ways.

The *smallest* semisimple Lie algebra is $\mathfrak{sl}(2,\mathbb{C})$. All others have higher dimension. So there are *multiple* raising/lowering operators, denoted $\mathbf{E}_\alpha$ and $\mathbf{F}_\alpha$ as α ranges over something called the *root system*. Similarly, there are multiple Hamiltonians $\mathbf{H}_\alpha$.

Exercise (Hard)**.** Consider $\mathfrak{sl}(3,\mathbb{C})$. By analogy with $\mathfrak{sl}(2,\mathbb{C})$, write down a basis

$$\mathbf{H}_{12}, \mathbf{H}_{23}, \mathbf{E}_{12}, \mathbf{E}_{13}, \mathbf{E}_{23}, \mathbf{F}_{12}, \mathbf{F}_{13}, \mathbf{F}_{23}$$

for $\mathfrak{sl}(3,\mathbb{C})$ such that:

1. $\{\mathbf{H}_{12}, \mathbf{E}_{12}, \mathbf{F}_{12}\}$ is a copy of $\mathfrak{sl}(2,\mathbb{C})$;
2. $\{\mathbf{H}_{23}, \mathbf{E}_{23}, \mathbf{F}_{23}\}$ is a copy of $\mathfrak{sl}(2,\mathbb{C})$;
3. $\{\mathbf{H}_{13}, \mathbf{E}_{13}, \mathbf{F}_{13}\}$ is a copy of $\mathfrak{sl}(2,\mathbb{C})$, where $\mathbf{H}_{13} = \mathbf{H}_{12} + \mathbf{H}_{23}$;
4. all the $\mathbf{H}$ commute with each other, i.e.

$$[\mathbf{H}_{12}, \mathbf{H}_{23}] = [\mathbf{H}_{12}, \mathbf{H}_{13}] = [\mathbf{H}_{13}, \mathbf{H}_{23}] = 0.$$

To understand an irrep V of $\mathfrak{sl}(2,\mathbb{C})$, we decomposed V into eigenspaces for $\mathbf{H}$, each labeled by an eigenvalue λ. Now there are multiple $\mathbf{H}_\alpha$ and they all commute, so we can decompose V into eigenspaces for *all* of them. Each such eigenspace will be labeled by eigenvalues $(\lambda_\alpha)_\alpha$, one for each $\mathbf{H}_\alpha$. Such a collection of eigenvalues is still called a **weight**.

Example 4.26. For $\mathfrak{sl}(3,\mathbb{C})$, irreps are classified by two integers (k_{12}, k_{13}) (the eigenvalues of $\mathbf{H}_{12}$ and $\mathbf{H}_{23}$).

- $\mathbf{E}_{12}$ and $\mathbf{F}_{12}$ raises/lowers k_{12} by ± 2.

- $\mathbf{E}_{23}$ and $\mathbf{F}_{23}$ raises/lowers k_{23} by ± 2.
- $\mathbf{E}_{13}$ and $\mathbf{F}_{13}$ raises/lowers (k_{12}, k_{23}) by $(\pm 1, \pm 1)$.

Hence, to understand a semisimple Lie algebra $\mathfrak{g}$, it suffices to understand how all these $\mathfrak{sl}(2,\mathbb{C})$'s are packaged together in $\mathfrak{g}$. It turns out one only needs to understand the collection of $\mathfrak{sl}(2,\mathbb{C})$'s corresponding to *linearly independent* $\mathbf{H}$'s. For example, in $\mathfrak{sl}(3,\mathbb{C})$, it suffices to understand how

$$\{\mathbf{H}_{12}, \mathbf{E}_{12}, \mathbf{F}_{12}\} \text{ and } \{\mathbf{H}_{23}, \mathbf{E}_{23}, \mathbf{F}_{23}\}$$

interact. How $\{\mathbf{H}_{13}, \mathbf{E}_{13}, \mathbf{F}_{13}\}$ fits into $\mathfrak{sl}(3,\mathbb{C})$ is fully determined by this.

Definition 4.27. Let $\mathfrak{g}$ be a semisimple Lie algebra, with a choice of linearly independent Hamiltonians $\mathbf{H}_\alpha$. The **Cartan matrix** of a semisimple Lie algebra $\mathfrak{g}$ is the matrix $\mathbf{C}$ whose $\alpha\beta$-th entry is the scalar $\mathbf{C}_{\alpha\beta}$ such that

$$[\mathbf{H}_\alpha, \mathbf{E}_\beta] = \mathbf{C}_{\alpha\beta}\mathbf{E}_\beta.$$

Note that this immediately implies

$$[\mathbf{H}_\alpha, \mathbf{E}_\beta] = -\mathbf{C}_{\alpha\beta}\mathbf{E}_\beta.$$

Example 4.28. The Cartan matrix for $\mathfrak{sl}(2,\mathbb{C})$ is the 1×1 matrix $\mathbf{C} = \begin{pmatrix} 2 \end{pmatrix}$, since there is only one Hamiltonian $\mathbf{H}$ and $[\mathbf{H}, \mathbf{E}] = 2\mathbf{E}$.

Exercise. Show that the Cartan matrix for $\mathfrak{sl}(3,\mathbb{C})$ is

$$\mathbf{C} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

Exercise (Hard)**.** Show that the Cartan matrix for $\mathfrak{sl}(4,\mathbb{C})$ is

$$\mathbf{C} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

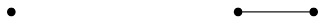
Any semisimple Lie algebra has an associated Cartan matrix, and the Cartan matrix determines everything about the Lie algebra. From the general structure theory of semisimple Lie algebras, Cartan matrices must satisfy certain *crystallographic relations*. These relations are very strict, to the point where it is straightforward to *classify* semisimple Lie algebras by writing down all possible Cartan matrices. This is usually expressed in a graphical form, as follows.

Definition 4.29. The **Dynkin diagram** associated to a Cartan matrix $\mathbf{C}$ is the graph drawn as follows.

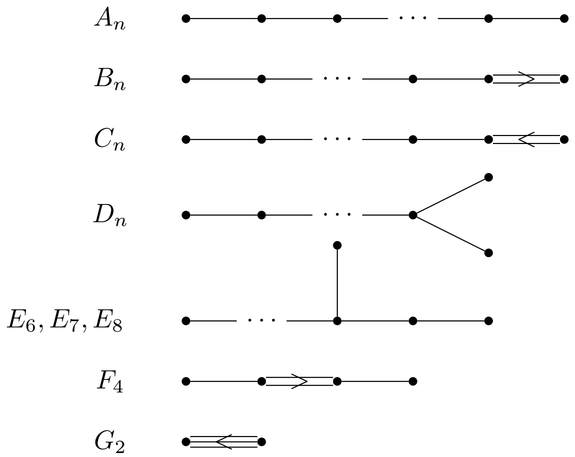
1. Draw a vertex corresponding to each row/column, i.e. each $\mathfrak{sl}(2,\mathbb{C})$.

