Geometry and Group Theory

ABSTRACT

In this course, we develop the basic notions of Manifolds and Geometry, with applications in physics, and also we develop the basic notions of the theory of Lie Groups, and their applications in physics.

Contents

The material in this course is intended to be more or less self contained. However, here is a list of some books and other reference sources that may be helpful for some parts of the course:

1. J.G. Hocking and G.S. Young, *Topology*, (Addison-Wesley, 1961). This is a very mathematical book on topological spaces, point-set topology, and some more advanced topics in algebraic topology. (Not for the faint-hearted!)

2. T. Eguchi, P.B. Gilkey and A.J. Hanson. *Gravitation, Gauge Theories and Differential Geometry*, Physics Reports, **66**, 213 (1980). This is a very readable exposition of the basic ideas, aimed at physicists. Some portions of this course are based fairly extensively on this article. It also has the merit that it is freely available for downloading from the web, as a PDF file. Go to http://www.slac.stanford.edu/spires/hep/, type "find a gilkey and a hanson", and follow the link to Science Direct for this article. Note that Science Direct is a subscription service, and you must be connecting from a URL in the tamu.edu domain, in order to get free access.

3. H. Georgi, *Lie Algebras and Particle Physics*, Perseus Books Group; 2nd edition (September 1, 1999). This is quite a useful introduction to some of the basics of Lie algebras and Lie groups, written by a physicist for physicists. It is a bit idiosyncratic in its coverage, but what it does cover is explained reasonably well.

4. R. Gilmore, *Lie Groups Lie Algebras and Some of Their Applications*, John Wiley & Sons, Inc (1974). A more complete treatment of the subject. Quite helpful, especially as a reference work.

1 Manifolds

One of the most fundamental constructs in geometry is the notion of a *Manifold*. A manifold is, in colloquial language, the arena where things happen. Familiar examples are the three-dimensional space that we inhabit and experience in everyday life; the surface of a ball, viewed as a two-dimensional closed surface on which, for example, an ant may walk; and the four-dimensional Minkowski spacetime that is the arena where special relativity may be formulated. In order to give a reasonably precise description of a manifold, it is helpful first to give a few rather formal definitions. It is not the intention in this course to make everything too formal and rigorous, so we shall try to strike a balance between formality and practical utility as we proceed. In particular, if things seem to be getting too abstract and rigourous at any stage, there is no need to panic, because it will probably just be a brief interlude before returning to a more intuitive and informal discussion.

In this spirit, let us begin with some formal definitions.

1.1 Some set-theoretic concepts

A set is a collection of objects, or elements; typically, for us, these objects will be points in a manifold. A set U is a subset of a set V if every element of U is also an element of V. If there exist elements in V that are not in the subset U, then U is called a *proper* subset of V.

If U is a subset of V then the *complement* of U in V, denoted by $V - U$, is the set of all elements of V that are not in U. If U is a subset but not a proper subset, then $V - U$ contains no elements at all. This set containing no elements is called the *empty set*, and is denoted by $\emptyset$. By definition, therefore, $\emptyset$ is a subset of every set.

The notion of the complement can be extended to define the *difference* of sets V and U, even when U is not a subset of V. Thus we have

$$V \backslash U = \{x : x \in V \text{ and } x \notin U\}\,. \tag{1.1}$$

If U is a subset of V then this reduces to the complement defined previously.

Two sets U and V are equal, $U = V$, if every element of V is an element of U, and vice versa. This is equivalent to the statement that U is a subset of V and V is a subset of U.

From two sets U and V we can form the *union*, denoted by $U \cup V$, which is the set of all elements that are in U *or* in V. The *intersection*, denoted by $U \cap V$, is the set of all elements that are in U *and* in V. The two sets U and V are said to be *disjoint* if $U \cap V = \emptyset$, i.e. they have no elements in common.

Some straightforwardly-established properties are:

$$
\begin{aligned}
&A \cup B = B \cup A\,, \qquad A \cap B = B \cap A\,,\\
&A \cup (B \cup C) = (A \cup B) \cup C\,, \qquad A \cap (B \cap C) = (A \cap B) \cap C\,, \qquad (1.2)\\
&A \cup (B \cap C) = (A \cup B) \cap (A \cup C)\,, \qquad A \cap (B \cup C) = (A \cap B) \cup (A \cap C)\,.
\end{aligned}
$$

If A and B are subsets of C, then

$$
\begin{aligned}
&C - (C - A) = A\,, \qquad C - (C - B) = B\,,\\
&A - (A \backslash B) = A \cap B\,,\\
&C - (A \cup B) = (C - A) \cap (C - B)\,,\\
&C - (A \cap B) = (C - A) \cup (C - B)\,. \qquad (1.3)
\end{aligned}
$$

1.2 Topological Spaces

Before being able to define a manifold, we need to introduce the notion of a topological space. This can be defined as a point set S, with open subsets $\mathcal{O}_i$, for which the following properties hold:

1. The union of any number of open subsets is an open set.

2. The intersection of a finite number of open subsets is an open set.

3. Both S itself, and the empty set $\emptyset$, are open.

It will be observed that the notion of an *open* set is rather important here. Essentially, a set X is open if every point x inside X has points round it that are also in X. In other words, every point in an open set has the property that you can wiggle it around a little and it is still inside X. Consider, for example, the set of all real numbers r in the interval $0 < r < 1$. This is called an *open interval*, and is denoted by $(0, 1)$. As its name implies, the open interval defines an open set. Indeed, we can see that for *any* real number r satisfying $0 < r < 1$, we can always find real numbers bigger than r, and smaller than r that still themselves lie in the open interval $(0, 1)$. By contrast, the interval $0 < r \le 1$ is *not* open; the point $r = 1$ lies inside the set, but if it is wiggled to the right by any amount, no matter how tiny, it takes us to a point with $r > 1$, which is not inside the set.

Given the collection $\{\mathcal{O}_i\}$ of open subsets of S, we can define the notion of the limit point of a subset, as follows. A point p is a limit point of a subset X of S provided that every open set containing p also contains a point in X that is distinct from p. This definition yields a topology for S, and with this topology, S is called a *Topological Space.*

Some further concepts need to be introduced. First, we define a *basis* for the topology of S as some subset of all possible open sets in S, such that by taking intersections and unions of the members of the subset, we can generate all possible open subsets in S. An *open cover* $\{U_i\}$ of S is a collection of open sets such that every point p in S is contained in at least one of the U_i. The topological space S is said to be *compact* if every open covering $\{U_i\}$ has a finite sub-collection $\{U_{i_1}, \cdots, U_{i_n}\}$ that also covers S.

Finally, we may define the notion of a *Hausdorff Space.* The topological space S is said to obey the Hausdorff axiom, and hence to be an Hausdorff Space, if, for any pair of distinct points p_1 and p_2 in S, there exist disjoints open sets $\mathcal{O}_1$ amd $\mathcal{O}_2$, each containing just one of the two points. In other words, for any distinct pair of points p_1 and p_2, we can find a small open set around each point such that the two open sets do not overlap.[1]

We are now in a position to move on to the definition of a manifold.

1.3 Manifolds

Before giving a formal definition of a manifold, it is useful to introduce what we will recognise shortly as some very simple basic examples. First of all, consider the real line, running from minus to plus infinity. Slightly more precisely, we consider the open interval $(-\infty, \infty)$, i.e. the set of points x such that $-\infty < x < \infty$. We denote this by the symbol $\mathbb{R}$ (the letter R representing the *real* numbers). In fact $\mathbb{R}$ is the prototype example of a manifold; it is a one-dimensional topological space parameterised by the points on the real line.

A simple extension of the above is to consider the n-dimensional space consisting of n copies of the real line. We denote this by $\mathbb{R}^n$. A familiar example is three-dimensional Euclidean space, with Cartesian coordinates (x, y, z). Thus our familiar three-dimensional space can be called the 3-manifold $\mathbb{R}^3$.

We can now give a formal definition of a smooth n-manifold, with a smooth atlas of charts, as

1. A topological space S

2. An open cover $\{U_i\}$, which are known as *patches*

[1]The great mathematician and geometer Michael Atiyah gave a nice colloquial definition: "A topological space is Hausdorff if the points can be housed off." One should imagine this being spoken in a rather plummy English accent, in which the word "off" is pronounced "orff."

3. A set (called an atlas) of maps $\phi_i : U_i \to \mathrm{I\!R}^n$ called charts, which define a 1-1 relation between points in U_i and points in an open ball in $\mathrm{I\!R}^n$, such that

4. If two patches U_1 and U_2 intersect, then both $\phi_1 \circ \phi_2^{-1}$ and $\phi_2 \circ \phi_1^{-1}$ are smooth maps from $\mathrm{I\!R}^n$ to $\mathrm{I\!R}^n$.

The set-up described here will be referred to as an n-dimensional manifold M. Sometimes we shall use a superscript or subscript n, and write M^n or M_n.

What does all this mean? The idea is the following. We consider a topological space S, and divide it up into patches. We choose enough patches so that the whole of S is covered, with overlaps between neighbouring patches. In any patch, say U_1, we can establish a rule, known as a mapping, between points in the patch and points in some open connected region (called an open ball) in $\mathrm{I\!R}^n$. This mapping will be chosen such that it is 1-1 (one to one), meaning that there is a unique invertible relation that associates to each point in U_1 a unique point in the open ball in $\mathrm{I\!R}^n$, and vice versa. We denote this mapping by ϕ_1. This is the notion of choosing *coordinates* on the patch U_1. See Figure 1.

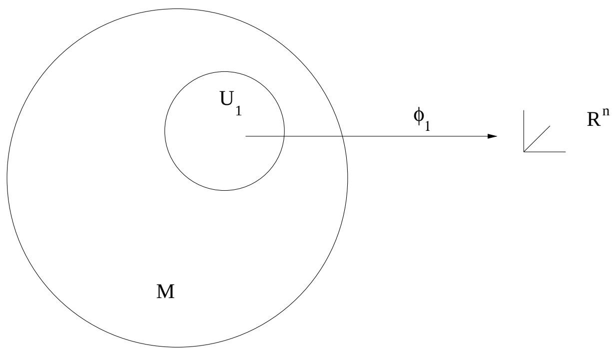

Figure 1: The map ϕ_1 takes points in U_1 into an open ball in $\mathrm{I\!R}^n$

Now consider another patch U_2, which has some overlap with U_1. For points in U_2 we make another mapping, denoted by ϕ_2, which again gives a 1-1 association with points in an open ball in $\mathrm{I\!R}^n$. Now, consider points in the topological space S that lie in the intersection of U_1 and U_2. For such points, we have therefore got two different rules for mapping into a copy of $\mathrm{I\!R}^n$: we have the mapping ϕ_1, and the mapping ϕ_2. We are therefore in a position to go back and forth between the two copies of $\mathrm{I\!R}^n$. Note that we can do this because each of ϕ_1 and ϕ_2 was, by definition, an invertible map.

We can start from a point in the open ball in the second copy of $\mathrm{I\!R}^n$, and then apply the inverse of the mapping ϕ_2, which we denote by ϕ_2^{-1}, to take us back to a point in the patch U_2 that is also in U_1. Then, we apply the map ϕ_1 to take us forward to the open ball in the first copy of $\mathbb{R}^n$. The composition of these two operations is denoted by $\phi_1 \circ \phi_2^{-1}$. Alternatively, we can go in the other order and start from a point in the open ball of the first copy of $\mathbb{R}^n$ that maps back using ϕ_1^{-1} to a point in U_1 that is also in U_2. Then, we apply ϕ_2 to take us into the second copy of $\mathbb{R}^n$. Going in this direction, the whole procedure is therefore denoted by $\phi_2 \circ \phi_1^{-1}$. See Figure 2.

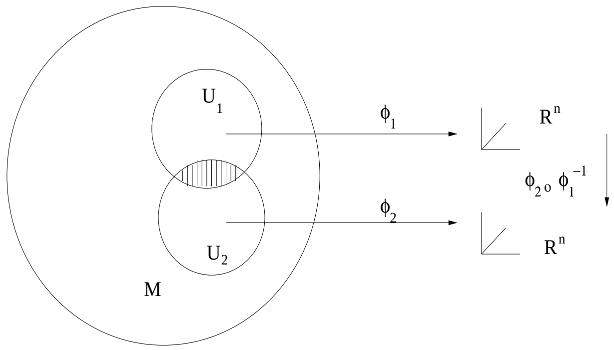

Figure 2: $\phi_2 \circ \phi_1^{-1}$ maps $\mathbb{R}^n$ into $\mathbb{R}^n$ for points in the intersection $U_1 \cap U_2$

Whichever way we go, the net effect is that we are mapping between a point in one copy of $\mathbb{R}^n$ and a point in another copy of $\mathbb{R}^n$. Suppose that we choose coordinates $(x^1, x^2, \cdots, x^n)$ on the first copy, and coordinates $(\tilde{x}^1, \tilde{x}^2, \cdots, \tilde{x}^n)$ on the second copy. Collectively, we can denote these by x^i, and $\tilde{x}^i$, where $1 \le i \le n$. In the first case, the composition $\phi_1 \circ \phi_2^{-1}$ is therefore giving us an expression for the x^i as functions of the $\tilde{x}^j$. In the second case, $\phi_2 \circ \phi_1^{-1}$ is giving us $\tilde{x}^i$ as functions of the x^j.

So far, we have discussed this just for a specific point that lies in the intersection of U_1 and U_2. But since we are dealing with open sets, we can always wiggle the point around somewhat, and thus consider an open set of points within the intersection $U_1 \cap U_2$. Thus our functions $x^i = x^i(\tilde{x}^j)$ and $\tilde{x}^i = \tilde{x}^i(x^j)$ can be considered for a range of x^i and $\tilde{x}^i$ values. This allows us to ask the question of whether the functions are smooth or not; in other words, are the x^i differentiable functions of the $\tilde{x}^j$, and vice versa? Thus we are led to the notion of a *Differentiable Manifold*, as being a manifold where the coordinates covering any pair of overlapping patches are smooth, differentiable functions of one another. One can, of course, consider different degrees of differentiability; in practice, we shall tend to assume that everything is C^∞ differentiable, meaning that we can differentiate infinitely many times.

The functions that describe how the x^i depend on the $\tilde{x}^j$, or how the $\tilde{x}^i$ depend on the x^j, are called the *transition functions* in the overlap region.

Two atlases are said to be *compatible* if, wherever there are overlaps, the transition functions are smooth.

It is worth emphasising at this point that all this talk about multiple patches is not purely academic. The reason why we have been emphasising this issue is that in general we *need* more than one coordinate patch to cover the whole manifold. To illustrate this point, it is helpful to consider some examples.

1.3.1 The circle; S^1

We have already met the example of the real line itself, as the one-dimensional manifold $\mathbb{R}$. This manifold can be covered by a single coordinate patch, namely we just use the coordinate x, $-\infty < x < \infty$.

There is another example of a one-dimensional manifold that we can consider, namely the circle, denoted by S^1. We can think of the circle as a real line interval, where the right-hand end of the line is identified with the left-hand end. Thus, for the unit circle, we can take a coodinate interval $0 \le \theta < 2\pi$, with the rule that the point $\theta = 2\pi$ is identified with the point $\theta = 0$. However, θ is not a good coordinate everywhere on the circle, because it has a discontinuity where it suddently jumps from 2π to 0. To cover the circle properly, we need to use (at least) two coordinate patches.

To see how this works, it is convenient to think of the standard unit circle $x^2 + y^2 = 1$ centred on the origin in the (x, y) plane, and to consider the standard polar angular coordinate θ running counter-clockwise around the circle. However, we shall only use θ to describe points on the circle corresponding to $0 < \theta < 2\pi$. Call this patch, or set of points, U_1. Introduce also another angular coordinate, called $\tilde{\theta}$, which starts from $\tilde{\theta} = 0$ (more precisely, we shall consider only $\tilde{\theta} > 0$, not allowing $\tilde{\theta} = 0$ itself) over on the left-hand side at $\theta = \pi$, and runs around counter-clockwise until it (almost) returns to its starting point as $\tilde{\theta}$ approaches 2π. We shall use $\tilde{\theta}$ only in the interval $0 < \tilde{\theta} < 2\pi$. This patch of S^1 will be called U_2. Thus we have the patch U_1, which covers all points on S^1 except $(x, y) = (1, 0)$, and the patch U_2, which covers all points on S^1 except $(x, y) = (-1, 0)$. The intersection of U_1 and U_2 therefore comprises all points on S^1 except for the two just mentioned. It therefore comprises two disconnected open intervals, one consisting of points on S^1 that lie above the x axis, and the other comprising points on S^1 that lie below the x axis. We may denote these two intervals by $(U_1 \cap U_2)_+$ and $(U_1 \cap U_2)_-$ respectively. See Figure 3.

The map ϕ_1 from points in U_1 into $\mathbb{R}$ is very simple: we have chosen just to use θ, lying in the open interval $0 < \theta < 2\pi$. For U_2, we have the map ϕ_2 into the open interval $0 < \tilde{\theta} < 2\pi$ in $\mathbb{R}$. The relation between the two coordinates in the overlap region defines

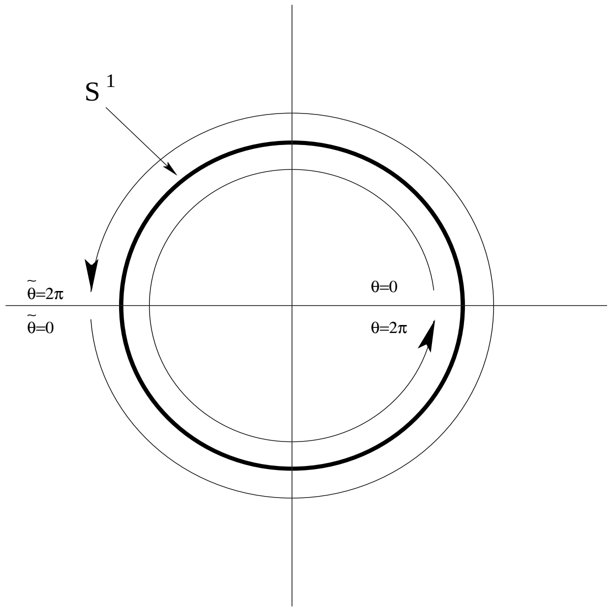

Figure 3: The coordinates θ and $\tilde{\theta}$ cover the two patches of S^1

the transition functions. Since we have an overlap region comprising two disconnected open intervals, we have to define the transition functions in each interval. This can be done easily by inspection, and we have

$$\begin{aligned} (U_1 \cap U_2)_+ : \qquad & \tilde{\theta} = \theta + \pi \\ (U_1 \cap U_2)_- : \qquad & \tilde{\theta} = \theta - \pi \,. \end{aligned} \tag{1.4}$$

It is obvious, in this example, that the transition functions are infinitely differentiable.

1.3.2 The 2-sphere; S^2

For a second example, consider the 2-sphere, denoted by S^2. We can think of this as the surface of the unit ball in Euclidean 3-space. Thus, if we introduce coordinates (x, y, z) on Euclidean 3-space $\mathbb{R}^3$, we define S^2 as the surface

$$x^2 + y^2 + z^2 = 1 \,. \tag{1.5}$$

We can think of using the spherical polar coordinates (θ, ϕ) on S^2, defined in the standard way:

$$x = \sin\theta\, \cos\phi \,, \qquad y = \sin\theta\, \sin\phi \,, \qquad z = \cos\theta \,. \tag{1.6}$$

However, these coordinates break down at the north pole N, and at the south pole S, since at these points $\theta = 0$ and $\theta = \pi$ there is no unique assignment of a value of ϕ. Instead, we can introduce stereographic coordinates, and define two charts:

For a point P on the sphere, take the straight line in $\mathbb{R}^3$ that starts at the north pole N, passes through P, and then intersects the $z = 0$ plane at (x, y) . A simple geometric calculation shows that if the point P has spherical polar coordinates (θ, ϕ), then the corresponding point of intersection in the $z = 0$ plane is at

$$x = \cot \tfrac{1}{2}\theta \, \cos\phi \,, \qquad y = \cot \tfrac{1}{2}\theta \, \sin\phi \,. \tag{1.7}$$

This mapping from points in S^2 into points in $\mathbb{R}^2$ works well except for the point N itself: the north pole gets mapped out to infinity in the (x, y) plane. Let us call U_- the patch of S^2 comprising all points except the north pole N.

We can get a well-defined mapping for a second patch U_+, consisting of all points in S^2 except the south pole S, by making an analogous stereographic mapping from the south pole instead. A simple geometric calculation shows that the straight line in $\mathbb{R}^3$ joining the south pole to the point on S^2 parameterised by (θ, ϕ) intersects the $z = 0$ plane at

$$\tilde{x} = \tan \tfrac{1}{2}\theta \, \cos\phi \,, \qquad \tilde{y} = \tan \tfrac{1}{2}\theta \, \sin\phi \,. \tag{1.8}$$

Thus we have a mapping given by (1.7) from U_- into $\mathbb{R}^2$, with coordinates (x, y), and a mapping given by (1.8) from U_+ into $\mathbb{R}^2$, with coordinates $(\tilde{x}, \tilde{y})$.

In the intersection $U_- \cap U_+$, which comprises all points in S^2 except the north and south poles, we can look at the relation between the corresponding coordinates. From (1.7) and (1.8), a simple calculation shows that we have

$$\tilde{x} = \frac{x}{x^2 + y^2} \,, \qquad \tilde{y} = \frac{y}{x^2 + y^2} \,. \tag{1.9}$$

Conversely, we may express the untilded coordinates in terms of the tilded coordinates, finding

$$x = \frac{\tilde{x}}{\tilde{x}^2 + \tilde{y}^2} \,, \qquad y = \frac{\tilde{y}}{\tilde{x}^2 + \tilde{y}^2} \,. \tag{1.10}$$

It is easy to see that these transition functions defining the relations between the tilded and the untilded coordinates are infinitely differentiable, provided that $x^2 + y^2$ is not equal to zero or infinity. In other words, the transition functions are infinitely differentiable provided we omit the north and south poles; i.e. , they are infinitely differentiable everywhere in the overlap of the two patches.

The construction we have just described has provided us with an atlas comprising two charts. Clearly there was nothing particularly special about the way we chose to do this, except that we made sure that our atlas was big enough to provide a complete covering of S^2. We could, for example, add some more charts by repeating the previous discussion for pairs of charts obtained by stereographic projection from $(x, y, z) = (1, 0, 0)$ and $(-1, 0, 0)$, and from $(0, 1, 0)$ and $(0, -1, 0)$ as well. We would then in total have a collection of six charts in our atlas. A crucial point, though, which was appreciated even in the early days of map-making, is that you cannot cover the whole of S^2 with a single chart.

1.3.3 Incompatible Atlases

It is not necessarily the case that the charts in one atlas are compatible with the charts in another atlas. A simple example illustrating this can be given by considering the one-dimensional manifold $\mathbb{R}$. We have already noted that this can be covered by a single chart. Let us take z to represent the real numbers on the interval $-\infty < z < \infty$. We can choose a chart given by the mapping

$$\phi : \qquad x = z \,. \tag{1.11}$$

We can also choose another chart, defined by the mapping

$$\tilde{\phi} : \qquad \tilde{x} = z^{1/3} \,. \tag{1.12}$$

Over the reals, each mapping gives a 1-1 relation between points z in the original manifold $\mathbb{R}$, and points in the copies of $\mathbb{R}$ coordinatised by x or $\tilde{x}$ respectively. However, these two charts are not compatible everywhere, since we have the relation $\tilde{x} = x^{1/3}$, which is not differentiable at $x = 0$.

1.3.4 Non-Hausdorff manifolds

In practice we shall not be concerned with non-Hausdorff manifolds, but is is useful to give an example of one, since this will illustrate that they are rather bizarre, and hence do not usually arise in situations of physical interest.