2. Connect vertex i and vertex j with $\mathbf{C}_{ij}\mathbf{C}_{ji}$ edges.
3. If $\mathbf{C}_{ij} \neq \mathbf{C}_{ji}$, label the edge(s) with an arrow pointing toward j if $\mathbf{C}_{ji} > \mathbf{C}_{ij}$, and an arrow pointing toward i otherwise.

Example 4.30. Here are the Dynkin diagrams associated to $\mathfrak{sl}(2,\mathbb{C})$ and $\mathfrak{sl}(3,\mathbb{C})$.



Theorem 4.31 (Classification of semisimple Lie algebras)**.** *Every semisimple Lie algebra is given by one of the following Dynkin diagrams.*



Here, n is the number of vertices in the Dynkin diagram, also called the **rank** of the Lie group/algebra. The A, B, C, D families are infinite and are called the **classical** Lie groups/algebras. They are easy to describe as matrix Lie groups:

$$A_n = \mathfrak{sl}(n+1,\mathbb{C}), \quad B_n = \mathfrak{so}(2n+1,\mathbb{C}), \quad C_n = \mathfrak{sp}(2n,\mathbb{C}), \quad D_n = \mathfrak{so}(2n,\mathbb{C}).$$

On the other hand, E, F, G are **exceptional**, and it is difficult to describe them in terms of matrices. For example, although G_2 is rank two, its smallest irrep is 7-dimensional.

The representation theory of semisimple Lie algebras is essentially a generalization of the $\mathfrak{sl}(2,\mathbb{C})$ picture. Every finite-dimensional irrep consists of “chains” of vectors obtained by raising/lowering from a vector with highest weight.

Definition 4.32. A vector $v \in V$ is **highest weight** of weight $(\lambda_\alpha)_\alpha$ if

$$\mathbf{H}_\alpha \mathbf{v} = \lambda_\alpha \mathbf{v}, \quad \mathbf{E}_\alpha \mathbf{v} = 0$$

for all α. A representation V is a **highest weight rep** if it has a highest weight vector.

Theorem 4.33 (Reps of semisimple Lie algebras)**.** *Every finite-dimensional irrep of a semisimple Lie algebra is a highest weight rep. Two such irreps with the same highest weight are always isomorphic.*

Hence to talk about a specific irrep of a semisimple Lie algebra, it suffices to specify a highest weight. For $\mathfrak{sl}(2,\mathbb{C})$, this means to specify an integer $k \geq 1$. For $\mathfrak{sl}(3,\mathbb{C})$, this means to specify a pair of integers (k_1, k_2) with $k_1, k_2 \geq 1$.

5 Particle physics

It is time to put together all the theory we have learned, in order to discuss modern particle physics. The current state-of-the-art model is called the **standard model**. From a representation-theoretic perspective, there is a 1-particle state space

$$\mathcal{H}_{\text{Standard Model}} = \mathcal{H}_{\text{quarks}} \oplus \mathcal{H}_{\text{electron}} \oplus \cdots$$

consisting of all known fundamental particles, and $\mathcal{H}$ is a representation of some group G which represents all known symmetries in nature obeyed by the fundamental particles. This means we must identify all the relevant symmetries. Such symmetries come in two types: *spacetime* symmetries and *gauge* symmetries. So $\mathcal{H}$ will be a rep of a huge group

$$G_{\text{Standard Model}} = G_{\text{spacetime}} \times G_{\text{gauge}}.$$

Each G-irrep in $\mathcal{H}$ is a fundamental particle, and is therefore labeled by quantum numbers for all the different "Hamiltonians" $\mathbf{H}$ in G. We'll see that these quantum numbers are:

- spin, coming from the group $\mathrm{SO}(3,1)$ of rotational symmetries in $G_{\text{spacetime}}$;
- energy/momentum, coming from the group $\mathbb{R}^{3,1}$ of translations in $G_{\text{spacetime}}$;
- hypercharge, weak isospin, and color, coming from the gauge group $G_{\text{gauge}} = U(1) \times \mathrm{SU}(2) \times \mathrm{SU}(3)$.

5.1 Spacetime symmetries

Definition 5.1. Let $\mathbb{R}^{3,1}$ denote a four-dimensional vector space with coordinates (x, y, z, t) where lengths of vectors are measured using the **Lorentzian** norm

$$\|\mathbf{v}\|_{3,1}^2 = x^2 + y^2 + z^2 - t^2.$$

(The superscript in $\mathbb{R}^{3,1}$ comes from there being 3 plus signs and 1 minus sign.) In the same way that orthogonal group $O(n)$ is the Lie group of all matrices preserving norms in $\mathbb{R}^n$, define the **Lorentz group** $O(3,1)$ to be the Lie group consisting of all 4×4 invertible matrices preserving the Lorentzian norm $\|\mathbf{v}\|^2$.

That (flat) spacetime is $\mathbb{R}^{3,1}$ instead of $\mathbb{R}^4$ is Einstein's great insight, forming the basis for the theory of special relativity. Hence we should think of $O(3,1)$ as the group of "relativistic" rotational symmetries. A caveat is that symmetries should preserve the *orientation* of space(time), so actually we want to take $\mathrm{SO}(3,1)$ instead of $O(3,1)$. This has to do with preserving left- vs right-handedness.

Definition 5.2. The **Poincaré group** is the Lie group of all symmetries of spacetime, generated by $\mathrm{SO}(3,1)$ and all translations (in both space and time).

Since translations all commute with each other and the representation theory of abelian groups is not very interesting, we'll focus on the $\mathrm{SO}(3,1)$ part of the Poincaré group. There is a subgroup $\mathrm{SO}(3) \subset \mathrm{SO}(3,1)$ which consists of rotations of just the space component $\mathbb{R}^3$. This is ordinary rotational symmetry. One can compute that

$$\dim \mathfrak{so}(3,1) - \dim \mathfrak{so}(3) = \dim \mathfrak{so}(4) - \dim \mathfrak{so}(3) = 3,$$

so there is an additional 3-dimensional collection of "relativistic rotations" which mix space and time. These are **Lorentz boosts**.

Exercise. Show that in $\mathbb{R}^2$, the rotations

$$\mathbf{J}(\theta) = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

are the elements of $\mathrm{SO}(2)$, by explicitly checking that $\|\mathbf{J}(\theta)\mathbf{v}\|^2 = \|\mathbf{v}\|^2$. Analogously, show that in $\mathbb{R}^{1,1}$, the "Lorentz boosts"

$$\mathbf{K}(\theta) = \begin{pmatrix} \cosh\theta & \sinh\theta \\ \sinh\theta & \cosh\theta \end{pmatrix}$$

are the elements of $\mathrm{SO}(1,1)$, by explicitly checking that $\|\mathbf{K}(\theta)\mathbf{v}\|_{1,1}^2 = \|\mathbf{v}\|_{1,1}^2$. Here sinh and cosh are the **hyperbolic** sine and cosine.