Consider the following one-dimensional manifold. We take the real lines $y = 0$ and $y = 1$ in the (x, y) plane $\mathbb{R}^2$. Thus we have the lines $(x, 0)$ and $(x, 1)$. Now, we identify the two lines for all points $x > 0$. Thus we have a one-dimensional manifold consisting of two lines for $x \le 0$, which join together to make one line for $x > 0$. Now, consider the two points $(0, 0)$ and $(0, 1)$. These two points are distinct, since we are only making the identification of $(x, 0)$ and $(x, 1)$ for points where x is strictly positive. However, any open neighbourhood of $(0, 0)$ necessarily intersects any open neighbourhood of $(0, 1)$, since slightly to the right of $x = 0$ for any x, no matter how small, the two lines are identified. Thus, in Atiyah's words, the points $(0, 0)$ and $(0, 1)$ cannot be "housed off" into separate disjoint subsets.

The only one-dimensional Hausdorff manifolds are $\mathbb{R}$ and S^1.

1.3.5 Compact *vs.* non-compact manifolds

When discussing topological spaces, we gave the definition of a compact topological space S as one for which *every* open covering $\{U_i\}$ has a finite sub-collection $\{U_{i_1}, \cdots, U_{i_n}\}$ that also covers S. The key point in this definition is the word "every." To illustrate this, let us consider as examples the two simple one-dimensional manifolds that we have encountered; $\mathbb{R}$ and S^1. As we shall see, $\mathbb{R}$ is non-compact, whilst S^1 is compact.

First, consider $\mathbb{R}$. Of course we can actually just use a single open set to over the whole manifold in this case, since if it is parameterised by the real number z, we just need to take the single open set $-\infty < z < \infty$. Clearly if we took this as our open covering U then there exists a finite sub-collection (namely U itself – no one said the sub-collection has to be a *proper* sub-collection) which also covers $\mathbb{R}$.

However, we can instead choose another open covering as follows. Let U_j be the open set defined by $j - \frac{1}{2} < z < j + \frac{3}{2}$. Thus U_j describes an open interval of length just less than 2. Clearly $\{U_j\}$ for all integers j provides us with an open covering for $\mathbb{R}$, since each adjacent pair U_j and U_{j+1} overlap. However, it is impossible to choose a *finite* subset of the U_j that still provides an open cover of $\mathbb{R}$. By exhibiting an open covering for which a finite sub-collection does *not* provide an open covering of the manifold, we have proved that $\mathbb{R}$ is not compact. A manifold that is not compact is called, not surprisingly, non-compact.

Now, consider instead the manifold S^1. We saw in section (1.3.1) that we can cover S^1 with a minimum of two open sets. We could, of course, use more, for example by covering the circumference of the circle in short little sections of overlapping open sets. However, no matter how short we take the intervals, they must always have a non-zero length, and so after laying a finite number around the circle, we will have covered it all. We could choose an infinity of open sets that covered S^1, for example by choosing intervals of length 1 (in the sense $0 < z < 1$) distributed around the circumference according to the rule that each sucessive interval starts at a point where the angle θ has advanced by $\frac{1}{2}$ relative to the start of the previous interval. Since the circumference of the circle is traversed by advancing θ by 2π, it follows from the fact that π is transcendental that none of these unit intervals will exactly overlap another. However, it will be the case that we can choose a finite subset of these open intervals that is already sufficient to provide an open cover.

No matter what one tries, one will always find that a finite collection of any infinite number of open sets covering S^1 will suffice to provide an open cover. Thus the manifold S^1 is compact.

Of the other examples that we have met so far, all the manifolds $\mathbb{R}^n$ are non-compact, and the manifold S^2 is compact.

1.3.6 Functions on manifolds

A real function f on a manifold M is a mapping

$$f : M \to \mathbb{R} \tag{1.13}$$

that gives a real number for each point p in M. If for some open set U in M we have a coordinate chart ϕ such that U is mapped by ϕ into $\mathbb{R}^n$, then we have a mapping

$$f \circ \phi^{-1} : \mathbb{R}^n \to \mathbb{R}\,. \tag{1.14}$$

If the coordinates in $\mathbb{R}^n$ are called x^i, then the mapping (1.14) can be written simply as $f(x^i)$. In colloquial language we can say that $f(x^i)$ represents the value of f at the point in M specified by the coordinates x^i. In other words, now that it is understood that we may well need different coordinate patches to cover different regions of the manifold, we can usually just think of the chosen coordinates in some patch as being "coordinates on the manifold," and proceed without explicitly reciting the full rigmarole about the mapping ϕ from M into $\mathbb{R}^n$.

The function $f(x^i)$ is called a smooth function if it is a differentiable function of the coordinates x^i in the patch where they are valid.

1.3.7 Orientable manifolds

A manifold is said to be *orientable* if it admits an atlas such that in all overlapping regions between charts, the Jacobian of the relation between the coordinate systems satisfies

$$\det\Big(\frac{\partial x^i}{\partial \tilde{x}^j}\Big) > 0\,. \tag{1.15}$$

1.4 Tangent vectors

We now turn to a discussion of vectors, and tensors, on manifolds.

We should begin this discussion by forgetting certain things about vectors that we learned in kindergarten. There, the concept of a vector was introduced through the notion of the *position vector*, which was an arrow joining a point A to some other point B in three-dimensional Euclidean space. This is fine if one is only going to talk about Euclidean space in Cartesian coordinates, but it is not a valid way describing a vector in general. If the space is curved, such as the sphere, or even if it is flat but described in non-cartesian coordinates, such as Euclidean 3-space described in spherical polar coordinates, the notion of a vector as a line joining two distant points A and B breaks down. What we *can* do is take the infinitesimal limit of this notion, and consider the line joining two points A and $A + \delta A$. In fact what this means is that we think of the *tangent plane* at a point in the space, and imagine vectors in terms of infinitesimal displacements in this plane.

To make the thinking a bit more concrete, consider a 2-sphere, such as the surface of the earth. A line drawn between Ney York and Los Angeles is not a vector; for example, it would not make sense to consider the "sum" of the line from New York to Los Angeles and the line from Los Angeles to Tokyo, and expect it to satisfy any meaningful addition rules. However, we *can* place a small flat sheet on the surface of the earth at any desired point, and draw very short arrows in the plane of the sheet; these are tangent vectors at that particular point on the earth.

The concept of a vector as an infinitesimal displacement makes it sound very like the derivative operator, and indeed this is exactly what a vector is. Suppose we consider some patch U in the manifold M, for which we introduce local coordinates x^i in the usual way. Now consider a path passing through U, which may therefore be described by specifying the values of the coordinates of points along the path. We can do this by introducing a parameter λ that increases monotonically along the path, and so points in M along the path are specified by

$$x^i = x^i(\lambda)\,. \tag{1.16}$$

Consider now a smooth function f defined on M. The values of f at points along the path are therefore given by $f(x^i(\lambda))$. By the chain rule, we shall have

$$\begin{aligned} \frac{df}{d\lambda} &= \sum_{i=1}^{n} \frac{\partial f}{\partial x^i}\, \frac{dx^i(\lambda)}{d\lambda}\,, \\ &= \frac{\partial f}{\partial x^i}\, \frac{dx^i(\lambda)}{d\lambda} \end{aligned} \tag{1.17}$$

Note that here, and throughout this course, we shall be using the Einstein summation convention, as is done in the second line, in which the summation over an index that appears exactly twice is understood.

We may define the directed derivative operator along the path by

$$V \equiv \frac{d}{d\lambda}\,, \tag{1.18}$$

which is a map taking smooth functions to $\mathbb{R}$:

$$f \to V f = \frac{df}{d\lambda} \tag{1.19}$$

This obeys the linearity property

$$V(f+g) = Vf + Vg \tag{1.20}$$

for any pair of smooth functions f and g, and also the Leibnitz property

$$V(fg) = (Vf)g + f(Vg)\,. \tag{1.21}$$

Such a map is called a *tangent vector* at the point p where the evaluation is made.

If we have two different tangent vectors at the point p (corresponding to directional derivatives along two different curves that intersect at p), let us call them $V = d/d\lambda$ and $\widetilde{V} = d/d\widetilde{\lambda}$, then linearity means that we shall have

$$(V + \widetilde{V})f = Vf + \widetilde{V}f\,. \tag{1.22}$$

We can also multiply tangent vectors by constants and they are again tangent vectors. Thus the space of tangent vectors at a point $p \in M$ is a vector space, which is called the *Tangent Space* at p, and denoted by $T_p(M)$. Its dimension is n, the dimension of the manifold M. This can be seen by considering Taylor's theorem in the local coordinate system x^i:

$$f(x) = f(x_p) + h^i\,\frac{\partial f}{\partial x^i} + \cdots\,, \tag{1.23}$$

where $h^i \equiv x^i - x_p^i$ and x_p^i denotes the coordinates corresponding to the point p. Therefore if we define

$$V^i \equiv V x^i = \frac{dx^i}{d\lambda}\,, \tag{1.24}$$

then we shall have

$$Vf = V^i\,\frac{\partial f}{\partial x^i}\,, \tag{1.25}$$

and so we can take $\partial/\partial x^i$ as a basis for the vector space of tangent vectors at the point p. This shows that the dimension of the tangent vector space is equal to the number of coordinates x^i, which is in turn equal to the dimension n of the manifold M. In order to abbreviate the writing, we shall commonly write

$$\partial_i \equiv \frac{\partial}{\partial x^i} \tag{1.26}$$

to denote the tangent vector basis.

To summarise, we can write the tangent vector $V = d/d\lambda$ as

$$V = V^i\,\partial_i\,, \tag{1.27}$$

where V^i are the *components* of the vector V with respect to the basis ∂_i;

$$V^i = \frac{dx^i(\lambda)}{d\lambda}\,. \tag{1.28}$$

(Of course here we are using the Einstein summation convention that any dummy index, which occurs twice in a term, is understood to be summed over the range of the index.)

Notice that there is another significant change in viewpoint here in comparison to the "kindergarten" notion of a vector. We make a clear distinction betwen the vector itself, which is the geometrical object V defined quite independently of any coordinate system by (1.18), and its *components* V^i, which are coordinate-dependent.[2] Indeed, if we imagine now changing to a different set of coordinates x'^i in the space, related to the original ones by $x'^i = x'^i(x^j)$, then we can use the chain rule to convert between the two bases:

$$V = V^j\,\frac{\partial}{\partial x^j} = V^j\,\frac{\partial x'^i}{\partial x^j}\,\frac{\partial}{\partial x'^i} \equiv V'^i\,\frac{\partial}{\partial x'^i}\,. \tag{1.29}$$

In the last step we are, by definition, taking V'^i to be the components of the vector V with respect to the primed coordinate basis. Thus we have the rule

$$V'^i = \frac{\partial x'^i}{\partial x^j}\,V^j\,, \tag{1.30}$$

which tells us how to transform the components of the vector V between the primed and the unprimed coordinate system. This is the fundamental defining rule for how a vector must transform under arbitrary coordinate transformations. Such transformations are called *General Coordinate Transformations.*

Let us return to the point alluded to previously, about the vector as a linear differential operator. We have indeed been writing vectors as derivative operators, so let's see why that is very natural. Suppose we have a smooth function f defined on M. As we discussed before, we can view this, in a particular patch, as being a function $f(x^i)$ of the local coordinates we are using in that patch. It is also convenient to suppress the i index on the coordinates x^i in the argument here, and just write $f(x)$. Now, if we wish to evaluate f at a nearby point $x^i + \xi^i$, where ξ^i is infinitesimal, we can just make a Taylor expansion:

$$f(x+\xi) = f(x) + \xi^i\,\partial_i\,f(x) + \cdots\,, \tag{1.31}$$

[2] However, it sometimes becomes cumbersome to use the longer form of words "the vector whose components are V^i," and so we shall sometimes slip into the way of speaking of "the vector V^i." One should remember, however, that this is a slightly sloppy way of speaking, and the more precise distinction between the vector and its components should always be borne in mind.

and we can neglect the higher terms since ξ is assumed to be infinitesimal. Thus we see that the change in f is given by

$$\delta f(x) \equiv f(x+\xi) - f(x) = \xi^i\, \partial_i\, f(x)\,, \tag{1.32}$$

and that the operator that is implementing the translation of $f(x)$ is exactly what we earlier called a vector field,

$$\xi^i\, \partial_i\,, \tag{1.33}$$

where

$$\delta x^i \equiv (x^i + \xi^i) - x^i = \xi^i\,. \tag{1.34}$$

Having defined $T_p(M)$, the tangent space at the point $p \in M$, we can then define the so-called "tangent bundle" as the space of all possible tangent vectors at all possible points:

$$T(M) = \cup_{p\in M}\, T_p(M)\,. \tag{1.35}$$

This space is a manifold of dimension $2n$, since to specify a point in it one must specify the n coordinates of M and also an n-dimensional set of basis tangent vectors at that point. It is sometimes called the "velocity space," since it is described by a specification of the positions and the "velocities" $\partial/\partial x^i$.

1.4.1 Non-coordinate bases for the tangent space

In the discussion above, we have noted that $\partial_i \equiv \partial/\partial x^i$ forms a basis for the tangent space $T_p(M)$ at a point p in M. This is called a *coordinate basis.* We can choose to use different bases; any choice of n basis vectors that span the vector space is equally valid. Thus we may introduce quantities $E_a{}^i$, where $1 \le a \le n$ (and, as usual, $1 \le i \le n$), and take our n basis vectors to be

$$E_a = E_a{}^i\, \partial_i\,. \tag{1.36}$$

As long as we have $\det(E_a{}^i) \ne 0$, this basis will span the tangent space. Note that $E_a{}^i$ need not be the same at each point in M; we can allow it to depend upon the local coordinates x^i:

$$E_a = E_a{}^i(x)\ \partial_i\,. \tag{1.37}$$

A common terminology is to refer to $E_a{}^i$ as the *inverse vielbein* (we shall meet the vielbein itself a little later). The coordinate index i is commonly also called a *world index*, while the index a is commonly called a *tangent space index.*

In addition to the general coordinate transformations $x^i \to x'^i = x'^i(x)$ that we discussed previously, we can also now make transformations on the tangent space index. In other words, we can make transformations from one choice of non-coordinate basis $E_a{}^i$ to another, say $E'_a{}^i$. This transformation can itself be different at different points in M:

$$E_a \to E'_a = \Lambda_a{}^b(x)\, E_b\,. \tag{1.38}$$

Note that if we have a vector $V = V^i\,\partial_i$, where V^i are its components in the coordinate basis ∂_i, we can also write it as

$$V = V^a\, E_a\,, \tag{1.39}$$

where V^a denotes the tangent-space components of V with respect to the basis E_a. Since V itself is independent of the coice of basis, it follows that the components V^a must transform in the inverse fashion to the transformation (1.38) of E_a, which we write as

$$V^a \to V'^a = \Lambda^a{}_b(x)\, V^b\,, \tag{1.40}$$

where

$$\Lambda^a{}_b\, \Lambda_c{}^b = \delta^a_c\,. \tag{1.41}$$

It is straightforward to see that (1.38) and (1.40), together with (1.41), implies that V given in (1.39) is invariant under these local tangent-space transformations. In matrix notation, we can associate $\Lambda^a{}_b$ with the matrix Λ, whose rows are labelled by a, and columns by b. Then from (1.41) we have that $\Lambda_a{}^b$ corresponds to the inverse, Λ^{-1}. If we view the set of n basis vectors E_a as a row vector denoted by $\mathcal{E}$, and the set of tangent-space components V^a as a column vector denoted by $\mathcal{V}$, then (1.38) and (1.40) can be written as

$$\mathcal{E}' = \mathcal{E}\,\Lambda^{-1}\,, \qquad \mathcal{V}' = \Lambda\,\mathcal{V}\,. \tag{1.42}$$

1.5 Co-vectors

We have so far met the concept of vectors V, which can be expanded in a coordinate basis ∂_i or a general tangent-space basis E_a: $V = V^i\,\partial_i = V^a\,E_a$. For every vector space X, there exists the notion of its *dual space* X^*, which is the space of linear maps

$$X^*: \qquad X \to \mathbb{R}\,. \tag{1.43}$$

What this means is that if V is any vector in X, and ω is any co-vector in X^*, then there exists a rule for making a real number from V and ω. We introduce the notation

$$\langle\omega|V\rangle \in \mathbb{R} \tag{1.44}$$

to denote this rule. The operation is linear, and so we have

$$\begin{aligned} \langle\omega|U+V\rangle &= \langle\omega|U\rangle+\langle\omega|V\rangle\,, \\ \langle\omega|\lambda\,V\rangle &= \lambda\,\langle\omega|V\rangle\,, \end{aligned} \tag{1.45}$$

where U and V are any two vectors, and λ is any real number.

Just as one expands vectors with respect to some basis E_a, namely $V = V^a\,E_a$, so one expands co-vectors with respect to a *dual basis*, which we shall denote by e^a. Thus we write $\omega = \omega_a\,e^a$. By definition, the basis and its dual satisfy

$$\langle e^a|E_b\rangle = \delta^a_b\,. \tag{1.46}$$

From the linearity of the mapping from X to X^*, we therefore have that

$$\begin{aligned} \langle\omega|V\rangle &= \langle\omega_a\,e^a|V^b\,E_b\rangle \\ &= \omega_a\,V^b\,\langle e^a|E_b\rangle = \omega_a\,V^b\,\delta^b_a \\ &= \omega_a\,V^a\,. \end{aligned} \tag{1.47}$$

Note that under the change of basis E^a given in (1.38), it follows that the dual basis e^a must transform inversely, namely

$$e^a \to e'^a = \Lambda^a{}_b\,e^b\,, \tag{1.48}$$

so that the defining property (1.46) is preserved for the primed basis and its dual. Correspondingly, the invariance of ω itself under the change of basis requires that its components ω_a transform as

$$\omega_a \to \omega'_a = \Lambda_a{}^b\,\omega_b\,. \tag{1.49}$$

At every point p in the manifold M we define the *cotangent space* $T^*_p(M)$ as the dual of the tangent space $T_p(M)$. The *cotangent bundle* $T^*(M)$ is then defined as the space of all possible co-vectors at all possible points:

$$T^*(M) = \cup_{p\in M}\,T^*_p(M)\,. \tag{1.50}$$

Like the tangent bundle $T(M)$, the cotangent bundle has dimension $2n$, since the manifold M is n-dimensional and there are n linearly independent co-vectors at each point.

An example of a co-vector is the differential of a function. Suppose $f(x)$ is a function on M. Its differential, df, is called a *differential 1-form*. It is also variously known as the *differential*, the *exterior derivative*, or the *gradient*, of f. It is defined by

$$\langle df|V\rangle = V f \tag{1.51}$$

for any vector V. Recall that Vf is the directional derivative of f along the vector V. If we work in a coordinate basis then the basis for tangent vectors is $\partial_i \equiv \partial/\partial x^i$. Correspondingly, the dual basis for co-vectors is dx^i. By definition, therefore, we have

$$\langle dx^i|\partial_j\rangle = \delta^i_j\,. \tag{1.52}$$

This all makes sense, and fits with our intuitive notion of taking the coordinate differential of f, namely

$$df = \partial_i f\, dx^i\,, \tag{1.53}$$

as can be seen by a simple calculation:

$$\begin{aligned}\langle df|V\rangle &\equiv Vf = V^i\partial_i f\\ &= \langle\partial_i f\, dx^i|V^j\,\partial_j\rangle = \partial_i f\, V^j\langle dx^i|\partial_j\rangle = \partial_i f\, V^j\,\delta^i_j\\ &= \partial_i f\, V^i\,.\end{aligned} \tag{1.54}$$

In a coordinate basis, a general co-vector or 1-form ω is expressed as

$$\omega = \omega_i\, dx^i\,. \tag{1.55}$$

As with a vector, the geometrical object ω itself is independent of any specific choice of coordinates, whilst its components ω_i will change when one changes coordinate frame. We can calculate how this happens by implementing a change of coordinate system, $x^i \to x'^i = x'^i(x^j)$, and applying the chain rule for differentiation:

$$\begin{aligned}\omega &= \omega_i\, dx^i = \omega_i\,\frac{\partial x^i}{\partial x'^j}\,\partial x'^j\\ &\equiv \omega'_j\, dx'^j\,,\end{aligned} \tag{1.56}$$

where in the second line this is simply the definition of what we mean by the components of ω in the primed frame. Thus we read off

$$\omega'_j = \frac{\partial x^i}{\partial x'^j}\,\omega_i\,. \tag{1.57}$$

This may be compared with the transformation rule (1.30) for the components of a vector.

Of course, if we form the scalar quantity $\langle\omega|V\rangle$ then we have

$$\langle\omega|V\rangle = \omega_i\, V^j\langle dx^i|\partial_j\rangle = \omega_i\, V^j\,\delta^i_j = \omega_i\, V^i\,, \tag{1.58}$$

and it is an immediate consequence of (1.30), (1.57) and the chain rule that this is independent of the choice of coordinates, as befits a scalar quantity:

$$\omega'_i\, V'^i = \frac{\partial x^j}{\partial x'^i}\,\frac{\partial x'^i}{\partial x^k}\,\omega_j\, V^k = \frac{\partial x^j}{\partial x^k}\,\omega_j\, V^k = \delta^j_k\,\omega_j\, V^k = \omega_j\, V^j\,. \tag{1.59}$$

1.6 An interlude on vector spaces and tensor products

For the sake of completeness, and by way of introduction to the next section, it is perhaps useful to pause here and define a couple of widely-used and important concepts.

Let us begin with the idea of a *Vector Space*. A vector space X is a set that is closed under finite vector addition and under scalar multiplication. In the general case, the elements are members of a field[3] F, in which case X is called a vector space over F. For now, at least, our interest lies in vector spaces over the real numbers.

The prototype example of a vector space is $\mathbb{R}^n$, with every element represented by an n-tuplet of real numbers $(a_1, a_2, \cdots, a_n)$, where the rule of vector addition is achieved by adding component-wise:

$$(a_1, a_2, \ldots, a_n) + (b_1, b_2, \ldots, b_n) = (a_1 + b_1, a_2 + b_2, \ldots, a_n + b_n)\,, \tag{1.60}$$

and scalar multiplication, for example by the real number r, is component-wise:

$$r\,(a_1, a_2, \ldots, a_n) = (r\,a_1, r\,a_2, \ldots, r\,a_n)\,. \tag{1.61}$$

In general, for any elements A, B and C in the vector space X, and any scalars r and s in the field F, one has the rules:

$$\begin{aligned}
\text{Commutativity:}\quad & A + B = B + A\,,\\
\text{Associativity of vector addition:}\quad & (A + B) + C = A + (B + C)\,,\\
\text{Additive identity:}\quad & 0 + A = A + 0 = A\,,\\
\text{Additive inverse:}\quad & A + (-A) = 0\,,\\
\text{Associativity of scalar multiplication:}\quad & r\,(s\,A) = (r\,s)\,A\,,\\
\text{Distributivity of scalar sums:}\quad & (r + s)\,A = r\,A + s\,A\,,\\
\text{Distributivity of vector sums:}\quad & r\,(A + B) = r\,A + r\,B\,,\\
\text{Identity for scalar multiplication:}\quad & 1\,A = A\,.
\end{aligned} \tag{1.62}$$

Now, let us turn to tensor products. The *Tensor Product* of two vector spaces X and Y, denoted by $X \otimes Y$, is again a vector space. It obeys a distributive law, in the sense that if X, Y and Z are vector spaces, then

$$X \otimes (Y + Z) = (X \otimes Y) + (X \otimes Z)\,. \tag{1.63}$$

[3]A Field is any set of elements that satisfies axioms of addition and multiplication, and is a commutative division algebra. Examples of fields are the real numbers $\mathbb{R}$, the complex numbers $\mathbb{C}$, and the rational numbers. By contrast, the integers are not a field, since division of integers by integers does not give the integers.