Exercise (Hard)**.** Use the preceding exercise to write down matrices for:

- the three rotations $\mathbf{J}_x(\theta), \mathbf{J}_y(\theta), \mathbf{J}_z(\theta)$ around x, y, z in $\mathrm{SO}(3,1)$;
- the three Lorentz boosts $\mathbf{K}_x(\theta), \mathbf{K}_y(\theta), \mathbf{K}_z(\theta)$ along x, y, z in $\mathrm{SO}(3,1)$.

Differentiate these generators at $\theta = 0$ to get a basis $\mathbf{j}_x, \mathbf{j}_y, \mathbf{j}_z, \mathbf{k}_x, \mathbf{k}_y, \mathbf{k}_z$ for the Lie algebra $\mathfrak{so}(3,1)$. Check that their commutators are

$$\begin{gathered} [\mathbf{j}_x, \mathbf{j}_y] = \mathbf{j}_z, \quad [\mathbf{k}_x, \mathbf{k}_y] = -\mathbf{j}_z, \\ [\mathbf{j}_x, \mathbf{k}_y] = \mathbf{k}_z, \quad [\mathbf{j}_y, \mathbf{k}_x] = -\mathbf{k}_z, \\ [\mathbf{j}_x, \mathbf{k}_x] = 0, \end{gathered}$$

and all cyclic permutations of (x, y, z) thereof.

As with $\mathfrak{su}(2)$ vs $\mathfrak{sl}(2,\mathbb{C})$, we should complexify $\mathfrak{so}(3,1)$ to get $\mathfrak{so}(3,1)_{\mathbb{C}}$. After complexification, i.e. after permitting the use of complex scalars, a clever change of basis to take is

$$\mathbf{A}_k = \frac{1}{2}(\mathbf{J}_k + i\mathbf{K}_k), \quad \mathbf{B}_k = \frac{1}{2}(\mathbf{J}_k - i\mathbf{K}_k).$$

This basis is nice because the $\mathbf{A}$'s become "unrelated" to the $\mathbf{B}$'s, as can be seen by checking commutators.

Exercise. Check that

$$[\mathbf{A}_x, \mathbf{A}_y] = i\mathbf{A}_z, \quad [\mathbf{B}_x, \mathbf{B}_y] = i\mathbf{B}_z, \quad [\mathbf{A}_x, \mathbf{B}_x] = 0 \tag{12}$$

and all cyclic permutations of (x, y, z) thereof.

Proposition 5.3.

$$\mathfrak{so}(3,1)_{\mathbb{C}} \cong \mathfrak{sl}(2,\mathbb{C}) \oplus \mathfrak{sl}(2,\mathbb{C}).$$

Proof. The commutation relations (12) show that the **A** form a copy of $\mathfrak{su}(2)_{\mathbb{C}}$, by comparing with the commutation relations (9). (One can always rescale the basis to get rid of scalars.) The complexification is necessary because we are allowing complex scalars. Finally, we showed previously that $\mathfrak{su}(2)_{\mathbb{C}} \cong \mathfrak{sl}(2,\mathbb{C})$. Hence the **A** form a copy of $\mathfrak{sl}(2,\mathbb{C})$, and so do the **B**. □

Since irreps of $\mathfrak{sl}(2,\mathbb{C})$ are specified by a single number k, it follows that irreps of $\mathfrak{so}(3,1)_{\mathbb{C}}$ are specified by a pair of numbers k_1, k_2. These numbers are known as **spin**.

5.2 Spin: fermions and bosons

Recall that we classified *all* irreps of $\mathfrak{sl}(2,\mathbb{C})$: there is one for every dimension $k \geq 1$, consisting of a chain of states which can be raised or lowered by **E** and **F**.

Definition 5.4. The k-dimensional irrep of $\mathfrak{sl}(2,\mathbb{C})$, which has highest weight vector of weight $k-1$, is called the **spin** $(k-1)/2$ **irrep**.

A fundamental particle is an irrep V in the standard model state space $\mathcal{H}$. In particular, V is a representation of the group of spacetime symmetries $\mathfrak{sl}(2,\mathbb{C}) \times \mathfrak{sl}(2,\mathbb{C})$. Hence, in addition to other quantum numbers, V is labeled by two spins s_L and s_R. (The L and R stand for "left" and "right".)

Definition 5.5. A fundamental particle V is a **left-handed Weyl spinor** if $(s_L, s_R) = (1/2, 0)$, and similarly is **right-handed** if $(s_L, s_R) = (0, 1/2)$.

Note that this whole business with having both s_L and s_R arises only because we take relativity into account, so that $\mathrm{SO}(3,1)$ is the rotational symmetry group. If we neglect relativity, then $\mathrm{SO}(3)$ is the rotational symmetry group, and one can show

$$\mathfrak{so}(3)_{\mathbb{C}} = \mathfrak{su}(2)_{\mathbb{C}} = \mathfrak{sl}(2,\mathbb{C}).$$

Then there is only one spin s.

Definition 5.6. For convenience, in the relativistic setting, we say that the spin (s_L, s_R) irrep of $\mathfrak{so}(3,1)_{\mathbb{C}}$ has **spin** $s = s_L + s_R$.

All known *matter* particles, like electrons and quarks, have spin $1/2$ in this sense. Since both $(1/2, 0)$ and $(0, 1/2)$ have spin $1/2$, this means all matter particles come in two flavors: left- and right-handed. So there is a "left-handed electron" and a "right-handed electron".

Spin is deeply related to the Pauli exclusion principle. To understand this, we take a short detour. Given a specific fundamental particle, we can take two of it and consider a *two-particle* state $\mathbf{v} \otimes \mathbf{w}$. This is distinct from $\mathbf{w} \otimes \mathbf{v}$, but only barely. Swapping two identical particles should not change any observable quantity of the system, so at least

$$\|\mathbf{v} \otimes \mathbf{w}\|^2 = \|\mathbf{w} \otimes \mathbf{v}\|^2.$$

This means that $\mathbf{v} \otimes \mathbf{w} = e^{i\theta} \cdot (\mathbf{w} \otimes \mathbf{v})$ for some θ. In other words, swapping particles multiplies the state by $e^{i\theta}$. Swapping twice multiplies by $e^{2i\theta}$. But swapping twice returns to the original state, so $e^{2i\theta} = 1$. It follows that $\theta = 0, \pi$, and so

$$\mathbf{v} \otimes \mathbf{w} = \pm \mathbf{w} \otimes \mathbf{v}.$$

Definition 5.7. Given a specific kind of fundamental particle:

1. if $\mathbf{v} \otimes \mathbf{w} = +\mathbf{w} \otimes \mathbf{v}$, the particle is a **boson**;

2. if $\mathbf{v} \otimes \mathbf{w} = -\mathbf{w} \otimes \mathbf{v}$, the particle is a **fermion**.