If elements of the vector spaces X and Y are denoted by x and y respectively, then the tensor-product vector space $X\otimes Y$ is spanned by elements of the form $x\otimes y$. The following rules are satisfied:

$$
\begin{aligned}
(x_1+x_2)\otimes y &= x_1\otimes y+x_2\otimes y\,,\\
x\otimes(y_1+y_2) &= x\otimes y_1+x\otimes y_2\,,\\
\lambda\,(x\otimes y) &= (\lambda\,x)\otimes y=x\otimes(\lambda\,y)\,,
\end{aligned}
\tag{1.64}
$$

where λ is any scalar. Note that $0\otimes y=x\otimes 0=0$.

If α_i is a basis of vectors for X, and β_j is a basis of vectors for Y, then $\alpha_i\otimes\beta_j$ for all (i,j) gives a basis for $X\otimes Y$. In other words, we can expand any vectors x and y in the vector spaces X and Y in the forms

$$x=\sum_i x_i\,\alpha_i\,,\qquad y=\sum_j y_j\,\beta_j\,,\tag{1.65}$$

and we can expand any vector z in the tensor-product vector space $Z=X\otimes Y$ as

$$z=\sum_{i,j} z_{ij}\,\alpha_i\otimes\beta_j\,.\tag{1.66}$$

Note that if the dimensions of the vector spaces X and Y are p and q, i.e. one needs a set of p basis vectors for X, and a set of q basis vectors for Y, then the tensor product $X\otimes Y$ has dimension pq. For example, if we take the tensor product $\mathbb{R}^p\otimes\mathbb{R}^q$, we get a tensor product vector space of dimension pq that is actually just $\mathbb{R}^{pq}$.

1.7 Tensors

Having introduced the notion of vectors and co-vectors, it is now straightforward to make the generalisation to tensors of arbitrary rank. By this is meant geometrical objects which live in a tensor product space, involving, say, p factors of the tangent space $T_p(M)$, and q factors of the cotangent space $T_p^*(M)$. Such a tensor is said to be of type (p,q), and to have rank $(p+q)$. Suppose T is such a tensor. We can then express it in terms of its components in a coordinate basis as

$$T=T^{i_1\cdots i_p}{}_{j_1\cdots j_q}\,\partial_{i_1}\otimes\partial_{i_2}\otimes\cdots\otimes\partial_{i_p}\otimes dx^{j_1}\otimes dx^{j_2}\otimes\cdots\otimes dx^{j_q}\,.\tag{1.67}$$

With the standard philosophy that the tensor T itself is a geometrical object which exists independently of any choice of frame, we therefore see by comparing with its expansion in a primed coordinate frame,

$$T=T'^{i_1\cdots i_p}{}_{j_1\cdots j_q}\,\partial'_{i_1}\otimes\partial'_{i_2}\otimes\cdots\otimes\partial'_{i_p}\otimes dx'^{j_1}\otimes dx'^{j_2}\otimes\cdots\otimes dx'^{j_q}\,,\tag{1.68}$$

where of course $\partial'_i \equiv \partial/\partial x'^i$, that the components will transform according to the rule

$$T'^{i_1\cdots i_p}{}_{j_1\cdots j_q} = \frac{\partial x'^{i_1}}{\partial x^{k_1}} \cdots \frac{\partial x'^{i_p}}{\partial x^{k_q}} \frac{\partial x^{\ell_1}}{\partial x'^{j_1}} \cdots \frac{\partial x^{\ell_q}}{\partial x'^{j_q}} T^{k_1\cdots k_p}{}_{\ell_1\cdots\ell_q} \,. \tag{1.69}$$

In other words, there is a factor of the type $\frac{\partial x'^i}{\partial x^k}$ for each vector index, just like the transformation for V^i in (1.30), and a factor of the type $\frac{\partial x^\ell}{\partial x'^j}$ for each co-vector index, just like in the transformation of ω_i in (1.57).

One can view (1.69) as the *defining* property of a tensor, or, more precisely, the defining property of a general-coordinate tensor, i.e. a tensor with respect to general coordinate transformations. Namely, we can say that T is a type (p,q) tensor under general-coordinate transformations if and only if its components $T^{i_1\cdots i_p}{}_{j_1\cdots j_q}$ transform like (1.69) under general coordinate transformations.

It is obvious that if T and U are two tensors of type (p,q), then $T+U$ is also a tensor of type (p,q). One proves this by the standard technique of showing that the components of $T+U$ transform in the proper way under general coordinate transformations.

It is rather obvious that we can take arbitrary products of tensors and thereby obtain new tensors. For example, if V is a $(1,0)$ tensor (i.e. a vector), and if ω is a $(0,1)$ tensor (i.e. a co-vector, or 1-form), then $W \equiv V \otimes \omega$ is a tensor of type $(1,1)$, with components

$$W^i{}_j = V^i\, \omega_j \,. \tag{1.70}$$

It is clear from the transformation rules (1.30) and (1.57) for V^i and ω_j that the components $W^i{}_j$ transform in the proper way, namely as in (1.69) with $p=q=1$. This product is called the *Outer Product* of V and ω. This terminology signifies that no index contractions are being made, and so the rank of the product tensor is equal to the sum of the ranks of the two tensor factors. In general, we can take the outer product of two tensor of types (p,q) and $p',q')$, thereby obtaining a tensor of type $(p+p', q+q')$.

Note that the Kronecker delta symbol δ^i_j is nothing but the set of components of a very specific tensor δ of type $(1,1)$. It is known as an *invariant tensor*, since it takes the identical form in any coordinate frame. Thus if we take the standard definition of the Kronecker delta in a particular coordinate frame, namely

$$\delta^i_j = 1\,, \quad \text{if } i=j\,, \qquad \delta^i_j = 0\,, \quad \text{if } i \ne j\,, \tag{1.71}$$

and then compute the components of δ in another coordinate frame, under the assumption that it is a tensor, then from (1.69) we obtain

$$\delta'^i{}_j = \delta^k_\ell \,\frac{\partial x'^i}{\partial x^k}\,\frac{\partial x^\ell}{\partial x'^j} = \frac{\partial x'^i}{\partial x'^j} = \delta^i_j \,, \tag{1.72}$$

and so it has the same numerical set of components in all coordinate frames.

Another operation that takes tensors into tensors is called *Contraction.* We can illustrate this with a tensor of type $(2,2)$; the generalisation to the arbitrary case is immediate. Suppose T is of type $(2,2)$, with components $T^{ij}{}_{k\ell}$. We can form a tensor of type $(1,1)$ by contracting, for example, the first upper index and the first lower index:

$$X^j{}_\ell \equiv T^{ij}{}_{i\ell}\,. \tag{1.73}$$

(Recall that as always, the summation convnetion is operating here, and so the repeated i index is understood to be summed over $1 \le i \le n$.) The proof that $X^j{}_\ell$ so defined is indeed a tensor is to verify that it transforms properly under general coordinate transformations:

$$\begin{aligned} X'^j{}_\ell &\equiv T'^{ij}{}_{i\ell} = T^{mn}{}_{pq}\,\frac{\partial x'^i}{\partial x^m}\,\frac{\partial x'^j}{\partial x^n}\,\frac{\partial x^p}{\partial x'^i}\,\frac{\partial x^q}{\partial x'^\ell} \\ &= T^{mn}{}_{pq}\,\delta^p_m\,\frac{\partial x'^j}{\partial x^n}\,\frac{\partial x^q}{\partial x'^\ell} = T^{mn}{}_{mq}\,\frac{\partial x'^j}{\partial x^n}\,\frac{\partial x^q}{\partial x'^\ell} \\ &= X^n{}_q\,\frac{\partial x'^j}{\partial x^n}\,\frac{\partial x^q}{\partial x'^\ell}\,. \end{aligned} \tag{1.74}$$

Note that the crucial point is that the transformation matrices for the upper and lower i indices are inverses of one another, and so in the second line we just obtain the Kronecker delta δ^p_m that implements the contraction of indices on the unprimed tensor $T^{mn}{}_{pq}$, giving back $X^n{}_q$. It is clear that the same thing will happen for a contraction of an upper and a lower index in any tensor.

A common example of an index contraction, and one which we have in fact already encountered, is in the formation of the so-called *Inner Product.* If V is a vector and ω is a co-vector or 1-form, then their inner product is given by

$$\langle\omega|V\rangle = \omega_i\,V^i\,, \tag{1.75}$$

as in (1.58). This can be viewed as taking the index contraction on their outer product $W^i{}_j \equiv V^i\,\omega_j$ defined as in (1.70): $W^i{}_i = V^i\omega_i$. Not surprisingly, since this produces a tensor of type $(0,0)$ (otherwise known as a scalar), it is invariant under general coordinate transformations, as we saw earlier.

Note that one can also perform the operations of symmetrisation or antisymmetrisation of a tensor, and this yields another tensor for which these properties are preserved under general coordinate transformations. For example, if T_{ij} is a general 2-index tensor we can define its symmetric and antisymmetric parts:

$$S_{ij} = \tfrac{1}{2}(T_{ij} + T_{ji})\,, \qquad A_{ij} = \tfrac{1}{2}(T_{ij} - T_{ji})\,, \tag{1.76}$$

and that $T_{ij} = S_{ij} + A_{ij}$. It is easy to see that S_{ij} and A_{ij} are both tensors, and that S_{ij} is symmetric in all coordinate frames, and A_{ij} is antisymmetric in all coordinate frames. It is useful to have a notation indicating a symmetrisation or antisymmetrisation over sets of indices. This is done by the use of round or square brackets, respectively. Thus we can rewrite (1.76) as

$$S_{ij} = T_{(ij)} \equiv \tfrac{1}{2}(T_{ij} + T_{ji})\,, \qquad A_{ij} = T_{[ij]} \equiv \tfrac{1}{2}(T_{ij} - T_{ji})\,. \tag{1.77}$$

More generally, symmetrisation and antisymmetrisation over n indices is defined by

$$\begin{aligned} T_{(i_1\cdots i_n)} &\equiv \frac{1}{n!}\Big(T_{i_1\cdots i_n} + \text{even permutations} + \text{odd permutations}\Big)\,, \\ T_{[i_1\cdots i_n]} &\equiv \frac{1}{n!}\Big(T_{i_1\cdots i_n} + \text{even permutations} - \text{odd permutations}\Big)\,. \end{aligned} \tag{1.78}$$

We shall see later that totally antisymmetric tensors of type $(0,p)$ play an especially important role in geometry. They are the p-index generalisation of the co-vector or 1-form, and are known as p-forms.

1.8 The Metric Tensor

At this point, we introduce an additional structure on the manifold M, namely the notion of a *metric*. As its name implies, this is a way of measuring distances in M. It should be emphasised from the outset that there is no unique way of doing this, although very often it may be the case that there is a natural preferred choice of metric (up to scaling), suggested by the symmetries of the problem.

Mathematically, we may simply define the metric as a smooth assignment to the tangent space at each point of the manifold a real inner product, or bilinear form, which is linear over functions. We shall also require that this bilinear form be *symmetric*. Thus if U and V are any vectors, then a metric g is a bilinear map from U and V into the reals

$$g(U,V) \in \mathbb{R}\,, \tag{1.79}$$

with the following properties

$$g(U,V) = g(V,U)\,, \qquad g(\lambda\,U, \mu\,V) = \lambda\,\mu\,g(U,V)\,, \tag{1.80}$$

where λ and μ are arbitrary real numbers. We shall also demand that the metric g be non-degenerate, which means that if

$$g(U,V) = 0 \tag{1.81}$$

for all V, then it must be that $U = 0$.

Stated in more prosaic terms, the definitions above amount to saying that we have a real type $(0, 2)$ symmetric tensor, with components g_{ij}, with the non-degeneracy condition that $\det(g_{ij}) \neq 0$. In terms of components, we have

$$g(U, V) = g_{ij}\, U^i\, V^j \,. \tag{1.82}$$

Since g_{ij} is symmetric, it will have real eigenvalues; in general it will have s positive eigenvalues and t negative eigenvalues. Since we are requiring that $\det(g_{ij}) \neq 0$, it follows that s and t will be the same for all points in the coordinate patch, since for an eignevalue to change sign it would have to pass through zero at some point, which would then give a vanishing determinant. The *signature* of the metric is defined to be $s - t$. The two cases that commonly arise are when $t = 0$ and so $s = n = \dim M$, and $s = n - 1$, $t = 1$. In the former case the associated geometry is called *Riemannian Geometry.* In the latter, (or indeed in any case where s and t are both non-vanishing), the associated geometry is called *Pseudo-Riemannian.* The situation where $t = 1$ arises in physics in special and general relativity, with the negative eigenvalue being associated with the time direction.

The physical interpretation of the metric is that it gives the separation between two infinitesimally-separated points in the manifold. Supposing that these points correspond to the local coordinate values x^i and $x^i + dx^i$, the separation ds between them is given by

$$ds^2 = g_{ij}\, dx^i\, dx^j \,. \tag{1.83}$$

Note that in the case of Riemannian geometry, $ds^2 \geq 0$, with $ds^2 = 0$ if and only if $dx^i = 0$. In pseudo-Riemannian geometry, on the other hand, ds^2 can have either sign, depending on whether the positive contribution from the spatial directions outweighs, or is outweighed by, the negative contribution from the time direction or directions. The separation of the neighbouring points is then said to be spacelike, timelike or null, depending on whether ds^2 is positive, negative or zero.

Probably the most familiar example of a metric is the rule for measuring distances in Euclidean space. If we have two infinitesimally-separated points in $\mathbb{R}^3$ with coordinates x^i and $x^i + dx^i$, then, as we know from the work of Pythagoras, the square of the distance ds between the points can be written as

$$ds^2 = \delta_{ij}\, dx^i\, dx^j \,. \tag{1.84}$$

In this case the metric tensor g has components $g_{ij} = \delta_{ij}$. Of course this instantly generalises to an arbitrary dimension.

If we use speherical polar coordinates (θ, ϕ) on the 2-sphere, then the standard metric, namely the one induced on the unit S^2 via its embedding in $\mathbb{R}^3$ that we discussed earlier, is

$$ds^2 = d\theta^2 + \sin^2\theta\, d\phi^2\,, \tag{1.85}$$

as is easily established by elementary geometry. It can also be dervived by direct substitution into (1.84) of the expressions

$$x^1 = \sin\theta\, \cos\phi\,, \qquad x^2 = \sin\theta\, \sin\phi\,, \qquad x^3 = \cos\theta \tag{1.86}$$

giving the Cartesian coordinates in $\mathbb{R}^3$ of points on the unit sphere.

Viewing g_{ij} as a symmetric $n \times n$ matrix, whose determinant is assumed to be non-zero, we can take its inverse, obtaining another symmetric tensor whose components we shall denote by g^{ij}. The statement that this corresponds to the inverse of the matrix with components g_{ij} is therefore that

$$g_{ij}\, g^{jk} = \delta_i{}^k\,, \tag{1.87}$$

which is just the component form of the matrix equation $\mathbf{g}\,\mathbf{g}^{-1} = \mathbf{1}$. It is easy to verify that g^{jk} is indeed a tensor, by verifying that with g_{ij} and δ^i_k transforming in their known tensorial ways, equation (1.87) transforms tensorially provided that g^{jk} transforms in the standard way for a tensor of type $(2, 0)$.

It is now obvious that if U and V are two vectors, then the quantity $g_{ij}\, U^i\, V^j$ transforms as a scalar, i.e. it is invariant under general coordinate transformations. This quantity is known as the inner product of the two vectors U and V.

Note that another way of viewing this is that we can think of g_{ij} as "lowering the index" on U^i or on V^i, so that we are then contracting the upper and the lower index on the components of a vector and a 1-form or co-vector, respectively. This then makes contact with the notion of the inner product of a vector and a 1-form, which we defined in section 1.7. Because g_{ij} is invertible, we do not "lose information" by lowering the index; we can always raise it back up again with the inverse metric g^{ij}, getting back to where we started, by virtue of equation (1.87). Because of this fact, it is conventional to use the same symbol for the quantity with the index lowered using g_{ij}, or raised using g^{ij}. Thus for example, we define

$$V_i \equiv g_{ij}\, V^j\,, \qquad W^i \equiv g^{ij}\, W_j\,. \tag{1.88}$$

It is obvious from the properties of tensors discussed in section (1.7) that if V is a vector with components V^i, then the downstairs components $V_i \equiv g_{ij}\, V^j$ transform as the components of a co-vector. More generally, if any indices on the components of any tensor are lowered or raised using the metric tensor or its inverse, one gets the compoents of a tensor again.

Note that if we are in the Riemannian case, where the eigenvalues of g_{ij} are all positive, then we must have that

$$g_{ij}\, V^i\, V^j \ge 0\,, \tag{1.89}$$

with equality achieved if and only if $V^i = 0$. By contrast, in the pseudo-Riemannian case where there is one or more time directions, the inner product $g_{ij}\, V^i\, V^j$ can in general have either sign, and there can exist so-called *null vectors* for which $g_{ij}\, V^i\, V^j = 0$, with $V^i \ne 0$.

1.9 Covariant differentiation

A familiar concept in Cartesian tensor analysis is that if one acts on the components of any tensor field with the partial derivatives $\partial_i \equiv \partial/\partial x^i$, one gets the components of another tensor field, with an additional index.[4] However, this property as it stands is very specific to the case of Cartesian tensors. The crucial point is that in Cartesian tensor analysis one does not allow general coordinate transformations between coordinate frames, but rather, one restricts to a very special subset, namely transformations with *constant coefficients*, namely

$$x^i \to x'^i = M^i{}_j\, x^j\,, \tag{1.90}$$

where $M^i{}_j$ are constants.

In order to retain the useful property of having a derivative operator that maps tensor fields into tensor fields in the case of arbitrary coordinate transformations, it will be necessary to introduce a new type of derivative, called the *Covariant Derivative.* To introduce this, let us begin by seeing what goes wrong if we just try to act with the partial derivative.

Suppose V^i is a vector under general coordinate transformations (so it transforms as in (1.30)). Let us consider the quantity

$$W^i{}_j \equiv \frac{\partial V^i}{\partial x^j}\,. \tag{1.91}$$

Is this a tensor? To test it, we calculate $W'^i{}_j$, to see if it is the proper tensorial transform of $W^i{}_j$. We get:

$$\begin{aligned} W'^i{}_j \equiv \frac{\partial V'^i}{\partial x'^j} &= \frac{\partial x^\ell}{\partial x'^j}\, \frac{\partial}{\partial x^\ell}\Big(\frac{\partial x'^i}{\partial x^k}\, V^k\Big) \\ &= \frac{\partial x^\ell}{\partial x'^j}\, \frac{\partial x'^i}{\partial x^k}\, \frac{\partial V^k}{\partial x^\ell} + \frac{\partial x^\ell}{\partial x'^j}\, \frac{\partial^2 x'^i}{\partial x^\ell\, \partial x^k}\, V^k\,, \end{aligned}$$

[4] We now use "tensor" as a generic term, which can include the particular cases of a scalar, and a vector.

$$= \frac{\partial x^\ell}{\partial x'^j}\,\frac{\partial x'^i}{\partial x^k}\,W^k{}_\ell + \frac{\partial x^\ell}{\partial x'^j}\,\frac{\partial^2 x'^i}{\partial x^\ell\,\partial x^k}\,V^k\,. \tag{1.92}$$

So the answer is no; the first term by itself would have been fine, but the second term here has spoiled the general coordinate transformation behaviour. Of course there is no mystery behind what we are seeing here; the second term has arisen because the derivative operator has not only landed on the vector V^k, giving us what we want, but it has also landed on the transformation matrix $\partial x'^i/\partial x^k$. This problem was avoided in the case of the Cartesian tensors, because we only required that they transform nicely under *constant* transformations (1.90).

Now, we shall define the covariant derivative ∇_j of a vector V^i as follows:

$$\nabla_j\,V^i \equiv \partial_j\,V^i + \Gamma^i{}_{jk}\,V^k\,. \tag{1.93}$$

It is defined to have precisely the correct transformation properties under general coordinate transformations to ensure that the quantity

$$T^i{}_j \equiv \nabla_j\,V^i \tag{1.94}$$

does transform like a tensor. The crucial point here is that $\Gamma^i{}_{jk}$ itself is *not* a tensor. It is called a *Connection*, in fact.

We may also impose on the quantities $\Gamma^i{}_{jk}$ the symmetry condition

$$\Gamma^i{}_{jk} = \Gamma^i{}_{kj}\,, \tag{1.95}$$

and usually this is done. It will be assumed that (1.95) holds in all our subsequent discussions, unless otherwise specified.

First, let us see how we would *like* $\Gamma^i{}_{jk}$ to transform, and then, we shall show how to construct such an object. By definition, we want it to be such that

$$\frac{\partial x'^i}{\partial x^k}\,\frac{\partial x^\ell}{\partial x'^j}\,\nabla_\ell\,V^k = \nabla'_j\,V'^i \equiv \partial'_j\,V'^i + \Gamma'^i{}_{jk}\,V'^k\,. \tag{1.96}$$

Wrtiting out the two sides here, we get the requirement that

$$\begin{aligned}\frac{\partial x'^i}{\partial x^k}\,\frac{\partial x^\ell}{\partial x'^j}\,\Big(\partial_\ell\,V^k + \Gamma^k{}_{\ell m}\,V^m\Big) &= \frac{\partial x^\ell}{\partial x'^j}\,\partial_\ell\Big(\frac{\partial x'^i}{\partial x^m}\,V^m\Big) + \Gamma'^i{}_{jk}\,\frac{\partial x'^k}{\partial x^m}\,V^m\\ &= \frac{\partial x^\ell}{\partial x'^j}\,\frac{\partial x'^i}{\partial x^m}\,\partial_\ell\,V^m + \frac{\partial x^\ell}{\partial x'^j}\,\frac{\partial^2 x'^i}{\partial x^\ell\,\partial x^m}\,V^m + \Gamma'^i{}_{jk}\,\frac{\partial x'^k}{\partial x^m}\,V^m\,.\end{aligned} \tag{1.97}$$

The required equality of the left-hand side of the top line and the right-hand side of the bottom line *for all vectors* V^m allows us to deduce that we must have

$$\frac{\partial x'^i}{\partial x^m}\,\frac{\partial x^\ell}{\partial x'^j}\,\Gamma^k{}_{\ell m} = \frac{\partial x'^k}{\partial x^m}\,\Gamma'^i{}_{jk} + \frac{\partial x^\ell}{\partial x'^j}\,\frac{\partial^2 x'^i}{\partial x^\ell\,\partial x^m}\,. \tag{1.98}$$

Multiplying this by $\partial x^m/\partial x'^n$ then gives us the result that

$$\Gamma'^i{}_{jn} = \frac{\partial x'^i}{\partial x^k}\,\frac{\partial x^\ell}{\partial x'^j}\,\frac{\partial x^m}{\partial x'^n}\,\Gamma^k{}_{\ell m} - \frac{\partial x^m}{\partial x'^n}\,\frac{\partial x^\ell}{\partial x'^j}\,\frac{\partial^2 x'^i}{\partial x^\ell\,\partial x^m}\,. \tag{1.99}$$

This dog's breakfast is the required transformation rule for $\Gamma^i{}_{jk}$. Notice that the first term on the right-hand side is the "ordinary" type of tensor transformation rule. The presence of the second term shows that $\Gamma^i{}_{jk}$ is not in fact a tensor, because it doesn't transform like one.