The distinction between bosons and fermions is extremely important, because of the case when $\mathbf{v} = \mathbf{w}$. A fermionic particle must then satisfy

$$\mathbf{v} \otimes \mathbf{v} = -\mathbf{v} \otimes \mathbf{v},$$

and the only way for this to happen is if $\mathbf{v} \otimes \mathbf{v} = 0$. In other words, it is impossible to have two fermionic particles in *exactly* the same state. This is often stated as the **Pauli exclusion principle**. On the other hand, there can be many bosons all in the exact same state.

Theorem 5.8 (Spin-statistics theorem)**.** *Particles with integer spin are bosons, and particles with half-integer spin are fermions.*

Since all (known) matter particles are spin 1/2, they must all obey the Pauli exclusion principle.

5.3 Gauge theories

So far, we have discussed *external* symmetries of fundamental particles; these symmetries come from symmetries of spacetime itself. It turns out that fundamental particles also have *internal* symmetries, much like how a sphere has a rotational symmetry because of its shape and not because of any spacetime symmetries. These internal symmetries are called **gauge symmetries**.

To understand what gauge symmetries are, it helps to revisit an piece of history. In the early 20th century, physicists were struggling to understand the **strong force**. This is the force which binds protons and neutrons together in the nucleus of an atom. It is strong only at short distances ($\approx 10^{-15}$ m) and overcomes the electric repulsion between the positively charged protons. Neutrons experience the strong force as well, since they are also bound to the nucleus. In 1932, Heisenberg proposed that the proton and neutron were actually two different states of the *same* particle, called a **nucleon**, to explain why the strong force acts on both (almost) identically. The state of a nucleon would then be

$$\alpha_1 \left|\text{proton}\right\rangle + \alpha_2 \left|\text{neutron}\right\rangle ,$$

and Heisenberg believed there were physical processes which could change protons into neutrons. Soon after, Yukawa predicted that such processes were mediated by another particle, now called the **pion**. Such a particle was discovered, and comes in three kinds: π^+ (positive charge), π^0 (neutral), and π^- (negative charge). Nucleons interact with pions in processes which preserve electric charge:

$$\pi^- + p \to n, \quad \pi^+ + n \to p, \quad \pi^0 + p \to p, \quad \pi^0 + n \to n.$$

Other physicists quickly caught on to this idea, and proposed that under this model for the strong force, protons and neutrons should be interchangeable. In other words, there must be a symmetry $p \leftrightarrow n$ which swaps protons with neutrons. Because of linearity (or superposition), this symmetry must actually allow (almost) *arbitrary* changes of basis

$$\alpha p + \beta n \leftrightarrow \alpha' p + \beta' n.$$

This change of basis can be written as a 2×2 matrix. Quantum mechanics dictates that such a physical process actually must be a matrix in SU(2), so the state space of a nucleon should have an SU(2) **gauge symmetry**. The nucleon was declared to be the spin-1/2 rep under this SU(2),

- the proton was the **isospin up** (or $I_3 = 1/2$) state, and
- the neutron was the **isospin down** (or $I_3 = -1/2$) state.

For this symmetry to truly be a symmetry of the model, isospin must be conserved across interactions. This means that for pions,

$$I_3(\pi^+) = 1, \quad I_3(\pi^0) = 0, \quad I_3(\pi^-) = 1.$$

If $\mathcal{H}_{\text{nucleon}} = \mathbb{C}^2$ is the state space of a nucleon, interaction with a pion is therefore an intertwiner

$$\mathcal{H}_{\text{nucleon}} \to \mathcal{H}_{\text{nucleon}}.$$

In fact, we can view the pions π^+, π^- as raising/lowering operators. In this way, pions act on $\mathcal{H}_{\text{nucleons}}$ in the same way that $\mathfrak{su}(2)_{\mathbb{C}} = \mathfrak{sl}(2, \mathbb{C})$ does. Actually,

$$\mathcal{H}_{\text{pion}} \cong \mathfrak{sl}(2, \mathbb{C}).$$

Note also that $\mathcal{H}_{\text{pion}}$ is itself an SU(2)-irrep in this model. Since it is 3-dimensional, it must be the spin-1 rep.

Exercise (Hard)**.** Show that the matrix Lie group SU(2) acts on its own Lie algebra $\mathfrak{su}(2)$ by **conjugation**:

$$g \cdot X = gXg^{-1},$$

where $g \in \mathrm{SU}(2)$ and $X \in \mathfrak{su}(2)$. (This actually works for *any* Lie group and associated Lie algebra and is called the **adjoint rep**.) Hence $\mathfrak{su}(2)$ is an irrep for SU(2). Explain why it is the spin-1 irrep.

This exercise is how one would go about predicting the existence of *three* pions, along with their isospins.

Today, we know that the isospin "symmetry" is only approximate, because protons and neutrons are *not* fundamental particles. Instead, they are made of **quarks**. Specifically,

$$p = uud, \quad n = udd$$

where u is an **up quark** and d is a **down quark**. Since protons have charge +1 and neutrons have charge 0, it is not hard to figure out that

$$\mathrm{charge}(u) = \frac{2}{3}, \quad \mathrm{charge}(d) = -\frac{1}{3}.$$

In the same manner as for the isospin model of protons and neutrons, it turns out that each type of quark has an SU(3) **gauge symmetry**. This is the true gauge symmetry group for the strong force. Each up/down quark therefore comes in three **colors**: r (red), g (green), and b (blue). In other words, the state of an up quark is of the form

$$\alpha_1 \left|u^r\right\rangle + \alpha_2 \left|u^g\right\rangle + \alpha_3 \left|u^b\right\rangle,$$

where u^r, u^g, u^b are red, green and blue up quarks, and similarly for down quarks. (The superscripts are just labels, not actual exponents.) Note that the naming is not because quarks have actual color; they are just labels for the various states.

The SU(3) strong force is mediated by fundamental particles called **gluons**, just like the isospin model had pions as mediators. They live in the adjoint rep $\mathfrak{su}(3)_{\mathbb{C}}$, just like pions lived in the adjoint rep $\mathfrak{su}(2)_{\mathbb{C}}$. Interactions between gluons and quarks are SU(3)-intertwiners $\mathcal{H}_{\text{quark}} \to \mathcal{H}_{\text{quark}}$.

Exercise. Check that $\dim \mathfrak{su}(n) = n^2 - 1$. Conclude that there are eight "distinct" gluons.