The above calculation is quite messy, but hopefully the essential point comes across clearly; the purpose of the ugly second term in the transformation rule for $\Gamma^i{}_{jk}$ is precisely to remove the ugly extra term that we encountered which prevented $\partial_j V^i$ from being a tensor.

Luckily, it is quite easy to provide an explicit construction for a suitable quantity $\Gamma^i{}_{jk}$ that has the right transformation properties. First, we need to note that we should like to define a covariant derivative for any tensor, and that it should satisfy Leibnitz's rule for the differentiation of products. Now the need for the covariant derivative arises because the transformation of the components of a vector or a tensor from one coordinate frame to another involves non-constant transformation matrices of the form $\partial x'^i/\partial x^j$. Therefore on a scalar, which doesn't have any indices, the covariant derivative must be just the same thing as the usual partial derivative. Combining this fact with the Leibnitz rule, we can work out what the covariant derivative of a vector with a downstairs index must be:

$$\begin{aligned}
\partial_j\,(V^i\,U_i) &= (\partial_j\,V^i)\,U_i + V^i\,\partial_j\,U_i\,, && \text{usual Leibnitz rule}\,,\\
&= \nabla_j\,(V^i\,U_i) = (\nabla_j\,V^i)\,U_i + V^i\,\nabla_j\,U_i\,, && \text{covariant Leibnitz rule}\,,\\
&= (\partial_j\,V^i + \Gamma^i{}_{jk}\,V^k)\,U_i + V^i\,\nabla_j\,U_i\,, && \text{from definition of } \nabla_j\,V^i\,.
\end{aligned} \tag{1.100}$$

Comparing the top line with the bottom line, the two $\partial_j V^i$ terms cancel, leaving

$$V^i\,\partial_j\,U_i = V^i\,\nabla_j\,U_i + \Gamma^i{}_{jk}\,V^k\,U_i\,. \tag{1.101}$$

Changing the labelling of dummy indices to

$$V^i\,\partial_j\,U_i = V^i\,\nabla_j\,U_i + \Gamma^k{}_{ji}\,V^i\,U_k\,, \tag{1.102}$$

we see that if this is to be true for all possible vectors V^i then we must have

$$\nabla_j\,U_i = \partial_j\,U_i - \Gamma^k{}_{ji}\,U_k\,. \tag{1.103}$$

This gives us what we wanted to know, namely how the covariant derivative acts on vectors with downstairs indices.

It is straightforward to show, with similar techniques to the one we just used, that the covariant derivative of an arbitrary tensor with p upstairs indices and q downstairs indices is given by using the two rules (1.93) and (1.103) for each index; (1.93) for each upstairs index, and (1.103) for each downstairs index. Thus we have

$$\begin{aligned}\nabla_i T^{j_1\cdots j_p}{}_{k_1\cdots k_q} &= \partial_i T^{j_1\cdots j_p}{}_{k_1\cdots k_q} + \Gamma^{j_1}{}_{i\ell}\, T^{\ell j_2\cdots j_p}{}_{k_1\cdots k_q} + \Gamma^{j_2}{}_{i\ell}\, T^{j_1\ell j_3\cdots j_p}{}_{k_1\cdots k_q} + \cdots \\ &\quad + \Gamma^{j_p}{}_{i\ell}\, T^{j_1\cdots j_{p-1}\ell}{}_{k_1\cdots k_q} - \Gamma^{\ell}{}_{ik_1}\, T^{j_1\cdots j_p}{}_{\ell k_2\cdots k_q} - \Gamma^{\ell}{}_{ik_2}\, T^{j_1\cdots j_p}{}_{k_1\ell k_3\cdots k_q} - \cdots \\ &\quad - \Gamma^{\ell}{}_{ik_q}\, T^{j_1\cdots j_p}{}_{k_1\cdots k_{q-1}\ell}\,. \end{aligned} \tag{1.104}$$

Note that a trivial case is when we apply the covariant derivative to a scalar. Since this has no indices of either type, it follows that the covariant derivative is exactly the same as the standard partial derivative:

$$\nabla_i f = \partial_i f\,, \tag{1.105}$$

for any scalar function f. Commonly, we may write $\nabla_i f$ rather than the more fundamental but identical expression $\partial_i f$, simply for the sake of uniformity of appearance in equations.

Now, recall that in section 1.8 we introduced the notion of the metric tensor g_{ij}. Calculating its covariant derivative using (1.103) for each downstairs index, we find

$$\nabla_k\, g_{ij} = \partial_k\, g_{ij} - \Gamma^{\ell}{}_{ki}\, g_{\ell j} - \Gamma^{\ell}{}_{kj}\, g_{i\ell}\,. \tag{1.106}$$

We can now give an explicit construction of the connection $\Gamma^i{}_{jk}$. We do this by making the additional requirement that we should like the metric tensor to be *covariantly constant*, $\nabla_k\, g_{ij} = 0$. This is a very useful property to have, since it means, for example, that if we look at the scalar product $V^i\, W^j\, g_{ij}$ of two vectors, we shall have

$$\nabla_k (V^i\, W^j\, g_{ij}) = (\nabla_k\, V^i)\, W^j\, g_{ij} + V^i\, (\nabla_k\, W^j)\, g_{ij}\,. \tag{1.107}$$

Remembering our rule that we shall in fact freely write $W^j\, g_{ij}$ as W_i, and so on, it should be clear that life would become a nightmare if the metric could not be taken freely through the covariant derivative!

Luckily, it turns out that all the things we have been asking for are possible. We can find a connection $\Gamma^i{}_{jk}$ that is symmetric in jk, gives us a covariant derivative that satisfies the Leibnitz rule, and for which $\nabla_k\, g_{ij} = 0$. We can find it just by juggling around the indices in equation (1.106). To do this, we write out $\nabla_k\, g_{ij} = 0$ using (1.106) three times, with different labellings of the indices:

$$\partial_k\, g_{ij} - \Gamma^{\ell}{}_{ki}\, g_{\ell j} - \Gamma^{\ell}{}_{kj}\, g_{i\ell} = 0\,,$$

$$\begin{aligned} \partial_i\, g_{kj} - \Gamma^\ell{}_{ik}\, g_{\ell j} - \Gamma^\ell{}_{ij}\, g_{k\ell} &= 0\,, \\ \partial_j\, g_{ik} - \Gamma^\ell{}_{ji}\, g_{\ell k} - \Gamma^\ell{}_{jk}\, g_{i\ell} &= 0\,, \end{aligned} \tag{1.108}$$

Now, add the last two equations and subtract the first one from this. Since we are requiring $\Gamma^i{}_{jk}$ to be symmetric in jk, we therefore get

$$\partial_i\, g_{kj} + \partial_j\, g_{ik} - \partial_k\, g_{ij} - 2\Gamma^\ell{}_{ij}\, g_{k\ell} = 0\,. \tag{1.109}$$

Multiplying this by the inverse metric g^{km}, we immediately obtain the following expression for $\Gamma^i{}_{jk}$ (after finally relabelling indices for convenience):

$$\Gamma^i{}_{jk} = \tfrac{1}{2} g^{i\ell}\, (\partial_j\, g_{\ell k} + \partial_k\, g_{j\ell} - \partial_\ell\, g_{jk})\,. \tag{1.110}$$

This is known as the *Christoffel Connection*, or sometimes the *Affine Connection.*

It is a rather simple matter to check that $\Gamma^i{}_{jk}$ defined by (1.110) does indeed have the required transformation property (1.99) under general coordinate transformations. Actually, there is really no need to check this point, since it is logically guaranteed from the way we constructed it that it must have this property. So we leave it as an "exercise to the reader," to verify by direct computation. The principle should be clear enough; one simply uses the expression for $\Gamma^i{}_{jk}$ given in (1.110) to calculate $\Gamma'^i{}_{jk}$, in terms of ∂'_i and g'_{ij} (which can be expressed in terms of ∂_i and g_{ij} using their standard tensorial transformation properties). It then turns out that $\Gamma'^i{}_{jk}$ is related to $\Gamma^i{}_{jk}$ by (1.99).

Notice that $\Gamma^i{}_{jk}$ is zero if the metric components g_{ij} are all constants. This explains why we never see the need for $\Gamma^i{}_{jk}$ if we only look at Cartesian tensors, for which the metric is just δ_{ij}. But as soon as we consider any more general situation, where the components of the metric tensor are functions of the coordinates, the Christoffel connection will become non-vanishing. Note that this does not necessarily mean that the metric has to be one on a curved space (such as the 2-sphere that we met earlier); even a flat metric written in "curvilinear coordinates" will have a non-vanishing Christoffel connection. As a simple example, suppose we take the metric on the plane,

$$ds^2 = dx^2 + dy^2\,, \tag{1.111}$$

and write it in polar coordinates (r, θ) defined by

$$x = r\, \cos\theta\,, \qquad y = r\, \sin\theta\,. \tag{1.112}$$

It is easy to see that (1.111) becomes

$$ds^2 = dr^2 + r^2\, d\theta^2\,. \tag{1.113}$$

If we label the (r, θ) coordinates as (x^1, x^2) then in the metric $ds^2 = g_{ij}\, dx^i\, dx^j$ we shall have

$$g_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix}, \qquad g^{ij} = \begin{pmatrix} 1 & 0 \\ 0 & r^{-2} \end{pmatrix}. \tag{1.114}$$

Using (1.110), simple algebra leads to the following results:

$$\begin{aligned} &\Gamma^1{}_{11} = 0\,, \qquad \Gamma^1{}_{12} = 0\,, \qquad \Gamma^1{}_{22} = -r\,, \\ &\Gamma^2{}_{11} = 0\,, \qquad \Gamma^2{}_{12} = \frac{1}{r}\,, \qquad \Gamma^2{}_{22} = 0\,. \end{aligned} \tag{1.115}$$

The covariant derivative allows us to obtain a new general-coordinate tensor by applying it to any tensor field. Since the metric is, by construction, covariantly constant, we can freely take it through covariant derivatives as and when we wish. Since the covariant derivative is tensorial, it follows that any contraction of indices, performed with the metric tensor or its inverse if the two contracted indices are both vector-like or both co-vector-like, will also be a tensor.

For example, if V is a vector, with components V^i, then the quantity $\nabla_i V^i$ is a scalar.

Recalling that $\nabla_i f$ (or equivalently $\partial_i f$) transforms as the components of a co-vector, it follows that $\nabla^i \nabla_i f$ is a scalar, where, of course, ∇^i is defined by $\nabla^i = g^{ij}\, \nabla_j$. In fact this second-order differential operator is very important, and since it arises frequently it is customary to use a special symbol to denote it:

$$\Box f \equiv \nabla^i \nabla_i f \tag{1.116}$$

for any scalar f. Since it clearly reduces to the traditional Laplace operator if one specialises to $\mathbb{R}^n$ with Cartesian coordinates, it is, not surprisingly, called the *Laplacian.* It is the natural generalisation of the Cartesian-space Laplacian, which, unlike $g^{ij}\, \partial_i \partial_j$, always maps a scalar into another scalar, in any manifold with any metric and choice of local coordinate system. Explicitly, written out using the affine connection, it can be written, when acting on f, as

$$g^{ij}\, \partial_i\, \partial_j\, f - g^{ij}\, \Gamma^k{}_{ij}\, \partial_k\, f\,. \tag{1.117}$$

It is evident looking at the expression (1.110) for the affine connection that in general it can be quite tiresome to calculate $\Gamma^i{}_{jk}$, especially if the dimension n is large, since there are so many components to evaluate. In certain cases, and in fact the calculation of the scalar Laplacian is one of them, the task can be greatly simplified because only a specific contracted subset of the $\Gamma^i{}_{jk}$ are needed, namely

$$g^{ij}\, \Gamma^k{}_{ij}\,, \tag{1.118}$$

as can be seen from (1.117). From (1.110) we have

$$\begin{aligned}
g^{ij}\,\Gamma^k{}_{ij} &= \tfrac{1}{2}g^{ij}\,g^{k\ell}\,(\partial_i\,g_{\ell j}+\partial_j\,g_{i\ell}-\partial_\ell\,g_{ij})\,,\\
&= g^{ij}\,g^{k\ell}\,\partial_i\,g_{\ell j}-\tfrac{1}{2}g^{k\ell}\,g^{ij}\,\partial_\ell\,g_{ij}\,,\\
&= -g^{ij}\,g_{\ell j}\,\partial_i\,g^{k\ell}-\tfrac{1}{2}g^{k\ell}\,g^{ij}\,\partial_\ell\,g_{ij}\,,\\
&= -\delta^i_\ell\,\partial_i\,g^{k\ell}-\tfrac{1}{2}g^{k\ell}\,g^{ij}\,\partial_\ell\,g_{ij}\,,\\
&= -\partial_\ell\,g^{k\ell}-\tfrac{1}{2}g^{k\ell}\,g^{ij}\,\partial_\ell\,g_{ij}\,.
\end{aligned}\tag{1.119}$$

Note that in getting to the third line, we have use that $g^{k\ell}\,g_{\ell j}=\delta^k_j$, which is constant, and so $(\partial_i\,g^{k\ell})\,g_{\ell j}+g^{k\ell}\,(\partial_i\,g_{\ell j})=0$.

Now we use one further trick, which is to note that as a matrix expression, $g^{ij}\,\partial_\ell\,g_{ij}$ is just $\mathrm{tr}(\mathbf{g}^{-1}\,\partial_\ell\,\mathbf{g})$. But for any symmetric matrix we can write[5]

$$\det\mathbf{g}=e^{\mathrm{tr}\log\mathbf{g}}\,,\tag{1.120}$$

and so

$$\partial_\ell\,\det\mathbf{g}=(\det\mathbf{g})\,\mathrm{tr}(\mathbf{g}^{-1}\,\partial_\ell\,\mathbf{g})\,.\tag{1.121}$$

Thus we have

$$\tfrac{1}{2}g^{ij}\,\partial_\ell\,g_{ij}=\frac{1}{\sqrt{g}}\,\partial_\ell\,\sqrt{g}\,,\tag{1.122}$$

where we use the symbol g here to mean the determinant of the metric g_{ij}.

Putting all this together, we have

$$g^{ij}\,\nabla_i\,\partial_j\,f=g^{ij}\,\partial_i\,\partial_j\,f+(\partial_i\,g^{ij})\,\partial_j\,f+g^{ij}\,\frac{1}{\sqrt{g}}\,(\partial_i\,\sqrt{g})\,\partial_j\,f\,,\tag{1.123}$$

after making some convenient relabellings of dummy indices. Now we can see that all the terms on the right-hand side assemble together very nicely, giving us the following simple expression for the Laplacian:

$$\Box f=\nabla_i\,\nabla_i\,f=\frac{1}{\sqrt{g}}\,\partial_i\Big(\sqrt{g}\,g^{ij}\,\partial_j\,f\Big)\,.\tag{1.124}$$

This general expression gives us the Laplacian in an arbitrary coordinate system, for an arbitrary metric.

As a first example, suppose we choose to use polar coordinates (r,θ) in the plane $\mathbb{R}^2$, for which the metric will be $ds^2=dr^2+r^2\,d\theta^2$. From (1.114) we instantly see that the

[5] Prove by diagonalising the matrix, so that $\mathbf{g}\longrightarrow\mathrm{diag}(\lambda_1,\lambda_2,\ldots,\lambda_n)$. This means that $\det\mathbf{g}=\prod_i\lambda_i$, while $e^{\mathrm{tr}\log\mathbf{g}}=e^{\sum_i\log\lambda_i}$, and so the result is proven.

determinant of the metric is $g = r^2$, so plugging into (1.124) we get

$$g^{ij}\,\nabla_i\,\partial_j\,f \;=\; \frac{1}{r}\,\partial_i\Big(r\,g^{ij}\,\partial_j\,f\Big)\,,$$
$$\;=\; \frac{1}{r}\,\frac{\partial}{\partial r}\Big(r\,\frac{\partial f}{\partial r}\Big) + \frac{1}{r^2}\,\frac{\partial^2 f}{\partial\theta^2}\,. \tag{1.125}$$

This can be recognised as the standard expression for the Laplacian in two-dimensional flat space, written in polar coordinates.

As a slightly less trivial example, consider Euclidean 3-space, written in terms of spherical polar coordinates (r, θ, ϕ). These, of course, are related to the Cartesian coordinates (X, Y, Z) by

$$X = r\,\sin\theta\,\cos\phi\,, \qquad Y = r\,\sin\theta\,\sin\phi\,, \qquad Z = \cos\theta\,. \tag{1.126}$$

The metric, written in terms of the spherical polar coordinates, is therefore

$$ds^2 = dr^2 + r^2\,d\theta^2 + r^2\,\sin^2\theta\,d\phi^2\,. \tag{1.127}$$

The determinant is given by $g = r^4\,\sin^2\theta$ and so from (1.124) we get that the Laplacian is

$$\frac{1}{r^2}\,\frac{\partial}{\partial r}\Big(r^2\,\frac{\partial\psi}{\partial r}\Big) + \frac{1}{r^2}\,\Big[\frac{1}{\sin\theta}\,\frac{\partial}{\partial\theta}\,\Big(\sin\theta\,\frac{\partial\psi}{\partial\theta}\Big) + \frac{1}{\sin^2\theta}\,\frac{\partial^2\psi}{\partial\phi^2}\Big]\,. \tag{1.128}$$

Again, we can recognise the familiar three-dimensional Laplacian, written in spherical polar coordinates.

We close this discussion of the covariant derivative with a few further remarks. First, a matter of notation. A fairly widespread abbreviated notation for indicating partial differentiation and covariant differentiation is as follows:

$$V_{i,j} \equiv \partial_j V_i\,, \qquad V_{i;j} \equiv \nabla_j V_i\,. \tag{1.129}$$

In this example, we have chosen the case of a co-vector, but obviously the same comma and semicolon notation can be used for any type of tensor.

One other remark concerns applications of the covariant derivative. We introduced it by demanding the existence of a generalisation of the partial derivative that had the property of mapping tensors into tensors. In fact it essentially replaces the partial derivative in all situations where one is generalising from Cartesian tensor analysis to general-coordinate tensor analysis. In other words, if one has a tensorial equation in the context of Cartesian tensor analysis, and one wants to know how to generalise it to a tensorial equation in the general-coordinate tensor case, the rule can be more or less universally stated as "replace ∂_i by ∇_i everywhere." Or, in the notation mentioned in the previous paragraph, "replace commas by semicolons." In particular, one can easily show that this is always the correct rule to follow if one wants to convert a tensorial equation in flat Euclidean space from one written using Cartesian coordinates to one written using arbitrary curvilinear coordinates. There can be certain subtleties that sometimes arise if one wants to generalise a tensorial equation written in flat space to a tensorial equation for an arbitrary curved space, and we shall be discussing these shortly. But to a good approximation, the rule of "comma goes to semicolon" is a pretty reliable one.

1.10 The Riemann curvature tensor

Since the covariant derivative maps tensors into tensors, it follows also that if we compute the commutator of two covariant derivatives, namely

$$[\nabla_i, \nabla_j] \equiv \nabla_i \nabla_j - \nabla_j \nabla_i\,, \tag{1.130}$$

then this operator will also map tensors into tensors. Ostensibly this is a second-order differential operator, but in fact it is actually a purely algebraic operator, with no derivatives at all. This may seem surprising at first sight, but recalling that in the case of Euclidean space written in Cartesian coordinates it is the same as $[\partial_i, \partial_j]$, which is identically zero since partial derivatives commute, it is not so remarkable after all. In fact, the commutator $[\nabla_i, \nabla_j]$ is an object that characterises the *curvature* of the metric g_{ij} (or more precisely, of the connection $\Gamma^i{}_{jk}$). In fact, it gives rise to the so-called *Riemann tensor.*

First, let us look at $[\nabla_i, \nabla_j]$ acting on scalars. From the expression (1.103) for the covariant derivative acting on a co-vector we have that

$$\nabla_i \nabla_j\, f = \nabla_i \partial_j f = \partial_i \partial_j f - \Gamma^k{}_{ij}\, \partial_k f\,. \tag{1.131}$$

From this it follows that

$$[\nabla_i, \nabla_j] f = -(\Gamma^k{}_{ij} - \Gamma^k{}_{ji})\, \partial_k f\,, \tag{1.132}$$

since partial derivatives commute. Recalling that we also imposed the symmetry condition $\Gamma^k{}_{ij} = \Gamma^k{}_{ji}$, it therefore follows that

$$[\nabla_i, \nabla_j] f = 0\,. \tag{1.133}$$

Things are different if we instead consider acting on a vector with $[\nabla_i, \nabla_j]$. Now, we have

$$\nabla_i \nabla_j V^k = \partial_i \nabla_j V^k + \Gamma^k{}_{i\ell}\, \nabla_j V^\ell - \Gamma^\ell{}_{ij} \nabla_\ell V^k \tag{1.134}$$

and so on calculating the commutator the symmetry of $\Gamma^k{}_{ij}$ in ij implies the last term will drop out, leaving

$$[\nabla_i, \nabla_j]V^k = \partial_i(\partial_j V^k + \Gamma^k{}_{j\ell}\, V^\ell) + \Gamma^k{}_{i\ell}\, (\partial_j V^\ell + \Gamma^\ell{}_{jm}\, V^m) - (i \leftrightarrow j) \tag{1.135}$$

which, after distributing the derivatives yields

$$[\nabla_i, \nabla_j]V^k = (\partial_i \Gamma^k{}_{jm} - \partial_j \Gamma^k{}_{im} + \Gamma^k{}_{i\ell}\, \Gamma^\ell{}_{jm} - \Gamma^k{}_{j\ell}\, \Gamma^\ell{}_{im})\, V^m\,. \tag{1.136}$$

We see that as promised, there are no derivative terms at all left acting on the components V^k of the vector V. Although it is not manifest, we know from general arguments that the quantity in brackets multiplying V^k here *must* be a tensor, and we can rewrite (1.136) as

$$[\nabla_i, \nabla_j]V^k = R^k{}_{\ell ij}\, V^\ell\,, \tag{1.137}$$

where we have defined the *Riemann tensor*

$$R^i{}_{jk\ell} = \partial_k \Gamma^i{}_{\ell j} - \partial_\ell \Gamma^i{}_{kj} + \Gamma^i{}_{km}\, \Gamma^m{}_{\ell j} - \Gamma^i{}_{\ell m}\, \Gamma^m{}_{kj}\,. \tag{1.138}$$

One could laboriously verify that the quantity $R^i{}_{jk\ell}$ is indeed a tensor, by evaluating it in a primed coordinate system and using the known transformation properties of ∂_i and $\Gamma^i{}_{jk}$, but there really is no need to do so, since as remarked above, we *know* from our construction that it must be a tensor.

The Riemann tensor has several symmetry properties, most of which are slightly non-obvious, but can be proven by simply grinding out the algebra. First of all, there is a symmetry that is trivial to see, just by inspection of the definition (1.138):

$$R^i{}_{jk\ell} = -R^i{}_{j\ell k}\,. \tag{1.139}$$

The non-obvious ones are the *cyclic identity*

$$R^i{}_{jk\ell} + R^i{}_{k\ell j} + R^i{}_{\ell jk} = 0\,, \tag{1.140}$$

and two symmetries that follow after lowering the first index with the metric:

$$R_{ijk\ell} = -R_{jik\ell}\,, \qquad R_{ijk\ell} = R_{k\ell ij}\,, \tag{1.141}$$

where $R_{ijk\ell} \equiv g_{im}\, R^m{}_{jk\ell}$. There is also a differential identity satisfied by the Riemann tensor, namely

$$\nabla_m R^i{}_{jk\ell} + \nabla_k R^i{}_{j\ell m} + \nabla_\ell R^i{}_{jmk} = 0\,. \tag{1.142}$$

This is called the *Bianchi identity.*

Whereas the antisymmetry in the last index-pair in (1.139) is obvious merely from the definition (1.138), the other symmetries and identities follow only after one uses the expression (1.110) for ${\Gamma^i}_{jk}$. Note that the Riemann tensor would have fewer symmetries if we did not impose the condition (1.95) on ${\Gamma^i}_{jk}$. We shall not give details here, since it would be a bit of a diversion from the main thread of the development. Note that using the definition of total antisymmetrisation in (1.78), the cyclic identity (1.140) and Bianchi identity (1.142) can be written as

$$R^i{}_{[jk\ell]} = 0\,, \qquad \nabla_{[m} R^i{}_{|j|k\ell]} = 0\,. \tag{1.143}$$

In writing the Bianchi identity in this way we have introduced another piece of standard notation, namely that indices enclosed by vertical lines, such as $|j|$ in the above, are omitted from the antisymmetrisation.