So the existence of gauge symmetries helps us in two ways: it organizes "equivalent" particles (e.g. red/green/blue quarks) together, but also helps describe fundamental *forces*. For example, when we say the strong force is mediated by the gluon, this means quarks are held together within the nucleus by an attractive force arising from the exchange of gluons between quarks. For this reason, we say the strong force is an SU(3) **gauge theory**.

5.4 The standard model

We can now describe the standard model. It describes all fundamental particles and all four fundamental forces except gravity. (There is no known quantum theory for gravity yet.) The remaining three fundamental forces are all described by gauge theories.

- Electromagnetism is an $U(1)$ gauge theory, mediated by the **photon** γ;
- The weak force is an SU(2) gauge theory, mediated by the W^+**,** W^- **and** Z **bosons**.
- The strong force, as we have seen, is an SU(3) gauge theory, mediated by eight **gluons** g. (There are no conventional names for the eight "distinct" states.)

All particles which mediate forces are bosons, and we do not consider them as matter particles. There is an additional boson which mediates an interaction by which particles gain mass; this is the **Higgs boson** and the **Higgs mechanism**. It has spin 0, and all the other bosons have spin 1. Hypothetically, a quantum theory of gravity would involve a spin-2 mediator particle called the **graviton**.

Al fundamental matter particles are irreps of the symmetry group $G = G_{\text{spacetime}} \times G_{\text{gauge}}$. Aside from energy/momentum, we have identified that G consists of four groups.

- The rotational symmetry SO(3, 1) of spacetime, whose irreps are labeled by spins (s_L, s_R). Since all particles is spin-1/2, namely $s_L + s_R = 1/2$, it suffices to specify whether they are left-handed or right-handed.
- The $U(1)$ gauge group of electromagnetism, whose irreps are labeled by the **hypercharge** Y.
- The SU(2) gauge group of the weak force, whose irreps are labeled by the **weak isospin** I_3.
- The SU(3) gauge group of the strong force, whose irreps are labeled by *two* quantum numbers. There is no name in physics for these quantum numbers: all matter either is unchanged by color symmetry and therefore belongs to the trivial rep $\mathbb{C}$, or is some kind of quark and therefore belongs to the standard rep $\mathbb{C}^3$.

At this point, it is important to recognize that the hypercharge Y is *not* the electric charge Q of a particle, but rather obeys the **Gell-Mann–Nishijima formula**

$$Q = I_3 + Y/2.$$

For example, the up and down quarks have $I_3 = 1/2$ and $I_3 = -1/2$ respectively, and therefore both have hypercharges $Y = 1/3$. The reason for this weird relationship between Q and Y stems from a small lie in the story so far: it is *not true* that the photon is a basis of $\mathfrak{u}(1)_{\mathbb{C}}$ and the $W^{\pm}, Z$ bosons form a basis of $\mathfrak{su}(2)_{\mathbb{C}}$. Instead, the generators are

$$B \in \mathfrak{u}(1)_{\mathbb{C}}, \quad W_1, W_2, W_3 \in \mathfrak{su}(2)_{\mathbb{C}},$$

and the more familiar particles are linear combinations

$$\gamma = W_3 + \frac{B}{2}, \quad Z = W_3 - \frac{B}{2}, \quad W^\pm = W_1 \mp iW_2$$

inside $\mathfrak{u}(1)_\mathbb{C} \oplus \mathfrak{su}(2)_\mathbb{C}$. This combined $U(1) \times \mathrm{SU}(2)$ gauge theory is called the **electroweak** or **Weinberg–Salam** theory, and is valid at high energies. At low energies, **electroweak symmetry breaking** “ruins” part of this symmetry, and preserves only a $U(1)$ subgroup corresponding to the photon. Importantly, this $U(1)$ is *not* the one corresponding to B, but rather to γ, and this is the source of the distinction between charge and hypercharge. Consequently, the $W^\pm$ and Z bosons are also not the “correct” states to realize the SU(2) symmetry.

The Standard Model. We can now write down the *first generation* of particles in the standard model. Some notation: $\mathbb{C}_Y$ denotes the one-dimensional $U(1)$-irrep with hypercharge Y.

Chirality	Name	Symbol	Y	I_3	$U(1) \times \mathrm{SU}(2)$ rep	SU(3) rep
Left	Neutrino	ν_L	-1	$+\frac{1}{2}$	$\mathbb{C}_{-1} \otimes \mathbb{C}^2$	$\mathbb{C}$
Left	Electron	e^-_L	-1	$-\frac{1}{2}$		$\mathbb{C}$
Left	Up quarks	u^r_L, u^g_L, u^b_L	$\frac{1}{3}$	$+\frac{1}{2}$	$\mathbb{C}_{\frac{1}{3}} \otimes \mathbb{C}^2$	$\mathbb{C}^3$
Left	Down quarks	d^r_L, d^g_L, d^b_L	$\frac{1}{3}$	$-\frac{1}{2}$		$\mathbb{C}^3$
Right	Neutrino	ν_R	0	0	$\mathbb{C}_0 \otimes \mathbb{C}$	$\mathbb{C}$
Right	Electron	e^-_R	-2	0	$\mathbb{C}_{-2} \otimes \mathbb{C}$	$\mathbb{C}$
Right	Up quarks	u^r_R, u^g_R, u^b_R	$\frac{4}{3}$	0	$\mathbb{C}_{\frac{4}{3}} \otimes \mathbb{C}$	$\mathbb{C}^3$
Right	Down quarks	d^r_R, d^g_R, d^b_R	$-\frac{2}{3}$	0	$\mathbb{C}_{-\frac{2}{3}} \otimes \mathbb{C}$	$\mathbb{C}^3$

An immediate observation is that the weak force only acts on left-handed particles! All the right-handed ones have zero isospin and belong to the trivial irrep for the SU(2) gauge group of the weak force. This was a shocking discovery called the **parity violation** of the weak force.

Another observation is that the right-handed neutrino is not acted on by any of the fundamental forces. Actually, the right-handed neutrino is a hypothetical particle that has not yet been experimentally observed, but is strongly believed to exist because all other particles in the standard model come in left- and right-handed pairs.

It turns out this table is *not* all known fundamental particles. In the 1930s, a fundamental particle much heavier but otherwise completely identical to the electron was discovered experimentally. It was called the **muon** μ^-, and it came with its corresponding **muon neutrino** ν_μ. A few decades later, a pair of overweight up/down quarks, also otherwise identical to the up/down quarks, were discovered and called the **charm** and **strange** quarks. Today we know of *three generations* of the standard model. Each generation consists of exactly the same reps, but successively heavier and heavier.