The Riemann tensor characterises the curvature of the metric g_{ij}. To see how this works, first let us consider the case of flat Euclidean space, with the metric $g_{ij} = \delta_{ij}$. Obviously the Riemann tensor vanishes for this metric, since it is constructed (see eqn (1.138)) from the affine connection ${\Gamma^i}_{jk}$ and its first derivatives, and the affine connection is itself zero since it is constructed (see eqn (1.110)) from the first derivatives of the components of the metric.

What about flat Euclidean space written in some other coordinate system, such as polar coordinates on $\mathbb{R}^2$? We saw earlier that the components of the affine connection are now non-zero (see eqn (1.115)), so one might think that the Riemann tensor now has the possibility to be non-zero. However, the crucial point is that the Riemann tensor is a *tensor*, which means that if its components vanish in *any* coordinate frame then they vanish in *all* coordinate frames. This is an immediate consequence of the linearlity of the transformation of the components of any tensor field; see equation (1.69). One could instead demonstrate explicitly that the Riemann tensor vanishes by thrashing out the calculation of substituting the affine connection (1.115) into (1.138), but aside from being educational there is no point, since the general argument about the linearity of the tensor transformation already proves it must vansih.

In fact it can be shown that conversely, if the Riemann tensor of a metric g_{ij} vanishes then locally, at least, there always exists a general coordinate transformation $x^i \to x'^i = x'^i(x^j)$ that puts it in the form $g'_{ij} = \delta_{ij}$.

By contrast, suppose we now consider the metric $ds^2 = d\theta^2 + \sin^2\theta\, d\phi^2$ on the unit

2-sphere. Taking the coordinates to be $x^1 = \theta$, $x^2 = \phi$, we have

$$g_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & \sin^2\theta \end{pmatrix}\,, \qquad g^{ij} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\sin^2\theta} \end{pmatrix}\,. \tag{1.144}$$

Simple algebra using (1.110) leads to the following results for the comonents of the Christoffel connection:

$$\begin{aligned} &\Gamma^1{}_{11} = 0\,, \qquad \Gamma^1{}_{12} = 0\,, \qquad \Gamma^1{}_{22} = -\sin\theta\,\cos\theta\,, \\ &\Gamma^2{}_{11} = 0\,, \qquad \Gamma^2{}_{12} = \cot\theta\,, \qquad \Gamma^2{}_{22} = 0\,. \end{aligned} \tag{1.145}$$

From the symmetries of the Riemann tensor given above, it follows that in two dimensions there is only one independent component, and one easily finds that this is given by

$$R^1{}_{212} = \sin^2\theta\,. \tag{1.146}$$

For the 2-sphere, therefore, with its standard metric, we find that the Riemann tensor is non-zero; the metric has curvature.

1.10.1 The meaning of curvature

We have introduced the curvature tensor $R^i{}_{jk\ell}$ as the thing that arises when taking the commutator of covariant derivatives acting on a vector or tensor. It is instructive also to examine what it means from a more geometrical standpoint. To do this, we first need to introduce the notion of *parallel transport*. Suppose we have a curve in the manifold M, parameterised by $x^i = x^i(\lambda)$, where λ increases monotonically along the path. Suppose that at some point along the path we have a vector V, with components V^i in the local coordinate system we are using. We now wish to carry this vector along the curve, by parallel transport. The easiest way of seeing what this means is by thinking first about the case of Euclidean space, with a Cartesian coordinate system. Parallel transport means picking up the vector as if it were a little arrow, and carrying it along the path keeping it parallel to its original orientation at all times. In other words, the components of V will remain unchanged along the path, and so

$$\frac{dV^i}{d\lambda} = 0\,. \tag{1.147}$$

Note that another way of writing this, using the chain rule, is

$$\frac{dx^j}{d\lambda}\,\partial_j V^i = 0\,. \tag{1.148}$$

Now let us consider the generalisation of this concept of parallel transport to an arbitrary manifold. What will be the analogue of (1.147) and (1.148)? Clearly (1.147) and (1.148) wouldn't make sense in general, since they are not tensorial equations. (They don't transform like vectors under general coordinate transformations, for the usual reason that the transformation matrix used in expressing $V'^i = \partial x'^i/\partial x^j\, V^j$ will get differentiated by the $d/d\lambda$.) It is immediately clear what we should do; replace the ∂_i in (1.148) by ∇_i; this is in fact the *only* possible change that can give us a covariantly-transforming equation. Thus we write the parallel-transport equation as

$$\frac{DV^i}{D\lambda} \equiv \frac{dx^j}{d\lambda}\, \nabla_j V^i = \frac{dx^j}{d\lambda}\, (\partial_j V^i + \Gamma^i{}_{jk}\, V^k) = 0\,. \tag{1.149}$$

One can easily see that if one is in Euclidean space, and one transforms from Cartesian coordinates to an arbitrary coordinate system, then the equation (1.149) is *derivable* from (1.148). When one is in a general curved space, rather than Euclidean space, it is necessary to *define* what one means by parallel transport. The expression in (1.149) provides that definition. It is in fact the only possible covariant equation one could write down, that is constructed purely from first derivatives of the vector field, and that specialises properly to the Euclidean space case.

Having defined parallel transport, let us look at what happens if we parallel transport a vector around an infinitesimal closed curve C in M, starting and finishing at the point $x^i = 0$. At some point $x^i(\lambda)$ along the path, it therefore follows from (1.149) that an infinitesimal further displacement δx^i along it will result in the following change in V^i:

$$\delta V^i = -\Gamma^i{}_{jk}(x)\, V^k(x)\, \delta x^j\,. \tag{1.150}$$

Since the entire closed curve is itself infinitesimal in size, we can evaluate $V^i(x)$ and $\Gamma^i{}_{jk}(x)$ in terms of their expressions at the origin of the curve, by using Taylor's theorem up to first order in x^i:

$$\begin{aligned} V^i(x) &= V^i(0) + x^j\, \partial_j V^i(0) = V^i(0) - x^j\, \Gamma^i{}_{jk}(0)\, V^k(0) + \mathcal{O}(x^2)\,, \\ \Gamma^i{}_{jk}(x) &= \Gamma^i{}_{jk}(0) + x^\ell\, \partial_\ell \Gamma^i{}_{jk}(0) + \mathcal{O}(x^2)\,. \end{aligned} \tag{1.151}$$

We want to see how the vector is changed after it has been carried all the way around the closed infinitesimal curve C by parallel transport. We evaluate this by integrating around the curve:

$$\Delta V^i = \oint_C \delta V^i = -\oint_C \Gamma^i{}_{jk}(x)\, V^k(x)\, dx^j\,. \tag{1.152}$$

Using the expressions in (1.151), and working just up to linear order in x^i, we therefore find

$$\Delta V^i = -\Gamma^i{}_{jk}(0)\, V^k(0) \oint_C dx^j - [\partial_\ell \Gamma^i{}_{jk}(0) - \Gamma^i{}_{jm}(0)\, \Gamma^m{}_{\ell k}(0)]\, V^k(0) \oint x^\ell\, dx^j\,. \qquad (1.153)$$

The first term is zero, because dx^j is an exact differential, and so it is equal to the difference between x^i at the start and the finish of the curve. But since the curve is closed, the start and finish are the same point and hence the integral gives zero.

For the remaining term in (1.153), we may note that the integral is antisymmetric in ℓ and j, since we have

$$\oint_C x^\ell\, dx^j = \oint_C d(x^\ell\, x^j) - \oint_C x^j\, dx^\ell\,, \qquad (1.154)$$

and the first term on the right-hand side is zero because $d(x^\ell\, x^j)$ is an exact differential. Thus we may rewrite (1.153) as

$$\Delta V^i = -\tfrac{1}{2}[\partial_\ell \Gamma^i{}_{jk} - \partial_j \Gamma^i{}_{\ell k} - \Gamma^i{}_{jm}\, \Gamma^m{}_{\ell k} + \Gamma^i{}_{\ell m}\, \Gamma^m{}_{jk}]\, V^k \oint x^\ell\, dx^j\,, \qquad (1.155)$$

where we have suppressed the (0) arguments on the connection and vector. Comparing with the expression (1.138) for the Riemann tensor, we see that, after some index reorganisation we have

$$\Delta V^i = -\tfrac{1}{2} R^i{}_{jk\ell}\, V^j \oint_C x^k\, dx^\ell\,. \qquad (1.156)$$

The integral $\oint_C x^k\, dx^\ell$ just gives the area element of the infinitesimal loop. Think, for example, of an infinitesimal loop taken to be a rectangle in the (x,y) plane, with its four vertices at

$$(x,y) = (0,0)\,, \quad (\Delta x, 0)\,, \quad (\Delta x, \Delta y)\,, \quad (0, \Delta y)\,. \qquad (1.157)$$

If we define $\Delta A^{ij} = \oint_C x^i\, dx^j$, then it is easy to see that

$$\Delta A^{11} = \Delta A^{22} = 0\,, \qquad \Delta A^{12} = -\Delta A^{21} = \Delta x\, \Delta y\,, \qquad (1.158)$$

where $x^1 \equiv x$ and $x^2 \equiv y$. Thus ΔA^{ij} is the area element of the loop, with its indices indicating the plane in which the loop lies. The upshot from (1.156) is that after parallel-transporting a vector V around an infinitesimal closed loop spanned by the area element ΔA^{ij}, the components of the vector are changed by an amount ΔV^i, given by

$$\Delta V^i = -\tfrac{1}{2} R^i{}_{jk\ell}\, V^j\, \Delta A^{k\ell}\,. \qquad (1.159)$$

Thus the Riemann tensor characterises the way in which vectors are modified by parallel transport around closed curves. In particular, if the space is flat, there will be no change.

1.10.2 The Ricci tensor, Ricci scalar and Weyl tensor

By contracting indices on the Riemann tensor, one can form tensors of lower rank, namely 2 and 0. First, by taking one contraction, we form the *Ricci tensor*

$$R_{ij} = R^k{}_{ikj}\,. \tag{1.160}$$

It follows from the prevsiously-discussed symmetries of the Riemann tensor that the Ricci tensor is symmetric, i.e. $R_{ij} = R_{ji}$.

A further contraction of the Ricci tensor, performed with the use of the inverse metric, yields the *Ricci scalar*

$$R = g^{ij}\, R_{ij}\,. \tag{1.161}$$

It follows from the Bianchi identity (1.142) that the divergence of the Ricci tensor is related to the gradient of the Ricci scalar:

$$\nabla^i R_{ij} = \tfrac{1}{2}\nabla_j R\,. \tag{1.162}$$

In several contexts, most notably in general relativity, another tensor that plays a very important role is the *Einstein tensor*, whose definition is

$$G_{ij} = R_{ij} - \tfrac{1}{2}R\, g_{ij}\,. \tag{1.163}$$

Note that from (1.162) it follows that

$$\nabla^i G_{ij} = 0\,. \tag{1.164}$$

This fact that the Einstein tensor is *conserved* is very crucial in general relativity.

Another important notion is a special type of metric called an *Einstein metric*. An Einstein metric is defined to be one whose Ricci tensor satisfies

$$R_{ij} = \lambda\, g_{ij}\,, \tag{1.165}$$

where λ is a constant. Note that if the dimension n is greater than 2, we can prove that λ *must* be a constant, if we merely begin by assuming that (1.165) holds for some function λ. The proof is as follows: Taking the divergence of (1.165), and using (1.162), we find that

$$\tfrac{1}{2}\nabla_j R = \nabla_j \lambda\,. \tag{1.166}$$

On the other hand, contracting (1.165) with g^{ij} we obtain

$$R = n\,\lambda\,. \tag{1.167}$$

Combining the two equations gives

$$(n-2)\,\nabla_j \lambda = 0\,, \tag{1.168}$$

and hence λ must be a constant if $n > 2$. Einstein metrics are of considerable importance in physics and mathematics, and we shall encounter them frequently later in the course.

Since they are obtained by contracting indcies on the Riemann tensor, the information contained in the Ricci tensor or Ricci scalar is in general less than that contained in the full Riemann tensor; the mapping is non-reversible and one cannot reconstruct the Riemann tensor from the Ricci tensor. In fact the "extra" information that is contained in the Riemann tensor but not in the Ricci tensor is characterised by a tensor called the *Weyl tensor*, defined in n dimensions by

$$C^i{}_{jk\ell} = R^i{}_{jk\ell} - \frac{1}{n-2}\,(R^i{}_k\, g_{j\ell} - R^i{}_\ell\, g_{jk} + R_{j\ell}\, \delta^i_k - R_{jk}\, \delta^i_\ell) + \frac{1}{(n-1)(n-2)}\, R\,(\delta^i_k\, g_{j\ell} - \delta^i_\ell\, g_{jk})\,. \tag{1.169}$$

The Weyl tensor has the property, as can easily be verified from (1.169), that the contraction $C^i{}_{jik}$ is zero.

Although it naturally arises as a $(1,3)$ tensor, the expression for the Weyl tensor in terms of the Riemann tensor looks a little more elegant if we write it with the upper index lowered, to give

$$C_{ijk\ell} = R_{ijk\ell} - \frac{1}{n-2}\,(R_{ik}\, g_{j\ell} - R_{i\ell}\, g_{jk} + R_{j\ell}\, g_{ik} - R_{jk}\, g_{i\ell}) + \frac{1}{(n-1)(n-2)}\, R\,(g_{ik}\, g_{j\ell} - g_{i\ell}\, g_{jk})\,. \tag{1.170}$$

One can show by a lengthy but straightforward calculation that the Weyl tensor $C^i{}_{jk\ell}$ is conformally invariant, in the following sense. Suppose we have two metrics, g_{ij} and $\tilde g_{ij}$, which are related to one another by what is called a conformal transformation:

$$\tilde g_{ij} = \Omega^2\, g_{ij}\,. \tag{1.171}$$

Here, the factor Ω is allowed to depend arbitrarily on the coordinates. After some algebra, involving first calculating the relation between the affine connections $\Gamma^i{}_{jk}$ and $\widetilde\Gamma^i{}_{jk}$ for the two metrics using (1.110), and then the relation between the two Riemann tensors $R^i{}_{jk\ell}$ and $\widetilde R^i{}_{jk\ell}$ using (1.138), one eventually finds that the two type $(1,3)$ Weyl tensors are identical,

$$\widetilde C^i{}_{jk\ell} = C^i{}_{jk\ell}\,. \tag{1.172}$$

1.10.3 Index-free notation: Torsion and curvature

It may not have escaped the reader's attention that the discussion in the last few sections has become somewhat more "index oriented" than in the earlier parts of these lecture notes.

This is largely because when it comes to doing practical calculations, the use of indices, and explicit coordinate frames, generally makes things easier. However, it is perhaps worthwhile to look at a couple of topics we have already covered from a more geometrical and abstract standpoint. If nothing else, this may help anyone who wants to look at textbooks or papers that adopt an abstract approach.

Let us begin with the covariant derivative. We can define a connection ∇ at a point p in the manifold M as a rule that assigns to each vector field X a differential operator ∇_X which maps a vector field Y to another vector field $\nabla_X Y$, with the following properties:

$$\begin{aligned}
\text{Tensor in } X: &\quad \nabla_{fX+gY} Z = f\nabla_X Z + g\nabla_Y Z\,,\\
\text{Linear in } Y: &\quad \nabla_X(\alpha Y + \beta Z) = \alpha\nabla_X Y + \beta\nabla_X Z\,,\\
\text{Leibnitz:} &\quad \nabla_X(fY) = X(f)\,Y + f\,\nabla_X Y\,,
\end{aligned} \tag{1.173}$$

where X, Y and Z are vector fields, f and g are functions on M, and α and β are constants. We can say that $\nabla_X Y$ is the covariant derivative of Y along the direction of X. In more familiar index notation, then if we give the vector $\nabla_X Y$ the name W, i.e. $W = \nabla_X Y$, then we shall have

$$W^i = X^j\,\nabla_j Y^i\,. \tag{1.174}$$

Of course once we write it out in components, the first property in (1.173), namely that $\nabla_X Y$ is tensorial in X, is obvious, since if we multiply X^i by a function in (1.174), clearly the expression is simply multiplied by that function. The point about being "tensorial" in X, which is not *a priori* obvious in the abstract definition, and thus needs to be stated as one of the defining properties, is the following. We have seen repreatedly that the thing that can stop something transforming as a tensor is if a derivative lands on a transformation matrix $\partial x^i/\partial x'^j$ or $\partial x'^i/\partial x^j$ when one transforms from one coordinate frame to another. The statement that $\nabla_{fX} Y = f\nabla_X Y$ is sufficient to ensure that we will not run into any trouble from the transformation matrix applied to the vector X getting differentiated when we change coordinates.

We now define the *torsion tensor* T by

$$T(X,Y) = \nabla_X Y - \nabla_Y X - [X,Y]\,, \tag{1.175}$$

where X and Y are arbitrary vector fields, and the bracket $[X,Y]$ is simply the commutator of vector fields, giving another vector. If we act with this vector on a function f in the usual way (recall that a vector V acting on a function f returns another function, namely

$Vf = V^i \partial_i f$, which is the directed derivative of f along V), we have

$$[X, Y]f = X(Yf) - Y(Xf)\,, \tag{1.176}$$

where $X(Yf)$ just means the directed derivative of Yf along X. If we define $W = [X, Y]$, then in components we have

$$W^i = X^j \partial_j Y^i - Y^j \partial_j X^i\,. \tag{1.177}$$

One easily verifies by explicitly changing to a new coordinate system that W^i indeed transforms in the proper manner for the components of a vector. (This exercise was on the first problem sheet.) Note that from (1.177) we can easily see that

$$[fX, Y] = f\,[X, Y] - (Yf)\,X\,, \qquad [X, fY] = f\,[X, Y] + (Xf)\,Y\,. \tag{1.178}$$

Returning now to the definition of the torsion tensor T in (1.175), we can check that this is indeed tensorial in X and Y, namely

$$T(fX, Y) = fT(X, Y)\,, \tag{1.179}$$

and similarly in Y, for any function f. For example, we have

$$\begin{aligned} T(fX, Y) &= \nabla_{fX} Y - \nabla_Y (fX) - [fX, Y]\,, \\ &= f\nabla_X Y - f\nabla_Y X - (Yf)\,X - f[X, Y] + (Yf)\,X\,, \\ &= f(\nabla_X Y - \nabla_Y X - [X, Y])\,, \\ &= fT(X, Y)\,, \end{aligned} \tag{1.180}$$

where we have made use of the last equation in (1.173), and the first equation in (1.178). Obviously, the calculation for $T(X, fY)$ proceeds identically.

Note that $T(X, Y)$ itself is a vector. Writing out (1.175) in terms of components, we see that it gives

$$\begin{aligned} [T(X, Y)]^i &= X^j \nabla_j Y^i - Y^j \nabla_j X^i - X^j \partial_j Y^i + Y^j \partial_j X^i\,, \\ &= X^j (\partial_j Y^i + \Gamma^i{}_{jk} Y^k) - Y^j (\partial_j X^i + \Gamma^i{}_{jk} X^k) - X^j \partial_j Y^i + Y^j \partial_j X^i\,, \\ &= (\Gamma^i{}_{jk} - \Gamma^i{}_{kj})\, X^j Y^k\,. \end{aligned} \tag{1.181}$$

We may define the components of the torsion tensor by

$$[T(X, Y)]^i \equiv T^i{}_{jk}\, X^j Y^k \tag{1.182}$$

for any vector fields X and Y, and so we have

$$T^i{}_{jk} = (\Gamma^i{}_{jk} - \Gamma^i{}_{kj})\,. \tag{1.183}$$

It is of course, from its definition (1.175), antisymmetric in its lower indices j and k, as we see in (1.183). If we make our usual assumption that $\Gamma^i{}_{jk}$ will be symmetric in its lower indices then the torsion vanishes, $T^i{}_{jk} = 0$. With a more general choice of connection, the torsion can be non-zero. Note that despite looking like a differential operator, the calculations above show that $T(X,Y)$ is actually purely algebraic.

The abstract way of defining the Riemman tensor is rather similar. Given arbitrary vector fields X, Y and Z we define

$$R(X,Y)Z = [\nabla_X, \nabla_Y]Z - \nabla_{[X,Y]}Z\,, \tag{1.184}$$

where, of course, $[\nabla_X, \nabla_Y]Z$ just means $\nabla_X(\nabla_Y Z) - \nabla_Y(\nabla_X Z)$. Again, one can verify from the previous definitions that $R(X,Y)Z$ is tensorial in X, Y and Z, which we could summarise in the single equation

$$R(fX, gY)(hZ) = fgh R(X,Y)Z \tag{1.185}$$

for any functions f, g and h. This property again means that despite superficial appearances, $R(X,Y)Z$ defined in (1.184) is not a differential operator, but is purely algebraic. Note that $R(X,Y)Z$ itself is a vector. If we define

$$[R(X,Y)Z]^i \equiv R^i{}_{jk\ell}\, Z^j X^k Y^\ell\,, \tag{1.186}$$

then a straightforward calculation from (1.184) shows that $R^i{}_{jk\ell}$ is precisely given by the same expression (1.138) that we obtained previously.

1.11 Differential Forms

1.11.1 Definition of a p-form

We have already remarked that totally-antisymmetric co-tensors play a particularly important rôle in mathematics and physics. Recall that when we expand any co-tensor ω of type $(0,p)$ in a coordinate basis, we shall have

$$\omega = \omega_{i_1\cdots i_p}\, dx^{i_1} \otimes \cdots \otimes dx^{i_p}\,. \tag{1.187}$$

If $\omega_{i_1\cdots i_p}$ should happen to be totally antisymmetric in its indices, then we can choose to antisymmetrise the basis itself. Thus, for the two-index case, we define

$$dx^i \wedge dx^j \equiv dx^i \otimes dx^j - dx^j \otimes dx^i\,, \tag{1.188}$$

for the three-index case, we define

$$\begin{aligned} dx^i \wedge dx^j \wedge dx^k &\equiv dx^i \otimes dx^j \otimes dx^k + dx^j \otimes dx^k \otimes dx^i + dx^k \otimes dx^i \otimes dx^j \\ &\quad - dx^i \otimes dx^k \otimes dx^j - dx^j \otimes dx^i \otimes dx^k - dx^k \otimes dx^j \otimes dx^i \,, \end{aligned} \tag{1.189}$$

and so on. In general we shall have

$$\begin{aligned} dx^{i_1} \wedge \cdots \wedge dx^{i_p} &\equiv dx^{i_1} \otimes \cdots \otimes dx^{i_p} + \text{even permutations} \\ &\quad - \text{odd permutations}\,. \end{aligned} \tag{1.190}$$

From its definition, we see that the *wedge product* is antisymmetric, and so, for example,

$$dx^i \wedge dx^j = -dx^j \wedge dx^i\,. \tag{1.191}$$

Suppose that A is a rank-p totally antisymmetric co-tensor. Then using the definition above we can write

$$A = \frac{1}{p!}\, A_{i_1 \cdots i_p}\, dx^{i_1} \wedge \cdots \wedge dx^{i_p}\,. \tag{1.192}$$

Such a co-tensor is called a p-form. Suppose that analogously, B is a q-form. It is straightforward to see, using the definitions above, that we must have

$$A \wedge B = (-1)^{pq}\, B \wedge A\,. \tag{1.193}$$

Note that a scalar field is a 0-form, and a co-vector field is a 1-form.

1.11.2 Exterior derivative

We now define the *exterior derivative*, which acts on a p-form field and produces from it a $(p+1)$-form. Acting on a 0-form field f, it gives the 1-form df defined by

$$\langle df | V \rangle = V f \tag{1.194}$$

where V is any vector, and as usual Vf just means $V^i \partial_i f$. Acting on a p-form A, expanded as in (1.192), the exterior derivative is defined by

$$dA = \frac{1}{p!}\, (dA_{i_1 \cdots i_p}) \wedge dx^{i_1} \wedge \cdots dx^{i_p}\,. \tag{1.195}$$

Note that, as we have already observed in section 1.5, the definition (1.194) is equivalent to the component equation

$$df = \partial_i f\, dx^i\,. \tag{1.196}$$

Likewise, we can re-express the definition (1.195) as

$$dA = \frac{1}{p!}\,(\partial_j A_{i_1\cdots i_p})\,dx^j \wedge dx^{i_1} \wedge \cdots \wedge dx^{i_p}\,. \tag{1.197}$$

Since dA is a $(p+1)$-form, which we can expand in terms of its components as

$$dA = \frac{1}{(p+1)!}\,(dA)_{j_1\cdots j_{p+1}}\,dx^{j_1} \wedge \cdots \wedge dx^{j_{p+1}}\,, \tag{1.198}$$

we see, by comparing with (1.197), that the components of dA are given by

$$(dA)_{j_1\cdots j_{p+1}} = (p+1)\,\partial_{[j_1} A_{j_2\cdots j_{p+1}]}\,, \tag{1.199}$$

where the square brackets, denoting total antisymmetrisation, were defined in (1.78).

It is straightforward to check, by performing the standard arbitrary change of coordinates from x^i to $x'^i = x'^i(x^j)$, that the components of the $(p+1)$-form dA do indeed transform in the correct way for the components of a co-tensor of rank $(p+1)$. In other words, the "undesirable" terms that arise when one simply takes the partial derivatives of the components of a general tensor all miraculously cancel out when one looks at the special case of the totally-antisymmetrised partial derivatives of the components of a totally-antisymmetric rank-p co-tensor.

Two very important properties of the exterior derivative are the following. First, it is easily seen from the definitions that if A is a p-form and B is a q-form, then the following Leibnitz rule holds:

$$d(A\wedge B) = dA\wedge B + (-1)^p\,A\wedge dB\,. \tag{1.200}$$

Secondly, it is also easy to see from the definition of d that if it acts twice, it automatically gives zero, i.e.

$$d^2 \equiv 0\,. \tag{1.201}$$

This just follows from (1.197), which shows that d is an *antisymmetric* derivative, while on the other hand partial derivatives *commute*. For example, if we apply d twice to a scalar function f, we get

$$d^2 f = d(\partial_i f\,dx^i) = \partial_j\partial_i f\,dx^j\wedge dx^i = \partial_{[j}\partial_{i]} f\,dx^j\wedge dx^i\,, \tag{1.202}$$

where, in the last step, we have placed the antisymmetrisation brackets around the i and j indices just to emphasise that this antisymmetry is being enforced by the contraction onto the wedge product $dx^j\wedge dx^i$. It is now manifest that $d^2 f = 0$, since obviously $\partial_i\partial_j f = \partial_j\partial_i f$. Similarly, if A is a 1-form then applying d twice gives

$$d^2 A = d(\partial_j A_i\,dx^j\wedge dx^i) = \partial_k\partial_j A_i\,dx^k\wedge dx^j\wedge dx^i = \partial_{[k}\partial_j A_{i]}\,dx^k\wedge dx^j\wedge dx^i\,, \tag{1.203}$$

and again the fact that the partial derivatives commute immediately implies that we must have $d^2 A = 0$.

It is worth remarking that in three dimensions, using Cartesian coordinates on $\mathbb{R}^3$, the statement $d^2 f = 0$ is probably more familiar as the statement that

$$\operatorname{curl}\operatorname{grad} f = 0\,, \tag{1.204}$$

i.e. $\vec{\nabla}\times\vec{\nabla} f = 0$: Recall that if one writes out the three components of this equation, it says

$$\partial_x\partial_y f - \partial_y\partial_x f = 0\,,\qquad \partial_y\partial_z f - \partial_z\partial_y f = 0\,,\qquad \partial_z\partial_x f - \partial_x\partial_z f = 0\,, \tag{1.205}$$

which is just the statement $\partial_{[i}\partial_{j]} f = 0$. In fact a bit later, after we have introduced the further concept of *Hodge dualisation*, we shall be able to give a more extensive comparison between the notation of differential forms and three-dimensional Cartesian tensor analysis.

There is another remark that can be made now, although we shall have much more to say about the matter later on. We have noted that $d^2 = 0$ when acting on *any* differential form of any degree. This means that if B is the p-form given by $B = dA$, where A is any $(p-1)$-form, then we shall have that $dB = 0$. Any differential form ω that satisfies $d\omega = 0$ is called a *closed* form. Any differential form B that is written as $B = dA$ is called an *exact* form. Thus we have the statement that *any exact form is closed.* What about the converse? Suppose we have a closed differential p-form ω, i.e. it satisfies $d\omega = 0$. Can we necessarily write it as $\omega = d\nu$, for some $(p-1)$-form ν? Tha answer is that *locally*, we can always find a $(p-1)$-form ν that gives $\omega = d\nu$. However, it might be that the $(p-1)$-form ν is singular somewhere on the manifold M, even though ω is completely non-singular.

If the manifold is $\mathbb{R}^n$, meaning that it is topologically trivial, then the local differential equations that one would solve in order to find the $(p-1)$-form ν whose exterior derivative produces the closed p-form ω will have a globally-defined non-singular solution (if ω is non-singular), and so we can say that in this case all closed forms are exact. But if the manifold has non-trivial topology, such as, for example, the 2-sphere, then not all closed forms are exact. This is an extremely important topic in differential geometry, and it is one to which we shall return in due course.

1.12 Integration, and Stokes' Theorem

Integration over manifolds is a natural generalisation of the familiar idea of integration. The most basic integral we could consider is the one-dimensional integral of $f(x)$:

$$\int_{f_1}^{f_2} df = f_2 - f_1 = [f]_{x_1}^{x_2}\,, \tag{1.206}$$

where f_1 and f_2 denote the values of the function f at the begining and end of the integration range. The expression (1.206) is known as the *fundamental theorem of calculus.* In the language of differential forms, we can view (1.206) as the integration of the 1-form df over the 1-dimensional manifold that is the line interval along which the integration is performed. If we call this manifold M, then its endpoints, at x_1 and x_2, correspond to the *boundary* of M. The boundary of any manifold is a manifold of one dimension less (for example the 2-sphere can be thought of as the boundary of the unit ball in $\mathbb{R}^3$). Thus in our example, the boundary of the 1-dimensional manifold of the line interval consists of the two points, x_1 and x_2; these are of dimension zero. In general, we denote the boundary of a manifold M by ∂M.

The one-dimensional integral (1.206) can then be written as

$$\int_M df = \int_{\partial M} f \,. \tag{1.207}$$

The "integral" on the right-hand side here is a bit degenerate, since it is an integral over the zero-dimensional manifold consisting of just the two endpoints of the line interval. A zero-dimensional integral is nothing but the difference of the values of the "integrand" at the points on the 0-manifold;

$$\int_{\partial M} f = [f]_{x_1}^{x_2} = f_2 - f_1 \,. \tag{1.208}$$

The reason for writing the integral in the somewhat esoteric way (1.207) is that it admits an immediate generalisation to the much more interesting case of integration of p-forms over p-manifolds.

Just as a 1-form, such as the differential df, can be integrated over a 1-manifold, so a p-form is integrated over a p-dimensional manifold. This is perfectly reasonable, since a p-form A is written as

$$A = \frac{1}{p!}\, A_{i_1 \cdots i_p}\, dx^{i_1} \wedge \cdots \wedge dx^{i_p} \,, \tag{1.209}$$

which involves a p-fold tensor product of coordinate differentials. The evaluation of the integral of A over a p-dimensional manifold M proceeds just like traditional multi-dimensional integrals. For example, if we have a 2-form $A = a(x,y) dx \wedge dy$, and we wish to integrate it over a 2-manifold M that is spanned by the local coordinates x and y, then we would just have

$$\int_M A = \int_{x_1}^{x_2} dx \int_{y_1}^{y_2} dy\, a(x,y) \,, \tag{1.210}$$

where the limits on the x and y integrals are such that the integration is over the entire 2-volume of the manifold M.

If we have a p-form A that is exact, that is to say that it can be expressed as $A = d\omega$, where ω is some $(p-1)$-form, then a very important theorem, called *Stokes' theorem*, says that

$$\int_M d\omega = \int_{\partial M} \omega\,. \tag{1.211}$$

In order for this to be valid, ω must be a smooth $(p-1)$-form on the manifold M over which the integration is performed. Note that (1.211 is a generalisation of our previous 1-dimensional integral in (1.207). The proof of Stokes' theorem is very analogous to the way Stokes' theorem and the divergence theorem are proved in 3-dimensional vector calculus. Essentially, one breaks the integration region up into little hypercubes, and shows that the "volume integral" over each hypercube can be turned into an integral over its boundary surface. We shall not pause to prove Stokes' theorem (1.211), but we shall take a look presently at special cases that reduce to the familiar Stokes' and diverence theorems of vector calculus.

First, a few remarks:

1. If M is an n-manifold without boundary (such as the surface of a sphere), then $\partial M = \emptyset$, and it follows that for any *exact* n-form σ, we must have $\int_M \sigma = 0$. This follows from Stokes' theorem. Suppose that the exact form σ is written as $\sigma = d\alpha$, where α is some $(n-1)$-form. Then using (1.211) we shall have

$$\int_M \sigma = \int_M d\alpha = \int_{\partial M} \alpha = 0\,, \tag{1.212}$$

where the last step follows from the fact that we are supposing M has no boundary; $\partial M = \emptyset$.

2. Just as the exterior derivative d has the property that $d^2 = 0$, so manifolds have the "dual" property that $\partial^2 = 0$, i.e. *the boundary of a boundary is zero.* We prove this by taking ω to be an arbitrary $(n-2)$-form, and applying Stokes' theorem twice:

$$0 = \int_M d^2\omega = \int_{\partial M} d\omega = \int_{\partial^2 M} \omega\,. \tag{1.213}$$

(The initial 0 of course comes from the fact that $d^2\omega$ vanishes identically.) Since (1.213) is true for any $(n-2)$-form ω, it follows that $\partial^2 M$ must be zero for any M. This is clearly a reasonable result. For example, we take the boundary of the unit ball in $\mathbb{R}^3$, and we get S^2. And sure enough, S^2 has no boundary.

As one would expect and hope, the integration of an n-form over an n-manifold is independent of the choice of coordinates that one uses. It may, of course, be convenient in practice to choose a specific coordinate system, but crucially, we will get the same answer if we change to any other coordinate system. Suppose we choose local coordinates x^i on an n-manifold M, and integrate the n-form[6]

$$\omega = f\, dx^1 \wedge dx^2 \wedge \cdots \wedge dx^n \tag{1.214}$$

over M. Under a change of coordinates $x^i \to x'^i = x'^i(x^j)$, we shall therefore have

$$\begin{aligned} \omega &= f\, \frac{\partial x^1}{\partial x'^{i_1}}\, \frac{\partial x^2}{\partial x'^{i_2}} \cdots \frac{\partial x^n}{\partial x'^{i_n}}\, dx'^{i_1} \wedge dx'^{i_2} \wedge \cdots \wedge dx'^{i_n} \\ &= f\, \frac{\partial x^1}{\partial x'^{i_1}}\, \frac{\partial x^2}{\partial x'^{i_2}} \cdots \frac{\partial x^n}{\partial x'^{i_n}}\, \varepsilon^{i_1 i_2 \cdots i_n}\, dx'^1 \wedge dx'^2 \wedge \cdots \wedge dx'^n\,, \end{aligned} \tag{1.215}$$

where $\varepsilon^{i_1 i_2 \cdots, i_n}$ is defined to be $+1$ if $(i_1, i_2, \ldots, i_n)$ is an even permutation of $(1, 2, \ldots, n)$, -1 if it is an odd permutation, and 0 if it is no permuation at all (meaning that at least two index values must be equal). With a bit of thought, one can recognise that

$$\frac{\partial x^1}{\partial x'^{i_1}}\, \frac{\partial x^2}{\partial x'^{i_2}} \cdots \frac{\partial x^n}{\partial x'^{i_n}}\, \varepsilon^{i_1 i_2 \cdots i_n} = \Big|\frac{\partial x}{\partial x'}\Big|\,, \tag{1.216}$$

where $\Big|\frac{\partial x}{\partial x'}\Big|$ means the Jacobian of the transformation from the x^i to the x'^j coordinates, i.e.

$$\Big|\frac{\partial x}{\partial x'}\Big| = \det\Big(\frac{\partial x^i}{\partial x'^j}\Big)\,. \tag{1.217}$$

This accords with what one knows from elementary mathematics, namely that if one changes variables in a multi-dimensional integral, one must multiply the integrand by the Jacobian of the transformation. If the reader is in doubt about the steps above, it is well worthwhile to look explicitly at the case of a 2-dimensional integral of a 2-form. Thus one has

$$\begin{aligned} \int f dx \wedge dy &= \int f\Big(\frac{\partial x}{\partial x'}\, dx' + \frac{\partial x}{\partial y'} dy'\Big) \wedge \Big(\frac{\partial y}{\partial x'}\, dx' + \frac{\partial y}{\partial y'}\, dy'\Big) \\ &= \int f\Big(\frac{\partial x}{\partial x'}\, \frac{\partial y}{\partial y'} - \frac{\partial x}{\partial y'}\, \frac{\partial y}{\partial x'}\Big)\, dx' \wedge dy'\,. \end{aligned} \tag{1.218}$$

So we see that the antisymmetry of the wedge product automatically handles the Jacobian transformation when changing variables.

Integration over a manifold probes properties that go beyond just the local structure in a neighbourhood. A very simple illustration is provided by the following example. Suppose we have a circle, S^1, for which we set up a local coordinate θ, with $0 < \theta < 2\pi$. We already saw in section 1.3.1 that this coordinate cannot be well-defined everywhere on S^1; here, we have omitted the point $\theta = 0$ (which is identified with $\theta = 2\pi$). However, suppose for a

[6]Note that *any* n-form in n dimensions must simply be of the form of a function times the wedge product of all the coordinate differentials.

moment we mistakenly thought that θ was well-defined everywhere on S^1, meaning that θ was a well-defined function on S^1. We could quickly discover the mistake by applying Stokes' theorem and encountering the following contradiction:

$$2\pi = \int_{S^1} d\theta = \int_{\partial S^1} \theta = 0\,. \tag{1.219}$$

On the left-hand side, we present the standard integration around the unit-radius circle; in the middle step we have used Stokes' theorem to convert the integral of $d\theta$ around S^1 into an integral of θ over the boundary of S^1; on the right-hand side we have used the fact that S^1 has no boundary, and therefore this integral vanishes.

The mistake in the above sequence of steps was, of course, in the application of Stokes' theorem. The point is that θ is not a well-defined smooth function on S^1; it undergoes a discontinuous jump from 2π to 0 as one rotates anticlockwise and passes the point $(x, y) = (1, 0)$ on the circle. Thus θ is not a smooth 0-form, and so Stokes' theorem cannot be used. Note that when we write $d\theta$, we are really using a bit of a short-hand. What is meant is the 1-form that is expressed locally as $\omega \equiv d\theta$ when $0 < \theta < 2\pi$, i.e. in the patch called U_1 in section 1.3.1. To cover the patch of S^1 that include $(x, y) = (1, 0)$ (but excludes $(x, y) = (-1, 0)$), i.e. the patch called U_2 in 1.3.1, we can use the coordinate $\tilde{\theta}$. The globally-defined 1-form can be written as $\omega \equiv d\tilde{\theta}$ in that patch. Note that everywhere in the overlap region $U_1 \cap U_2$, the two expressions $d\theta$ and $d\tilde{\theta}$ for the 1-form agree. The essential point to note here is that there exists a globally-defined 1-form ω, but there exists no globally-defined 0-form whose exterior derivative gives ω. Thus ω is a closed 1-form that is not exact. It is in fact the volume form on S^1; its integral over S^1 gives the 1-dimensional "volume" of the unit circle; i.e. 2π.

To see this using Stokes' theorem, we can do the following. Divide the unit circle into two hemispheres (or, perhaps, we should say "hemicircles,"), namely the H_1 defined by points on $x^2 + y^2 = 1$ in $\mathbb{R}^2$ with $x < 0$, and H_2 defined by points with $x > 0$. In other words, H_1 is the left-hand half of the circle, and H_2 is the right-hand half. On H_1 we can use θ as coordinate, since H_1 lies entirely within the patch U_1, whilst on H_2 we can use $\tilde{\theta}$ as coordinate, since H_2 lies entirely within the patch U_2. Then we may calculate as follows:

$$\begin{aligned}
\int_{S^1} \omega &= \int_{H_1} \omega + \int_{H_2} \omega = \int_{H_1} d\theta + \int_{H_2} d\tilde{\theta} \\
&= \int_{\partial H_1} \theta + \int_{\partial H_2} \tilde{\theta} \\
&= [\theta]_{\pi/2}^{3\pi/2} + [\tilde{\theta}]_{\pi/2}^{3\pi/2} = \pi + \pi = 2\pi\,.
\end{aligned} \tag{1.220}$$

Note that our applications of Stokes' theorem are completely valid here, since in each of the patches where we are using it, ω is written as the exterior derivative of a function that is well-defined and non-singular within that patch.

As another example, consider the 2-form

$$\omega = \sin\theta\, d\theta \wedge d\phi \tag{1.221}$$

on S^2, where we use spherical polar coordinates (θ, ϕ) in a patch (excluding the north and south poles, as discussed in section 1.3.2). This is another example of a form that exists everywhere, but which cannot be written globally as a the exterior derivative of a globally-defined 1-form. Obviously, we could write it locally as $\omega = d\nu$, where

$$\nu = -\cos\theta\, d\phi\,, \tag{1.222}$$

but this is singular at $\theta = 0$ and $\theta = \pi$, since at these points (the north and south poles) the 1-form $d\phi$ is ill-defined, since the cicle parameterised by ϕ has shrunk to zero radius at the poles. Note, however, that because $d^2 = 0$ when applied to *any* p-form, we can always add df to ν, where f is any function, and the exterior derivative of the modified ν will again give ω. Thus, we may define the two 1-forms

$$\nu_\pm = \nu \pm d\phi = (-\cos\theta \pm 1)d\phi\,. \tag{1.223}$$

These are well-defined within the patches called $U_\pm$ in section 1.3.2 respectively. Thus ν_+ is well-defined at the north pole, $\theta = 0$, since the coefficient of $d\phi$ vanishes there. However, it is ill-defined at the south pole, $\theta = \pi$, because the coefficient of $d\phi$ does not vanish there. Thus ν_+ is well-defined everywhere in the patch U_+. The situation for ν_- is similar, except that it is well-defined everywhere in U_- (the sphere minus the north pole).

With these preliminaries, we can now see what happens if we apply (or misapply) Stokes' theorem. First, the naive misapplication: If we just say $\omega = d\nu$, and mistakenly assume $\nu = -\cos\theta\, d\phi$ is globally-defined on S^2 we get

$$4\pi = \int_{S^2} \sin\theta\, d\theta \wedge d\phi = \int_{S^2} \omega = \int_{S^2} d\nu = \int_{\partial S^2} \nu = 0\,, \tag{1.224}$$

where in the last step we have used that S^2 has no boundary. Now, let's see how we can instead use Stokes' theorem correctly, by being careful about where the various 1-forms are well-defined. To do this, introduce the notation $H_\pm$ to denote the northern and southern hemispheres of S^2. Now we can write

$$\int_{S^2} \omega \;\; = \;\; \int_{H_+} \omega + \int_{H_-} \omega = \int_{H_+} d\nu_+ + \int_{H_-} d\nu_-$$

$$
\begin{aligned}
&= \int_{\partial H_+} \nu_+ + \int_{\partial H_-} \nu_- = \int_{S^1} \nu_+ + \int_{(-S^1)} \nu_- \\
&= \int_{S^1} d\phi + \int_{(-S^1)} (-d\phi) = 2\pi + 2\pi = 4\pi\,. \qquad (1.225)
\end{aligned}
$$

Here, we have split the volume integral over S^2 into the sum over the two hemispheres, and in each case we have replaced the volume-form ω by its expression as the exterior derivative of a 1-form that is globally-defined within that hemisphere. Now, we we apply Stokes' theorem, we convert the volume integrals over hemispheres into integrals around their boundaries (i.e. the equatorial circle). We must be careful about the orientations of the circles; we have that ∂H_+ is the positively-oriented equatorial circle, but ∂H_- has the opposite orientation. Thus, when we put the two contributions together, we correctly recover the 2-dimensional "volume" of the unit S^2.

1.13 The Levi-Civita Tensor and Hodge Dualisation

1.13.1 The Levi-Civita Tensor

The totally-antisymmetric tensor ε_{ijk} in 3-dimensional Cartesian tensor calculus is a familiar object. It is defined by saying that ε_{ijk} is $+1$, -1 or 0 depending on whether ijk is an even permutation of 123, an odd permutation, or no permutation at all (such as 112). We already introduced an analogous n-dimensional totally-antisymmetric object $\varepsilon^{i_1 \cdots i_n}$ in equation (1.215). However, we must be careful; this object is *not* a tensor under general coordinate transformations.

Let us first of all define $\varepsilon_{i_1 \cdots i_n}$ with downstairs indices. We shall say

$$
\varepsilon_{i_1 \cdots i_n} = \pm 1, 0\,, \qquad (1.226)
$$

where we have $+1$ if $\{i_1 \ldots, i_n\}$ is an even permutation of the numerically-ordered index values $\{1, \ldots, n\}$, we have -1 if it is an odd permutation, and we have 0 if it is no permutation at all. We define $\varepsilon_{i_1 \cdots i_n}$ to have these values in *all* coordinate frames, which means that, *by definition*, we have

$$
\varepsilon'_{i_1 \cdots i_n} = \varepsilon_{i_1 \cdots i_n}\,. \qquad (1.227)
$$

Is it a tensor? The answer is no, and we can prove this by showing that it does not transform as a tensor. Suppose it did, and so we start in a coordinate frame x^i with the components being ± 1 and 0, as defined above. We could then work out its components in a primed frame, giving

$$
\tilde{\varepsilon}_{i_1 \cdots i_n} = \frac{\partial x^{j_1}}{\partial x'^{i_1}} \cdots \frac{\partial x^{j_n}}{\partial x'^{i_n}}\, \varepsilon_{j_1 \cdots j_n}\,. \qquad (1.228)
$$

(We avoid using $\varepsilon'_{i_1\cdots i_n}$ to denote the transformed components in the primed frame because we are currently *testing* whether the transformed components, calculated assuming that $\varepsilon_{i_1\cdots i_n}$ is a tensor, agree with our *definition* of $\varepsilon'_{i_1\cdots i_n}$ given in (1.227). As we shall see, they do not agree.)