First generation		Second generation		Third generation	
Name	Symbol	Name	Symbol	Name	Symbol
Up quark	u	Charm quark	c	Top quark	t
Down quark	d	Strange quark	s	Bottom quark	b
Electron neutrino	ν_e	Muon neutrino	ν_μ	Tau neutrino	ν_τ
Electron	e^-	Muon	μ^-	Tau	τ^-

Finally, it turns out every fermion has an associated **anti-particle**. If a particle is given by an irrep V, then the anti-particle corresponds to the *dual* representation $V^\vee$. In a very precise sense, an anti-particle can be viewed as the particle itself but traveling backward in time.

6 Beyond the standard model

The standard model successfully provides a unified description of three out of four fundamental forces, except gravity (which does not have a quantum theory yet), and all known fundamental particles. In the quest to find a unified "theory of everything", it is important to ask:

> what *constraints* are imposed by quantum field theory (QFT) on gauge groups and fundamental particles?

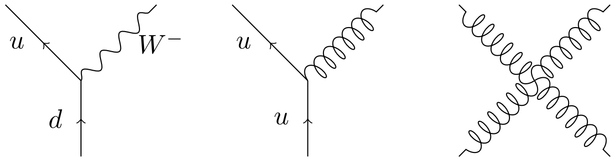
We can then try to expand or generalize the standard model while still satisfying these theoretical constraints, and see if the resulting theories can be supported by experiment. Such constraints come in two forms:

1. constraints on quantum numbers (like spin) of fundamental particles;
2. constraints on possible gauge groups of new or (further) unified forces.

To understand these constraints requires an understanding of how *interacting* QFTs work. Here, "interacting" means that every type of particle/field described by the theory has a non-trivial interaction with at least one other particle/field.

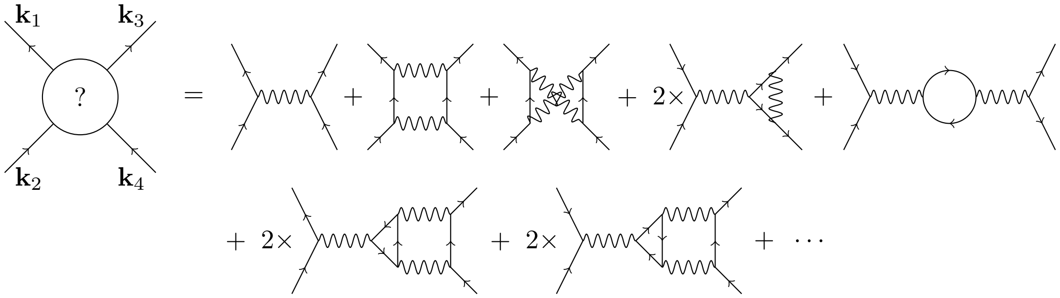
6.1 Perturbation theory and renormalization

Interacting QFTs work as follows. Every probability amplitude that we want to compute can be expressed as a sum of amplitudes of **Feynman diagrams**. A Feynman diagram is a pictorial representation of certain interactions between certain fundamental particles. Usually the vertical axis is time, and the horizontal axis is space. Here are Feynman diagrams for some fundamental interactions in the standard model.



Mathematically, Feynman diagrams are objects called *graphs*: they consist of **vertices** (the interactions) which are connected by **edges** (the propagating particles). One can make a list of all allowed vertices in the standard model, and they will all involve either three or four particles for reasons we will explain shortly.

The output of a QFT is are probability amplitudes for given physical scenarios. Usually one feeds into the QFT a set of incoming particles with some momenta, and asks for the probability amplitude of a set of outgoing particles with some other momenta. (Here, “momentum” means 4-*momentum*, i.e. a velocity (k_0, k_1, k_2, k_3) in space*time*.) A popular example is **electron-electron scattering**: we want to find the probability amplitude of two electrons with momenta $\mathbf{k}_1, \mathbf{k}_2$ going in and two electrons with momenta $\mathbf{k}_3, \mathbf{k}_4$ going out. The key idea is that, because we don't know what happens between “going in” and “going out”, we are allowed to fill in any Feynman diagram using allowed vertices in the standard model.



This is analogous to the Taylor series expansion of functions:

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots .$$

For Feynman diagrams, the number of **loops** is like the degree of x. The more loops, the less a diagram contributes to the total sum. In the Feynman diagram expansion of Compton scattering above, the first diagram has no loops (called a *tree-level* diagram), the next four are one-loop, and the remaining are two-loop. The process of computing amplitudes in this way is called **perturbation theory**.

The **Feynman rules** describe how to compute the amplitude of a given Feynman diagram. The recipe is simple.

- Since we don't know the momenta of all internal edges, we need to integrate over *all* possible momenta for those edges, keeping in mind that each interaction must conserve momentum.

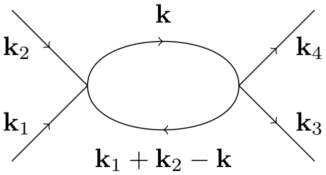
- Every edge contributes a **propagator** to the integrand. If the edge is a spin-s particle with momentum $\mathbf{k}$ and (rest) mass m, its propagator is roughly of the form
$$\frac{\mathbf{k}^{2s}}{|\mathbf{k}|^2 - m^2}. \tag{13}$$

However, it is very easy for such integrals to *diverge*! We can pinpoint two causes of divergences in the following examples.

Example 6.1 (Propagator divergence)**.** For particles of spin $s = 1$, there will be integrals that look like
$$\int_{\mathbb{R}^4} \frac{|\mathbf{k}|^2}{|\mathbf{k}|^2 - m^2}\, d\mathbf{k}.$$
As $|\mathbf{k}|^2 \to \infty$, the mass m^2 is negligible, and we can approximate the integrand as $|\mathbf{k}|^2/|\mathbf{k}|^2 = 1$. But the integral of 1 over an infinite volume diverges. Of course, this problem gets worse for spin $s > 1$.

We'll explain how to deal with this kind of divergence, for spin ≥ 1 particles, after we deal with the next kind, which is a more sophisticated problem.

Example 6.2 (Loop divergence)**.** Consider a toy QFT, called ϕ^4 **theory**, where the only type of particle is a spin-0 massive particle which interacts with itself via 4-particle vertices. The only 1-loop diagram is



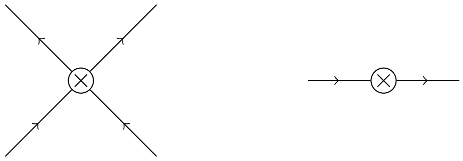
The amplitude of this diagram is (up to some constant factors)
$$\delta(\mathbf{k}_1 + \mathbf{k}_2 - \mathbf{k}_3 - \mathbf{k}_4) \int_{\mathbb{R}^4} \frac{1}{|\mathbf{k}|^2 + m^2} \frac{1}{|\mathbf{k}_1 + \mathbf{k}_2 - \mathbf{k}|^2 + m^2}\, d\mathbf{k}.$$
(The δ-function is just enforcing the condition that $\mathbf{k}_1 + \mathbf{k}_2 = \mathbf{k}_3 + \mathbf{k}_4$.) But this integral can be approximated by
$$\int_{\mathbb{R}^4} \frac{1}{|\mathbf{k}|^4}\, d\mathbf{k} \propto \left(\int_{-\infty}^{\infty} \frac{1}{k}\, dk\right)^4 = \infty.$$

The arduous work to systematically remove the second type of divergence was done in the 1970s, under the name of **renormalization**. The key idea behind renormalization is that the naive ideas of propagators and interactions are not accurate. A particle propagating from point a to point b can undergo all kinds of journeys, including *self-interactions*. These self-interactions change the propagator, and also the **coupling constants** of the theory, which describe things like electric charge and mass of particles.