The right-hand side of (1.228) can be recognised as giving

$$\Big|\frac{\partial x}{\partial x'}\Big|\, \varepsilon_{i_1\cdots i_n}\,, \tag{1.229}$$

where $\Big|\frac{\partial x}{\partial x'}\Big|$ is the Jacobian of the transformation, i.e. the determinant of the transformation matrix $\partial x^j/\partial x'^i$. This follows from the identity that

$$M^{j_1}{}_{i_1}\, \cdots M^{j_n}{}_{i_n} \varepsilon_{j_1\cdots j_n} = \det(M)\, \varepsilon_{i_1\cdots i_n} \tag{1.230}$$

for any $n\times n$ matrix. (Check it for $n=2$, if you doubt it.) Since (1.229) is not simply equal to $\varepsilon_{i_1\cdots i_n}$, we see that $\varepsilon_{i_1\cdots i_n}$, defined to be ± 1 and 0 in *all* frames, does not transform as a tensor. Instead, it is what is called a *Tensor Density.*

A quantity with components $H_{i_1\cdots i_p}$ is said to be a tensor density of weight w if it transforms as

$$H'_{i_1\cdots i_p} = \Big|\frac{\partial x'}{\partial x}\Big|^w\, \frac{\partial x^{j_1}}{\partial x'^{i_1}} \cdots \frac{\partial x^{j_1}}{\partial x'^{i_1}}\, H_{j_1\cdots j_p}\,, \tag{1.231}$$

under general coordinate transformations. Of course ordinary tensors, for which $w=0$, are the special case of tensor densities of weight 0.

Noting that $\Big|\frac{\partial x'}{\partial x}\Big| = \Big|\frac{\partial x}{\partial x'}\Big|^{-1}$, we see from (1.229) that $\varepsilon_{i_1\cdots i_n}$ transforms as a tensor density of weight 1 under general coordinate transformations, namely

$$\varepsilon'_{i_1\cdots i_n} = \Big|\frac{\partial x'}{\partial x}\Big|\, \frac{\partial x^{j_1}}{\partial x'^{i_1}} \cdots \frac{\partial x^j_n}{\partial x'^{i_n}}\, \varepsilon_{j_1\cdots j_n}\,. \tag{1.232}$$

Furthermore, it is indeed an *invariant* tensor density, i.e. $\varepsilon'_{i_1\cdots i_n} = \varepsilon_{i_1\cdots i_n}$; it takes the same numerical values in all coordinate frames.

We can make an honest tensor by multiplying $\varepsilon_{i_1\cdots i_n}$ by a scalar density of weight -1. Such an object can be built from the metric tensor. Consider taking the determinant of the inverse metric. Since we have already introduced the notation that $g\equiv \det(g_{ij})$, it follows that we shall have $\det(g^{ij}) = 1/g$. Thus we may write

$$\frac{1}{g} = \frac{1}{n!}\, g^{i_1 j_1} \cdots g^{i_n j_n}\, \varepsilon_{i_1\cdots i_n}\, \varepsilon_{j_1\cdots j_n}\,. \tag{1.233}$$

(Again, if this is not obvious to you, check it for the case $n=2$.) Changing to a primed coordinate system, and recalling that $\varepsilon_{i_1\cdots i_n}$ is an invariant tensor density, we therefore have

$$\frac{1}{g'} \;=\; \frac{1}{n!}\, g'^{i_1 j_1} \cdots g'^{i_n j_n}\, \varepsilon_{i_1\cdots i_n}\, \varepsilon_{j_1\cdots j_n}$$

$$
\begin{aligned}
&= \frac{1}{n!}\, g^{k_1\ell_1}\cdots g^{k_n\ell_n}\, \frac{\partial x'^{i_1}}{\partial x^{k_1}}\cdots\frac{\partial x'^{i_n}}{\partial x^{k_n}}\,\frac{\partial x'^{j_1}}{\partial x^{\ell_1}}\cdots\frac{\partial x'^{j_n}}{\partial x^{\ell_n}}\,\varepsilon_{i_1\cdots i_n}\,\varepsilon_{j_1\cdots j_n}\\
&= \Big|\frac{\partial x'}{\partial x}\Big|^2\,\frac{1}{g}\,. \qquad (1.234)
\end{aligned}
$$

This shows that $g' = \Big|\frac{\partial x'}{\partial x}\Big|^{-2} g$; i.e. that g is a scalar density of weight -2. Hence $\sqrt{|g|}$ is a scalar density of weight -1, and so we may define the tensor (i.e. with weight 0)

$$
\epsilon_{i_1\cdots i_n} \equiv \sqrt{|g|}\,\varepsilon_{i_1\cdots i_n}\,. \qquad (1.235)
$$

We shall universally use the notation $\epsilon_{i_1\cdots i_n}$ for the honest tensor, and $\varepsilon_{i_1\cdots i_n}$ for the tensor density whose components are $\pm 1, 0$. The totally-antisymmetric tensor $\epsilon_{i_1\cdots i_n}$ is called the *Levi-Civita tensor.*

Some further remarks are in order at this point. First, we shall *always* define $\varepsilon_{i_1\cdots i_n}$ to be $+1$ if its indices are an even permutation of the numerically-ordered index values $1,\ldots,n$, to be -1 for an odd permutation, and 0 for no permutation. For the tensor density with upstairs indices, we *define* them to be numerically given by

$$
\varepsilon^{i_1\cdots i_n} \equiv (-1)^t\,\varepsilon_{i_1\cdots i_n}\,, \qquad (1.236)
$$

where t is the number of negative eigenvalues of the metric g_{ij}. The typical cases will be $t=0$ if we are doing Riemannian geometry, and $t=1$ in special or general relativity.

The second remark is to note that $\varepsilon^{i_1\cdots i_n}$ is *not* given by raising the indices on $\varepsilon_{i_1\cdots i_n}$ using inverse metrics. This is the *one and only* exception to the otherwise universal rule that when we use the same symbol on an object with upstairs indices and an object with downstairs indices, the former is related to the latter by raising the indices with inverse metrics.

The third remark is that $\varepsilon^{i_1\cdots i_n}$ is a tensor density of weight -1. Thus we have tensors $\epsilon_{i_1\cdots i_n}$ and $\epsilon^{i_1\cdots i_n}$ related to the corresponding tensor-densities by

$$
\epsilon_{i_1\cdots i_n} = \sqrt{|g|}\,\varepsilon_{i_1\cdots i_n}\,,\qquad \epsilon^{i_1\cdots i_n} = \frac{1}{\sqrt{|g|}}\,\varepsilon^{i_1\cdots i_n}\,. \qquad (1.237)
$$

Note that $\epsilon^{i_1\cdots i_n}$ *is* obtained by raising the indices on $\epsilon_{i_1\cdots i_n}$ with inverse metrics. This accords with our second remark above.

The fourth remark is that if the number of negative eigenvalues t of the metric is odd, then the determinant g is negative. This is why we have written $\sqrt{|g|}$ in the definitions of the totally-antisymmetric tensors $\epsilon_{i_1\cdots i_n}$ and $\epsilon^{i_1\cdots i_n}$. If we know we are in a situation where $t=0$ (or more generally $t=$even), we typically just write $\sqrt{g}$. If on the other hand we know we are in a situation where $t=1$ (or more generally $t=$odd), we typically write $\sqrt{-g}$.

There are some very important identities that are satisfied by the product of two Levi-Civita tensors. Firstly, one can establish that

$$\epsilon^{i_1\cdots i_n}\,\epsilon_{j_1\cdots j_n} = n!\,(-1)^t\,\delta^{i_1\cdots i_n}_{j_1\cdots j_n}\,, \tag{1.238}$$

where as usual t is the number of negative eigenvalues of the metric, and we have defined

$$\delta^{i_1\cdots i_p}_{j_1\cdots j_p} \equiv \delta^{[i_1}_{[j_1}\cdots\delta^{i_n]}_{j_n]}\,. \tag{1.239}$$

Note that for any antisymmetric tensor $A_{i_1\cdots i_p}$ we have

$$A_{i_1\cdots i_p}\,\delta^{i_1\cdots i_p}_{j_1\cdots j_p} = A_{j_1\cdots j_p}\,. \tag{1.240}$$

It is quite easy to prove (1.238) by enumerating the possible sets of choices for the index values on the left-hand side and on the right-hand side, and verifying that the two expressions agree. Of course one need not verify every single possible set of index assignments, since both the left-hand side and the right-hand side are manifestly totally antisymmetric in the i indices, and in the j indices. In fact this means one really only has to check one case, which could be, for example, $\{i_1,\ldots,i_n\} = \{1,\ldots,n\}$ and $\{j_1,\ldots,j_n\} = \{1,\ldots,n\}$. With a little thought, it can be seen that once the two sides are shown to agree for this set of index choices, they *must* agree for any possible set of index choices.

It is also useful to record the expression one gets if one contracts p of the indices on a pair of Levi-Civita tensors. The answer is

$$\epsilon^{i_1\cdots i_q k_1\cdots k_p}\,\epsilon_{j_1\cdots j_q k_1\cdots k_p} = p!\,q!\,(-1)^t\,\delta^{i_1\cdots i_q}_{j_1\cdots j_q}\,, \tag{1.241}$$

where we have defined $q \equiv n-p$ in n dimensions. The proof is again just a matter of enumerating inequivalent special cases, and checking the equality of the two sides of the equation for each such case. Again, if one spends enough time thinking about it, one eventually sees that it is almost trivially obvious. Note that (1.238) is just the special case of (1.241) when $p=0$.

As an example, in three dimensions with positive-definite metric signature, we have

$$\begin{aligned}
\epsilon^{ijk}\,\epsilon_{\ell mn} &= 6\delta^{ijk}_{\ell mn} = \delta^i_\ell\delta^j_m\delta^k_n + \delta^j_\ell\delta^k_m\delta^i_n + \delta^k_\ell\delta^i_m\delta^j_n - \delta^i_\ell\delta^k_m\delta^j_n - \delta^j_\ell\delta^i_m\delta^k_n - \delta^k_\ell\delta^j_m\delta^i_n\,,\\
\epsilon^{ijm}\,\epsilon_{k\ell m} &= 2\delta^{ij}_{k\ell} = \delta^i_k\delta^j_\ell - \delta^j_k\delta^i_\ell\,.
\end{aligned} \tag{1.242}$$

These, or at least the second identity, should be very familiar from Cartesian tensor analysis.

1.13.2 The Hodge dual

Suppose we have a p-form ω in n dimensions. It is easy to count the number N_p of independent components $\omega_{i_1\cdots i_p}$ in a general such p-form: the antisymmetry implies that the answer is

$$N_p = \frac{n!}{p!\,(n-p)!}\,. \tag{1.243}$$

For example, for a 0-form we have $N_0 = 1$, and for a 1-form we have $N_1 = n$. These are exactly what one expects for a scalar and a co-vector. For a 2-form we have $N_2 = \frac{1}{2}n(n-1)$, which again is exactly what one expects for a 2-index antisymmetric tensor (it is just like counting the independent components of a general $n \times n$ antisymmetric matrix).

It will be noticed from (1.243) that we have

$$N_p = N_{n-p}\,, \tag{1.244}$$

i.e. the number of independent components of a p form is the same as the number of independent components of an $(n-p)$-form in n dimensions. This suggests the possibility that there could exist a 1-1 mapping between p-forms and $(n-p)$-forms, and indeed precisely such a mapping exists. It is called *Hodge Duality*, and it is implemented by means of the Levi-Civita tensor.

Suppose a p-form ω is expanded in a coordinate basis in the usual way, as

$$\omega = \frac{1}{p!}\,\omega_{i_1\cdots i_p}\,dx^{i_1}\wedge\cdots\wedge dx^{i_p}\,. \tag{1.245}$$

We can define a *Hodge dual* basis for $q = n-p$ forms, as

$$*(dx^{i_1}\wedge\cdots\wedge dx^{i_p}) = \frac{1}{q!}\,\epsilon_{j_1\cdots j_q}{}^{i_1\cdots i_p}\,dx^{j_1}\wedge\cdots dx^{j_q}\,. \tag{1.246}$$

We can then read off the Hodge dual of ω, namely

$$*\omega = \frac{1}{p!\,q!}\,\epsilon_{j_1\cdots j_q}{}^{i_1\cdots i_p}\,\omega_{i_1\cdots i_p}\,dx^{j_1}\wedge\cdots dx^{j_q}\,. \tag{1.247}$$

Comparing with the standard definition of a q-form, we can therefore read off the components of the q-form $*\omega$, whose expansion is

$$*\omega = \frac{1}{q!}\,(*\omega)_{j_1\cdots j_q}\,dx^{j_1}\wedge\cdots\wedge dx^{j_q}\,. \tag{1.248}$$

Thus from (1.247) we read off

$$(*\omega)_{j_1\cdots j_q} = \frac{1}{p!}\,\epsilon_{j_1\cdots j_q}{}^{i_1\cdots i_p}\,\omega_{i_1\cdots i_p}\,. \tag{1.249}$$

Equation (1.249) gives the mapping from the p-form ω to its Hodge dual, the $q = n - p$ form $*\omega$. It was said earlier that this is a 1-1 mapping, and so we must be able to invert it. This is easily done, by making use of the identity (1.241) for the contraction of two Levi-Civita tensors on some of their indices. Thus, taking the Hodge dual of the Hodge dual of ω, making use of the basic defining equation (1.249), we shall have

$$\begin{aligned}
(**\omega)_{i_1\cdots i_p} &= \frac{1}{p!\,q!}\,\epsilon_{i_1\cdots i_p}{}^{j_1\cdots j_q}\,\epsilon_{j_1\cdots j_q}{}^{k_1\cdots k_p}\,\omega_{k_1\cdots k_p} \\
&= \frac{(-1)^{pq}}{p!\,q!}\,\epsilon_{i_1\cdots i_p}{}^{j_1\cdots j_q}\,\epsilon^{k_1\cdots k_p}{}_{j_1\cdots j_q}\,\omega_{k_1\cdots k_p} \\
&= \frac{(-1)^{pq+t}}{p!\,q!}\,p!\,q!\,\delta^{k_1\cdots k_p}_{i_1\cdots i_p}\,\omega_{k_1\cdots k_p} \\
&= (-1)^{pq+t}\,\omega_{i_1\cdots i_p}\,.
\end{aligned} \tag{1.250}$$

In getting to the second line, the shifting of the block of q indices $(j_1 \cdots j_q)$ through the block of p indices $(k_1 \cdots k_p)$ on the second Levi-Civita tensor has given rise to the $(-1)^{pq}$ factor, since each interchange of an index pair produces a minus sign. In getting to the third line, we have used the identity (1.241). In getting to the fourth line, we have used the basic property (1.240) of the multi-index Kronecker delta tensor. The upshot, therefore, is that applying the Hodge dual operation twice to a p-form ω in n dimensions, we get

$$**\omega = (-1)^{pq+t}\,\omega\,, \tag{1.251}$$

where $q = n - p$, and where t is the number of time directions (i.e. the number of negative eigenvalues of the metric tensor).

In cases where $pq+t$ is even, we shall have that $**\omega = \omega$, which means that the operator $*$ itself has eigenvalues ± 1. If the dimension n is even, say $n = 2m$, an m-form ω is mapped into another m-form by the Hodge $*$ operator, and so if $m^2 + t$ is even, we can make $\pm$ eigenstates under $*$, defined by

$$\omega_\pm = \tfrac{1}{2}(\omega \pm *\omega)\,. \tag{1.252}$$

these have the property that

$$*\omega_\pm = \pm\omega_\pm\,, \tag{1.253}$$

and they are known as self-dual or anti-self-dual forms respectively. This possibility therefore arises in Riemannian geometry (i.e. $t = 0$) in dimensions $n = 4, 8, 12, \ldots$. In pseudo-Riemannian geometry with a single time dimension (i.e. $t = 1$), (anti)-self-duality is instead possible in dimensions $n = 2, 6, 10, \ldots$.

The Hodge dual provides a nice way of taking the inner product of two p-forms. Suppose we have two p-forms, A and B in an n-dimensional manifold M. Defining $q = n - p$ as usual, we shall have

$$\begin{aligned}
*A\wedge B &= \frac{1}{(p!)^2\,q!}\,\epsilon_{i_1\cdots i_q}{}^{j_1\cdots j_p}\,A_{j_1\cdots j_p}\,B_{k_1\cdots k_p}\,dx^{i_1}\wedge\cdots\wedge dx^{i_q}\wedge dx^{k_1}\wedge\cdots\wedge dx^{k_p}\\
&= \frac{(-1)^t}{(p!)^2\,q!}\,\epsilon_{i_1\cdots i_q}{}^{j_1\cdots j_p}\,A_{j_1\cdots j_p}\,B_{k_1\cdots k_p}\,\varepsilon^{i_1\cdots i_q k_1\cdots k_p}\,dx^1\wedge dx^2\wedge\cdots\wedge dx^n\\
&= \frac{(-1)^t}{(p!)^2\,q!}\,\epsilon_{i_1\cdots i_q}{}^{j_1\cdots j_p}\,A_{j_1\cdots j_p}\,B_{k_1\cdots k_p}\,\epsilon^{i_1\cdots i_q k_1\cdots k_p}\,\sqrt{|g|}\,dx^1\wedge dx^2\wedge\cdots\wedge dx^n\\
&= \frac{1}{p!}\,A_{j_1\cdots j_p}\,B_{k_1\cdots k_p}\,\delta^{k_1\cdots k_q}_{i_1\cdots i_q}\,\sqrt{|g|}\,dx^1\wedge dx^2\wedge\cdots\wedge dx^n\\
&= \frac{1}{p!}\,A_{i_1\cdots i_p}\,B^{i_1\cdots i_p}\,\sqrt{|g|}\,dx^1\wedge dx^2\wedge\cdots\wedge dx^n\,. \qquad (1.254)
\end{aligned}$$

Thus we can write

$$*A\wedge B = \frac{1}{p!}\,A_{i_1\cdots i_p}\,B^{i_1\cdots i_p}\,*\mathbb{1}\,, \qquad (1.255)$$

where

$$*\mathbb{1} = \frac{1}{n!}\,\epsilon_{i_1\cdots i_n}\,dx^{i_1}\wedge\cdots\wedge dx^{i_n} = \sqrt{|g|}\,dx^1\wedge dx^2\wedge\cdots\wedge dx^n\,. \qquad (1.256)$$

Note that $*\mathbb{1}$, which is the Hodge dual of the constant 1, calculated using the standard rule (1.246) applied to a 0-form, is the *volume form.* For example, in Cartesian coordinates on Euclidean 2-space, where the metric is just $ds^2 = dx^2 + dy^2$, we would have $*\mathbb{1} = dx\wedge dy$, whilst in polar coordinates, where the metric is $ds^2 = dr^2 + r^2 d\theta^2$, we would have $*\mathbb{1} = rdr\wedge d\theta$. Thus equation (1.255) shows that $*A\wedge B$ is equal to $1/p!$ times the volume form, multiplied by the inner product

$$|A\cdot B| \equiv A_{i_1\cdots i_p}\,B^{i_1\cdots i_p} \qquad (1.257)$$

of the two p-forms A and B. The inner product is manifestly symmetric under the exchange of A and B, and so we have

$$*A\wedge B = *B\wedge A = \frac{1}{p!}\,|A\cdot B|\,*\mathbb{1}\,. \qquad (1.258)$$

Of course if the metric has all positive eigenvalues (i.e. $t=0$), then the inner product is positive semi-definite, in the sense that

$$|A\cdot A| \ge 0\,, \qquad (1.259)$$

with equality if and only if $A=0$.

1.14 The δ Operator and the Laplacian

1.14.1 The adjoint operator δ; covariant divergence

Let A and B be two p-forms. We may define the quantity (A, B) by

$$(A, B) \equiv \int_M {*A} \wedge B\,, \tag{1.260}$$

where, by (1.258), the integrand is the n-form proportional to the volume form times the inner product of A and B. Like the unintegrated inner product, it is the case that if the metric has all positive eigenvalues, then (A, B) is positive semi-definite, in the sense that

$$(A, A) \geq 0\,, \tag{1.261}$$

with equality if and only if A vanishes everywhere in M. Note that from (1.258) we also have that

$$(A, B) = (B, A)\,. \tag{1.262}$$

Suppose now we have a p-form ω and $(p-1)$-form ν. Using the definition (1.260) we may form the quantity $(\omega, d\nu)$. Let us assume that the n-manifold M has no boundary. By using Stokes' theorem, we can perform the following manipulation:

$$\begin{aligned}
(\omega, d\nu) &= \int_M {*\omega} \wedge d\nu = (-1)^q \int_M d({*\omega} \wedge \nu) - (-1)^q \int_M d{*\omega} \wedge \nu \\
&= (-1)^q \int_{\partial M} {*\omega} \wedge \nu - (-1)^q \int_M d{*\omega} \wedge \nu \\
&= (-1)^{q+1} \int_M d{*\omega} \wedge \nu = (-1)^{pq+p+t} \int_M {*({*d{*\omega}})} \wedge \nu \\
&= (-1)^{pq+p+t}\, ({*d{*\omega}}, \nu)\,,
\end{aligned} \tag{1.263}$$

where as usual we have defined $q \equiv n - p$. Thus it is natural to define the *adjoint* of the exterior derivative, which is called δ, to be such that for any p-form ω and any $(p-1)$-form ν, we shall have

$$(\omega, d\nu) = (\delta\omega, \nu)\,, \tag{1.264}$$

with

$$\delta \equiv (-1)^{pq+p+t}\, {*d*} = (-1)^{np+t}\, {*d*}\,. \tag{1.265}$$

Of course from (1.262) we shall also have

$$(\nu, \delta\omega) = (d\nu, \omega)\,. \tag{1.266}$$

Note that using (1.265) and (1.251) we can immediately see that δ has the property that

$$\delta^2 = 0 \tag{1.267}$$

when acting on any p-form. Note that δ maps a p-form to a $(p-1)$-form.