The actual propagators, electric charges, masses, etc., called **dressed** quantities, are *not* equal to the original **bare** quantities. But the bare quantities are what appear in the theory. Hence it is necessary to "shift" the parameters of the original theory to get the physically correct theory. In terms of Feynman diagrams, this "shift" of the parameters of the theory essentially creates artificial vertices called **counterterm vertices** which turn out to exactly cancel the divergences in the original theory.

The counterterm vertices themselves are controlled by some fixed but unknown coupling constant, which then must be measured experimentally. For a given QFT to be **renormalizable**, then, the "shifting" of the theory must only produce finitely many counterterm vertices. Heuristically, a renormalizable theory has divergences which all stem from a finite number of Feynman diagrams. If there are infinitely many counterterms, we would have to run an infinite number of experiments to obtain all the coupling constants, and hence the theory loses predictive power.

Example 6.3. For the ϕ^4 theory of the previous example, one can show combinatorially that the only way for a Feynman diagram to diverge is if it contains the one-loop diagram in the previous example as a sub-diagram. Hence the ϕ^4 theory has two counterterm vertices.



The first cancels the divergence from the 4-particle interaction, and the second corrects the propagator.

Both the electroweak and the strong force turn out to have finitely many counterterms. However, one can do a fairly straightforward count of the number of **k**'s in the numerator vs denominator and show that for spin ≥ 2 particles, infinitely many counterterms will be necessary. This leads us to conclude that

a renormalizable QFT cannot involve fundamental particles of spin ≥ 2.

The problem with developing a quantum theory of gravity is that its mediator particle, the graviton, is necessarily spin 2. This comes from the theory of general relativity, where gravity is described by the Einstein field equations

$$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} + \Lambda g_{\mu\nu} = 8\pi T_{\mu\nu}.$$

On the left hand side are terms representing the curvature of spacetime, which dictates the gravitational field. The rhs is a quantity called the **stress-energy tensor**, which represents the matter/energy content in spacetime. In particular, it is a $(2,0)$-tensor, which means that

$$T_{\mu\nu} \in \underbrace{(V \otimes \cdots \otimes V)}_{\text{2 times}} \otimes \underbrace{(V^* \otimes \cdots \otimes V^*)}_{\text{0 times}}.$$

Here V is the natural 4-dimensional representation of $\mathfrak{so}(3,1) = \mathfrak{su}(2) \oplus \mathfrak{su}(2)$, called the **vector** representation or the spin-1 rep. It follows that elements of $V \otimes V$ are either spin-0 or spin-2. Spin-0 quantities are scalars, and so $T_{\mu\nu}$ is spin-2. In other words, the gravitational field is spin-2, and therefore the mediator particle must be as well.

In contrast, the classical theory of electromagnetism is described by Maxwell's equations

$$\partial_\nu F^{\mu\nu} = \mu_0 J^\mu$$

where F is the electromagnetic field strength, J^μ is the **four-current**, and μ_0 is a coupling constant called the **permeability** of space. The current J^μ carries only one index, and therefore is spin-1.

Now we can return to discussing the divergence in the propagator (13) for spin ≥ 1. It turns out this divergence can be canceled out as long as there is a corresponding conserved quantity, called a **current**, of the same spin in the theory. We say that the particle *couples* to the conserved current. For example, the divergence in the spin-1 propagator for the photon is dealt with by the conserved current J^μ. There is a similar current for the weak force, and the stress-energy tensor is the conserved current for the spin-2 graviton. In principle, a spin-3/2 particle is also renormalizable, but there is no current for it to couple to. This explains why all matter in the standard model is spin-1/2.

6.2 Grand unification

One direction in which we can explore physics beyond the standard model is to ask why the symmetry group of the standard model is

$$G_{\text{Standard model}} = G_{\text{spacetime}} \times G_{\text{gauge}}, \quad G_{\text{gauge}} = U(1) \times \mathrm{SU}(2) \times \mathrm{SU}(3).$$

Why are symmetries nicely separated into *external* and *internal* ones, and why are there no symmetries which mix them? For example, in the world of finite groups, there are two groups of order four: $\mathbb{Z}_2 \times \mathbb{Z}_2$ and $\mathbb{Z}_4$. Could it be that $G_{\text{Standard Model}}$ is more like $\mathbb{Z}_4$ than $\mathbb{Z}_2 \times \mathbb{Z}_2$? The famous "no-go theorem" of Coleman and Mandula in 1967 gives a negative answer to this question.

Theorem 6.4 (Coleman–Mandula)**.** *A non-trivial interacting quantum field theory (satisfying very mild assumptions) must have symmetry group*

$$G = G_{spacetime} \times G_{gauge},$$

i.e. with no "mixing" of the two.

Since we have already identified $G_{\text{spacetime}}$, the only remaining possibility is to generalize G_{gauge} to something bigger than but still containing $U(1) \times \mathrm{SU}(2) \times \mathrm{SU}(3)$. This has the added benefit of unifying different forces into a single force with a single gauge group, in the same way that $U(1) \times \mathrm{SU}(2)$ is the gauge group of the unified electroweak force. Starting in the 1970s, physicists called such a hypothetical theory a **grand unified theory (GUT)**.

Call the GUT gauge group G_{GUT}. In gauge theory, the only gauge groups we can use are semisimple Lie groups, whose Lie algebras are conveniently classified in Theorem 4.31. Whatever G_{GUT} is, it is built from these possibilities and contains the standard model's $U(1) \times \mathrm{SU}(2) \times \mathrm{SU}(3)$. We discuss some proposed GUTs.

$\mathrm{SU}(5)$ **model.** This is the most famous GUT, also called the **Georgi–Glashow model**. The simple Lie group $\mathrm{SU}(5)$ is the smallest which contains the standard model, via

$$\begin{aligned} U(1) \times \mathrm{SU}(2) \times \mathrm{SU}(3) &\to \mathrm{SU}(5) \\ (\gamma, \mathbf{A}, \mathbf{B}) &\mapsto \begin{pmatrix} \gamma^3\mathbf{A} & 0 \\ 0 & \gamma^{-2}\mathbf{B} \end{pmatrix}. \end{aligned}$$

Strictly speaking, this is *not* an inclusion because it is not one-to-one: all elements of the form

$$(\gamma, \gamma^{-3}\mathbf{I}, \gamma^2\mathbf{I}) \in U(1) \times \mathrm{SU}(2) \times \mathrm{SU}(3)$$

are sent to $\mathbf{I} \in \mathrm{SU}(5)$. These form a $\mathbb{Z}_6$. So, for the $\mathrm{SU}(5)$ theory to work, this $\mathbb{Z}_6$ must act trivially on everything in the standard model. Incredibly, the way hypercharges are arranged makes it so that the standard model passes this stringent test perfectly.

Unfortunately, the $\mathrm{SU}(5)$ theory was eventually discarded because of the following reason. The standard model representation embeds into an rep of $\mathrm{SU}(5)$ called $\wedge^*\mathbb{C}^5$, the exterior algebra of the defining representation. In particular, this means quarks are combined with electrons and neutrinos in a single irrep. Different particle types living in the same irrep must have mechanisms, coming from the adjoint rep $\mathfrak{su}(5)$, which transform one type into another, like how the weak force's $\mathfrak{su}(2)$ contains W bosons which transform up quarks to down quarks and vice versa. Hence the $\mathrm{SU}(5)$ allows for quarks to decay into electrons and neutrinos. Specifically, this allows for **proton decay**, a phenomenon which looks schematically like

$$p \to e^+ + 2\gamma.$$

However, proton decay has never been observed, and experiments show that the half-life of a proton is at least 10^{34} years. But the $\mathrm{SU}(5)$ theory has a maximum proton half-life of 10^{31} years.

$\mathrm{SO}(10)$ **theory.** There is a natural embedding $\mathrm{SU}(5) \to \mathrm{SO}(10)$, by viewing $\mathbb{C}^5$ as $\mathbb{R}^{10}$. Then the irrep $\wedge^*\mathbb{C}^5$ for the $\mathrm{SU}(5)$ theory embeds into the spin-1/2 irrep V of $\mathrm{SO}(10)$. This $\mathrm{SO}(10)$ theory has several advantages over the $\mathrm{SU}(5)$ theory.

- Since V is an *irrep*, it explains why a right-handed neutrino is necessary. Strictly speaking, a right-handed neutrino is not necessary at all for either the standard model or the $\mathrm{SU}(5)$ theory.

- It incorporates the earlier **Pati–Salam model** based on $\mathrm{SU}(4) \times \mathrm{SU}(2)_L \times \mathrm{SU}(2)_R$, which unifies the electrons/neutrinos with quarks by adding a new "color" for electrons/neutrinos and enlarging SU(3) to SU(4). The two copies of SU(2) correspond to a "left-handed weak force" and a "right-handed weak force".

- Its maximum proton half-life is 10^{35} years, which is above the current experimental lower bound.

Unfortunately, there are other reasons why both the SU(5) and SO(10) theories are questionable. By unifying different particles into the same irrep, both theories predict non-trivial relations between masses of various fundamental particles:

$$m_{\text{down quark}} \approx 9m_{\text{electron}}, \quad m_{\text{strange quark}} \approx m_{\text{muon}}, \quad m_{\text{bottom quark}} \approx 3m_{\text{tau muon}},$$

called the **Georgi–Jarlskog mass relations**. These masses have been measured very precisely, and only satisfy these relations very approximately.

E_6 **model.** Recall from the classification of semisimple Lie algebras that there are four infinite families A, B, C, D, and a few exceptional ones E_6, E_7, E_8, F_4, G_2. Out of the exceptional Lie algebras, only E_6 is a viable gauge group for GUTs. This is a constraint from representation theory.

Theorem 6.5. *Any irrep V of E_7, E_8, F_4, G_2 is isomorphic to its complex conjugate $\bar{V}$.*

But such irreps are necessary in order to have a concept of chirality for fermions, so that the weak force can act only on left-handed fermions. The only remaining possibility, E_6, actually naturally arises in the context of a type of string theory called $E_8 \times E_8$ heterotic. It naturally contains the SO(10) theory.

6.3 Supersymmetry

One theoretical solution to both the renormalization problem and the failure of unified models is to introduce supersymmetry. Supersymmetry is a hypothetical (external) symmetry which is *not* forbidden by the Coleman–Mandula theorem simply because it is *not* a Lie group symmetry. We must extend the concept of a Lie algebra to a super Lie algebra. In a super Lie algebra, there are additional generators Q_α, called **supercharges**, such that $Q_\alpha^2 = 0$. Instead of the commutator, we must take the *anti-commutator*

$$\{Q_\alpha, Q_\beta\} = Q_\alpha Q_\beta + Q_\beta Q_\alpha.$$

The supercharges have spin-1/2, and therefore change bosons to fermions, i.e.

$$Q\,|\text{boson}\rangle = |\text{fermion}\rangle,$$

and vice versa. Hence each fundamental particle has a **superpartner**. None have been experimentally observed yet. However, this immediately has consequences for a supersymmetric extension of the standard model.

- There is now a spin-3/2 current which is the superpartner of the spin-2 stress-energy tensor. Hence spin-3/2 superparticles can exist and be renormalizable.

- The existence of supersymmetry often takes care of non-renormalizability.

Definition 6.6. If there are n independent supercharges, we say the QFT has $\mathcal{N} = n$ **supersymmetry**. The resulting Lie algebra

$$\mathfrak{g}_{\mathrm{SUSY}} = \mathfrak{g}_{\mathrm{spacetime}} \ltimes \mathfrak{g}_{\mathrm{supercharges}}$$

is called the $\mathcal{N} = n$ **supersymmetry algebra**.

Irreps of the resulting Lie supergroup or superalgebra are **multiplets**, meaning that they consist of ordinary particles along with some super-particles. The supercharges do *not* commute with Lorentz boosts, so within a single multiplet can be different (super-)particles of different spins. But they do commute with everything else, so all particles in a multiplet have the same mass and internal quantum numbers.

One can ask: can the Lie super-algebra of symmetries be more complicated than just adjoining some spin-1/2 supercharges? The answer is no: there is an analogue of the Coleman–Mandula theorem in the supersymmetric setting.

Theorem 6.7 (Haag–Łopuszański–Sohnius)**.** *A non-trivial interacting supersymmetric QFT (satisfying very mild assumptions) must have symmetry group*

$$G = G_{SUSY} \times G_{gauge},$$

with all supercharges being spin-1/2*.*

Although there is no experimental evidence for supersymmetry, it remains a productive field of *mathematical* study, because the non-rigorous math of QFT can often be made fully rigorous in the supersymmetric setting. The study of supersymmetric QFT has yielded a lot of rich mathematics over the past half-century.