We know that d maps a p-form ω to a $(p+1)$-form, and that the Hodge dual $*$ maps a p-form to an $(n-p)$-form in n dimensions. It is easy to see, therefore, that the operator $*d*$ applied to a p-form gives a $(p-1)$-form. What is the object $*d*\omega$? It is actually related to something very simple, namely the divergence of ω, with components $\nabla^k\omega_{ki_1\cdots i_{p-1}}$. To show this is straightforward, although a little lengthy. For the sake of completeness, we shall give the derivation here. Those steps in the argument that are analogous to ones that have already been spelt out in previous derivations will be performed this time without further comment. We shall have

$$\begin{aligned}
\omega &= \frac{1}{p!}\,\omega_{i_1\cdots i_p}\,dx^{i_1}\wedge\cdots\wedge dx^{i_p}\,,\\
*\omega &= \frac{1}{p!\,q!}\,\omega_{i_1\cdots i_p}\,\epsilon_{j_1\cdots j_q}{}^{i_1\cdots i_p}\,dx^{j_1}\wedge\cdots\wedge dx^{j_q}\,,\\
d{*\omega} &= \frac{1}{p!\,q!}\,\partial_k(\omega_{i_1\cdots i_p}\,\epsilon_{j_1\cdots j_q}{}^{i_1\cdots i_p})\,dx^k\wedge dx^{j_1}\wedge\cdots\wedge dx^{j_q}\,,\\
d{\omega} &= \frac{1}{p!\,q!\,(p-1)!}\,\partial_k(\omega_{i_1\cdots i_p}\,\epsilon_{j_1\cdots j_q}{}^{i_1\cdots i_p})\,\epsilon_{\ell_1\cdots\ell_{p-1}}{}^{kj_1\cdots j_q}\,dx^{\ell_1}\wedge\cdots\wedge dx^{\ell_{p-1}}\\
&= \frac{1}{p!\,q!\,(p-1)!}\,\partial_k(\omega^{i_1\cdots i_p}\,\epsilon_{j_1\cdots j_q i_1\cdots i_p})\,\epsilon_{\ell_1\cdots\ell_{p-1}}{}^{kj_1\cdots j_q}\,dx^{\ell_1}\wedge\cdots\wedge dx^{\ell_{p-1}}\\
&= \frac{(-1)^{pq}}{p!\,q!\,(p-1)!}\,\partial_k(\omega^{i_1\cdots i_p}\,\epsilon_{i_1\cdots i_p j_1\cdots j_q})\,\epsilon_{\ell_1\cdots\ell_{p-1}}{}^{kj_1\cdots j_q}\,dx^{\ell_1}\wedge\cdots\wedge dx^{\ell_{p-1}}\\
&= \frac{(-1)^{pq}}{p!\,q!\,(p-1)!}\,\partial_k(\omega^{i_1\cdots i_p}\,\sqrt{|g|}\,\varepsilon_{i_1\cdots i_p j_1\cdots j_q})\,\epsilon_{\ell_1\cdots\ell_{p-1}}{}^{kj_1\cdots j_q}\,dx^{\ell_1}\wedge\cdots\wedge dx^{\ell_{p-1}}\\
&= \frac{(-1)^{pq}}{p!\,q!\,(p-1)!}\,\partial_k(\omega^{i_1\cdots i_p}\,\sqrt{|g|})\,\varepsilon_{i_1\cdots i_p j_1\cdots j_q}\,\epsilon_{\ell_1\cdots\ell_{p-1}}{}^{kj_1\cdots j_q}\,dx^{\ell_1}\wedge\cdots\wedge dx^{\ell_{p-1}}\\
&= \frac{(-1)^{pq}}{p!\,q!\,(p-1)!}\,\frac{1}{\sqrt{|g|}}\,\partial_k(\omega^{i_1\cdots i_p}\,\sqrt{|g|})\,\epsilon_{i_1\cdots i_p j_1\cdots j_q}\,\epsilon_{\ell_1\cdots\ell_{p-1}}{}^{kj_1\cdots j_q}\,dx^{\ell_1}\wedge\cdots\wedge dx^{\ell_{p-1}}\,,
\end{aligned}\tag{1.268}$$

where the only new type of manipulation so far is to replace the Levi-Civita tensor $\epsilon_{i_1\cdots i_n}$ by $\sqrt{|g|}\,\varepsilon_{i_1\cdots i_n}$, take the Levi-Civita tensor density $\varepsilon_{i_1\cdots i_n}$ outside the partial derivative (which can be done since it has constant components ±1 and 0), and then restore it to the Levi-Civita tensor by dividing out by $\sqrt{|g|}$ once it is outside the partial derivative. It is helpful at this point to define the object

$$Y_k{}^{i_1\cdots i_p} \equiv \frac{1}{\sqrt{|g|}}\,\partial_k(\sqrt{|g|}\,\omega^{i_1\cdots i_p})\,,\tag{1.269}$$

which we will shortly be able to turn into something recognisable. Continuing to the next step that follows on from the last line in (1.268), we can write

$$*d{*\omega} = \frac{(-1)^{pq}}{p!\,q!\,(p-1)!}\,Y^k{}_{i_1\cdots i_p}\,\epsilon^{i_1\cdots i_p}{}_{j_1\cdots j_q}\,\epsilon_{\ell_1\cdots\ell_{p-1}\,k}{}^{j_1\cdots j_q}\,dx^{\ell_1}\wedge\cdots dx^{\ell_{p-1}}$$

$$
\begin{aligned}
&= \frac{(-1)^{pq+t}}{(p-1)!}\, Y^k{}_{i_1\cdots i_p}\, \delta^{i_1\cdots i_p}_{\ell_1\cdots\ell_{p-1}\,k}\, dx^{\ell_1}\wedge\cdots dx^{\ell_{p-1}} \\
&= \frac{(-1)^{pq+t}}{(p-1)!}\, Y^k{}_{\ell_1\cdots\ell_{p-1}\,k}\, dx^{\ell_1}\wedge\cdots dx^{\ell_{p-1}} \\
&= \frac{(-1)^{pq+p+1+t}}{(p-1)!}\, Y^k{}_{k\ell_1\cdots\ell_{p-1}}\, dx^{\ell_1}\wedge\cdots dx^{\ell_{p-1}}\,. \qquad (1.270)
\end{aligned}
$$

Now, we have

$$
\begin{aligned}
Y^k{}_{k\ell_1\cdots\ell_{p-1}} &= Y_k{}^{km_1\cdots m_{p-1}}\, g_{\ell_1 m_1}\cdots g_{\ell_{p-1}\,m_{p-1}} \\
&= \frac{1}{\sqrt{|g|}}\,\partial_k\big(\sqrt{|g|}\,\omega^{km_1\cdots m_{p-1}}\big)\, g_{\ell_1 m_1}\cdots g_{\ell_{p-1}\,m_{p-1}} \\
&= (\nabla_k\,\omega^{km_1\cdots m_{p-1}})\, g_{\ell_1 m_1}\cdots g_{\ell_{p-1}\,m_{p-1}} \\
&= \nabla^k\omega_{k\ell_1\cdots\ell_{p-1}}\,, \qquad (1.271)
\end{aligned}
$$

where the step of passing to the third line involves using results derived in section 1.9, and the symmetry of $\Gamma^i{}_{jk}$ in its two lower indices. (A special case, for a 2-index antisymmetric tensor, was on Problem Sheet 2.)

Finally, we are approaching the bottom line, namely that we have found

$$
{*d{*\omega}} = \frac{(-1)^{pq+p+t+1}}{(p-1)!}\,\nabla^k\omega_{ki_1\cdots i_{p-1}}\, dx^{i_1}\wedge\cdots\wedge dx^{i_{p-1}}\,. \qquad (1.272)
$$

In other words, we have shown that the components of the $(p-1)$-form ${*d{*\omega}}$ are given by[7]

$$
({*d{*\omega}})_{i_1\cdots i_{p-1}} = (-1)^{pq+p+t+1}\,\nabla^k\omega_{ki_1\cdots i_{p-1}}\,. \qquad (1.273)
$$

Comparing this with (1.265), we see that for any p-form ω, we shall have

$$
(\delta\omega)_{i_1\cdots i_{p-1}} = -\nabla^k\omega_{ki_1\cdots i_p}\,. \qquad (1.274)
$$

[7]Note that although this derivation may have seemed like a bit of a long song and dance, much of this was because, for pedagodgic reasons, all the logical steps have been spelt out. Additionally, we presented rather carefully the mechanism by which the partial derivative turned into a covariant divergence. We could have short-circuited quite a few of those steps by making the following argument: We know that the exterior derivative d maps a p-form to a $(p+1)$-form, and we know that the Hodge $*$ maps a p-form to an $(n-p)$-form. Therefore we *know* that ${*d{*\omega}}$ must be a $(p-1)$-form, and therefore that it must be an honest tensorial object. Thus, as soon as we saw the ∂_k appear in the expression for $d{*\omega}$, we know on the grounds of covariance, that we *must* be able to replace the partial derivative by a covariant one, since the answer must be covariant, so what else could it be? All we are doing by replacing ∂_k by ∇_k is making a "hidden" non-manifest covariance into an explicit manifest covariance. If we allow ourselves to make that replacement, we more quickly end up at the same conclusion.

1.14.2 The Laplacian

We have already met the covariant Laplacian operator that acts on scalars. Here, we give the generalisation to a Laplacian operator that acts on p-forms of any rank. It is defined by

$$\Delta \equiv d\delta + \delta d\,. \tag{1.275}$$

Since d maps p-forms to $(p+1)$-forms, and δ maps p-forms to $(p-1)$-forms, we see that each of the two terms in Δ maps a p-form back into a p-form, and thus so does Δ itself.

If we apply Δ to a scalar f, then, noting that $\delta f \equiv 0$ (since δf would be a (-1)-form, which doesn't exist), we shall have

$$\Delta f = \delta df = -\nabla^i \nabla_i f\,. \tag{1.276}$$

Thus when acting on scalars, Δ is the *negative* of what one commonly calls the Laplacian in more elementary contexts. It is actually rather natural to include the minus sign in the definition, because $\Delta = -\nabla^i\nabla_i$ is then a positive operator when acting on scalars, in the case that the metric has all positive eigenvalues.

In fact, more generally, we can see that Δ defined by (1.275) is a positive operator when acting on any p-form, in the case that the metric has all positive eigenvalues (i.e. $t = 0$). To see this, let ω be an arbitrary p-form, and assume that M is a compact n-manifold equipped with a positive-definite metric. Then we shall have

$$(\omega \Delta\omega) = (\omega, d\delta\omega) + (\omega, \delta d\omega) = (d\omega, d\omega) + (\delta\omega, \delta\omega)\,. \tag{1.277}$$

As noted previously, we have $(A, A) \ge 0$, with equality if and only if $A = 0$, and so we conclude that

$$(\omega, \Delta\omega) \ge 0\,, \tag{1.278}$$

with equality if and only if $\Delta\omega = 0$. A p-form ω that satisfies $\Delta\omega = 0$ is called an *harmonic* p-form. Furthermore, (1.277) shows that $\Delta\omega = 0$ if and only if

$$d\omega = 0\,, \qquad \delta\omega = 0\,. \tag{1.279}$$

We already met the notion of a *closed* p-form ω, as being one for which $d\omega = 0$. We can also introduce the notion of a *co-closed* p-form ω, as being one for which $\delta\omega = 0$. Thus we have seen that on a manifold without boundary, equipped with a positive-definite metric, a p-form is harmonic if and only if it is both closed and co-closed.

We have already seen that when acting on scalars f (i.e. 0-forms), the Laplacian operator is given by

$$\Delta f = -\Box f\,, \tag{1.280}$$

where we define

$$\Box \equiv \nabla^i \nabla_i \,. \tag{1.281}$$

It is straightforward to evaluate the Laplacian acting on forms of higher degree, by making use of the expressions (1.199) and (1.274) for the components of $d\omega$ and $\delta\omega$. For example, acting on a 1-form V, and on a 2-form ω, one finds

$$\begin{aligned} (\Delta V)_i &= -\Box V_i + R_{ij}\, V^j \,, \\ (\Delta\omega)_{ij} &= -\Box \omega_{ij} - 2R_{ikj\ell}\, \omega^{k\ell} + R_{ik}\, \omega^k{}_j + R_{jk}\, \omega_i{}^k \,. \end{aligned} \tag{1.282}$$

Note that the curvatures arise because terms in the expression for Δ give rise to commutators of covariant derivatives.

1.15 Spin connection and curvature 2-forms

When we introduced the notations of the covariant derivative, in section 1.9, and the Riemann tensor, in section 1.10, this was done in the framework of a choice of coordinate basis. We have already discussed the idea of using a non-coordinate basis for the tangent and co-tangent frames, and here we return to this, in order to introduce a different way of defining the connection and curvature. It is, in the end, equivalent to the coordinate-basis description, but it has various advantages, including (relative) computational simplicity.

We begin by "taking the square root" of the metric g_{ij}, by introducing a vielbein, which is a basis of 1-forms $e^a = e^a_i\, dx^i$, with the components e^a_i having the property

$$g_{ij} = \eta_{ab}\, e^a_i\, e^b_j \,. \tag{1.283}$$

Here the indices a are local-Lorentz indices, or tangent-space indices, and η_{ab} is a "flat" metric, with constant components. The language of "local-Lorentz" indices stems from the situation when the metric g_{ij} has Minkowskian signature (which is $(-,+,+,\ldots,+)$ in sensible conventions). The signature of η_{ab} must be the same as that of g_{ij}, so if we are working in general relativity with Minkowskian signature we will have

$$\eta_{ab} = \mathrm{diag}\,(-1,1,1,\ldots,1) \,. \tag{1.284}$$

If, on the other hand, we are working in a space with Euclidean signature $(+,+,\ldots,+)$, then η_{ab} will just equal the Kronecker delta, $\eta_{ab} = \delta_{ab}$, or in other words

$$\eta_{ab} = \mathrm{diag}\,(1,1,1,\ldots,1) \,. \tag{1.285}$$

Of course the choice of vielbeins e^a as the square root of the metric in (1.283) is to some extent arbitrary. Specifically, we could, given a particular choice of vielbein e^a, perform an orthogonal-type transformation to get another equally-valid vielbein e'^a, given by

$$e'^a = \Lambda^a{}_b\, e^b\ , \tag{1.286}$$

where $\Lambda^a{}_b$ is a matrix satisfying the (pseudo)orthogonality condition

$$\eta_{ab}\, \Lambda^a{}_c\, \Lambda^b{}_d = \eta_{cd}\ . \tag{1.287}$$

Note that $\Lambda^a{}_b$ can be coordinate dependent. If the n-dimensional manifold has a Euclidean-signature metric then $\eta = \mathbb{1}$ and (1.287) is literally the orthogonality condition $\Lambda^T\, \Lambda = \mathbb{1}$. Thus in this case the arbitrariness in the choice of vielbein is precisely the freedom to make local $O(n)$ rotations in the tangent space, where $O(n)$ denotes the group of $n \times n$ orthogonal matrices. If the metric signature is Minkowskian, then instead (1.287) is the condition for Λ to be an $O(1, n-1)$ matrix; in other words, one then has the freedom to perform local Lorentz transformations in the tangent space. We shall typically use the words "local Lorentz transformation" regardless of whether we are working with metrics of Minkowskian or Euclidean signature.

The vielbein e^a is a local-Lorentz vector-valued 1-form. That is, it is a 1-form carrying in addition a local-Lorentz vector index. It transforms covariantly under local-Lorentz transformations, i.e. in the manner given in equation (1.286). It is natural, at this point, to introduce the general notion of local-Lorentz tensor-valued p-forms. Thus we could consider the object $V^{a_1\cdots a_r}{}_{b_1\cdots b_s}$, which is a p-form carrying in addition r upstairs local-Lorentz indices and s downstairs local-Lorentz indices. By definition, under local-Lorentz transformations, it transforms as

$$V^{a_1\cdots a_r}{}_{b_1\cdots b_s} \longrightarrow V'^{a_1\cdots a_r}{}_{b_1\cdots b_s} = \Lambda^{a_1}{}_{c_1} \cdots \Lambda^{a_r}{}_{c_r}\, \Lambda_{b_1}{}^{d_1} \cdots \Lambda_{b_s}{}^{d_s}\, V^{c_1\cdots c_r}{}_{d_1\cdots d_s}\ , \tag{1.288}$$

where we define

$$\Lambda_a{}^b \equiv \eta_{ac}\, \eta^{bd}\, \Lambda^c{}_d\,. \tag{1.289}$$

The transformation in (1.288) is exactly like an old-fashioned Lorentz transformation of a Lorentz tensor $V^{a_1\cdots a_r}{}_{b_1\cdots b_s}$, except that here $\Lambda^a{}_b$ can be position-dependent, and also $V^{a_1\cdots a_r}{}_{b_1\cdots b_s}$ is also a p-form.

What happens if we take the exterior derivative of the local-Lorentz tensor-valued p-form $V^{a_1\cdots a_r}{}_{b_1\cdots b_s}$? Obviously, for reasons that are now very familiar, we do not get a local-Lorentz tensor-valued $(p+1)$-form, because when we test its transformation under the appropriate analogue of (1.288), we run into trouble from the exterior derivative landing on the local-Lorentz transformation matrix. To illustrate the point, while avoiding the clutter of large numbers of indices, consider the case of a local-Lorentz vector-valued p-form, V^a. It transforms as

$$V'^a = \Lambda^a{}_b\, V^b\,. \tag{1.290}$$

Now check the transformation of dV^a:

$$dV'^a = d(\Lambda^a{}_b\, V^b) = \Lambda^a{}_b\, dV^b + d\Lambda^a{}_b \wedge V^b\,. \tag{1.291}$$

The second term has spoiled the covariant transformation law.

The remedy, as in our previous discussion of the covariant derivative, is to introduce a modified "covariant exterior derivative." Note that the covariance we are speaking of here is local-Lorentz covariance. To do this, we introduce the spin connection, or connection 1-forms, $\omega^a{}_b = \omega^a{}_{bi}\, dx^i$, and the torsion 2-forms $T^a = \frac{1}{2} T^a{}_{ij}\, dx^i \wedge dx^j$, by defining

$$T^a = de^a + \omega^a{}_b \wedge e^b\ . \tag{1.292}$$

We shall require *by definition* that T^a transform covariantly as a local-Lorentz vector-valued 2-form, and we shall deduce the necessary transformation rule of $\omega^a{}_b$ accordingly. Thus we shall have

$$\begin{aligned} T'^a &= \Lambda^a{}_b\, T^b = \Lambda^a{}_b\, de^b + \Lambda^a{}_b\, \omega^b{}_c \wedge e^c \\ &= de'^a + \omega'^a{}_b \wedge e'^b = d(\Lambda^a{}_b\, e^b) + \omega'^a{}_b \wedge \Lambda^b{}_c\, e^c \\ &= \Lambda^a{}_b\, de^b + d\Lambda^a{}_b \wedge e^b + \omega'^a{}_b \wedge \Lambda^b{}_c\, e^c\,. \end{aligned} \tag{1.293}$$

Equating the first and the last lines gives, after an index relabelling,

$$\Lambda^a{}_b\, de^b + \Lambda^a{}_b\, \omega^b{}_c \wedge e^c = \Lambda^a{}_b\, de^b + d\Lambda^a{}_c \wedge e^c + \omega'^a{}_b \wedge \Lambda^b{}_c\, e^c\,, \tag{1.294}$$

from which we can read off that

$$\Lambda^a{}_b\, \omega^b{}_c = d\Lambda^a{}_c + \omega'^a{}_b\, \Lambda^b{}_c\,. \tag{1.295}$$

Multiplying by $\Lambda_d{}^c$ to remove the Λ factor on the right-hand side, we have, after a further relabelling of indices,

$$\omega'^a{}_b = \Lambda^a{}_c\, \omega^c{}_d\, \Lambda_b{}^d - d\Lambda^a{}_c\, \Lambda_b{}^c\,. \tag{1.296}$$

Noting that from (1.287) and (1.289) we have $\Lambda^a{}_c\, \Lambda_b{}^c = \delta^a_b$, which in the obvious matrix notation reads $\Lambda\, \Lambda^{-1} = 1\!\!1$, we can write (1.296) in a matrix notation as

$$\omega' = \Lambda\, \omega\, \Lambda^{-1} - d\Lambda\, \Lambda^{-1}\,. \tag{1.297}$$

Equivalently, this can be written as

$$\omega' = \Lambda\,\omega\,\Lambda^{-1} + \Lambda\,d\Lambda^{-1}\,, \tag{1.298}$$

or, back in indices,

$$\omega'^a{}_b = \Lambda^a{}_c\,\omega^c{}_d\,\Lambda_b{}^d + \Lambda^a{}_c\,d\Lambda_b{}^c\,. \tag{1.299}$$

This is the transformation rule that we shall use, telling us how the spin connection transforms under local-Lorentz transformations. As we would expect, it does not transform covariant under local-Lorentz transformations, owing to the presence of the second term. This is exactly what is needed in order to ensure that the torsion T^a *does* transform covariantly.

The notion of a Lorentz-covariant exterior derivative, which we shall call D, can now be extended to the general case of the Lorentz tensor-valued p-form $V^{a_1\cdots a_r}{}_{b_1\cdots b_s}$ that we introduced earlier. Thus we define

$$\begin{aligned} DV^{a_1\cdots a_r}{}_{b_1\cdots b_s} &\equiv dV^{a_1\cdots a_r}{}_{b_1\cdots b_s} + \omega^{a_1}{}_c\wedge V^{ca_2\cdots a_r}{}_{b_1\cdots b_s} + \cdots + \omega^{a_r}{}_c\wedge V^{a_1\cdots a_{r-1}c}{}_{b_1\cdots b_s} \\ &\quad -\omega^c{}_{b_1}\wedge V^{a_1\cdots a_r}{}_{cb_2\cdots b_s} - \cdots - \omega^c{}_{b_s}\wedge V^{a_1\cdots a_r}{}_{b_1\cdots b_{s-1}c}\,. \end{aligned} \tag{1.300}$$

The pattern here should now be very familiar; there is one spin-connection term to covariantise each of the local-Lorentz indices on $V^{a_1\cdots a_r}{}_{b_1\cdots b_s}$. It is now just a straightfoward exercise to verify that $DV^{a_1\cdots a_r}{}_{b_1\cdots b_s}$ as defined here does indeed transform covariantly under local-Lorentz transformations. In other words, we have

$$D'V'^{a_1\cdots a_r}{}_{b_1\cdots b_s} = \Lambda^{a_1}{}_{c_1}\cdots\Lambda^{a_r}{}_{c_r}\,\Lambda_{b_1}{}^{d_1}\cdots\Lambda_{b_s}{}^{d_s}\,DV^{c_1\cdots c_r}{}_{d_1\cdots d_s}\,. \tag{1.301}$$

In order to prove this, it is helpful to look just at a simple case of a Lorentz tensor-valued p-form $V^a{}_b$, in order to avoid getting bogged down in a morass of indices. It is obvious, once one has checked for $V^a{}_b$, that the proof will go in just the same way if there are more indices.

In fact, one can avoid the need for indices at all by writing the $V^a{}_b$ example in a matrix notation. We note first that

$$V'^a{}_b = \Lambda^a{}_c\,\Lambda^d{}_b\,V^c{}_d\,, \tag{1.302}$$

which translates into $V' = \Lambda V \Lambda^{-1}$ in matrix notation. Next, we rewrite $DV^a{}_b$ in matrix notation. Thus

$$\begin{aligned} DV^a{}_b &= dV^a{}_b + \omega^a{}_c\wedge V^c{}_b - \omega^c{}_b\wedge V^a{}_c \\ &= dV^a{}_b + \omega^a{}_c\wedge V^c{}_b - (-1)^p\,V^a{}_c\wedge\omega^c{}_b\,, \end{aligned} \tag{1.303}$$

and in this latter form it can be re-expressed in the obvious matrix notation as

$$DV = dV + \omega \wedge V - (-1)^p\, V \wedge \omega\,. \tag{1.304}$$

Following a few simple steps, and using (1.298), one easily shows that

$$\begin{aligned} D'V' &\equiv dV' + \omega' \wedge V' - (-1)^p V' \wedge \omega' \\ &= \Lambda (DV) \Lambda^{-1}\,, \end{aligned} \tag{1.305}$$

which establishes the covariance of the transformation.

Next, we define the curvature 2-forms $\Theta^a{}_b$, *via* the equation

$$\Theta^a{}_b = d\omega^a{}_b + \omega^a{}_c \wedge \omega^c{}_b\ . \tag{1.306}$$

It is straightforward to show, by the same techniques as we used above, that in the obvious matrix notation, in which (1.306) is written as

$$\Theta = d\omega + \omega \wedge \omega\,, \tag{1.307}$$

then Θ transforms covariantly under local-Lorentz transformations, *viz.*

$$\Theta' = \Lambda \Theta \Lambda^{-1}\,. \tag{1.308}$$

To summarise, the vielbein, spin-connection, torsion and curvature forms transform under local-Lorentz transformations as

$$\begin{aligned} e' &= \Lambda e\,, \qquad \omega' = \Lambda\, \omega\, \Lambda^{-1} + \Lambda\, d\Lambda^{-1}\ , \\ T' &= \Lambda\, T\ , \qquad \Theta' = \Lambda\, \Theta\, \Lambda^{-1}\ . \end{aligned} \tag{1.309}$$

The covariant exterior derivative D will commute nicely with the process of contracting tangent-space indices with η_{ab}, provided we require that the local-Lorentz metric η_{ab} be Lorentz-covariantly constant, $D\eta_{ab} = 0$. From (1.300), we therefore have

$$D\,\eta_{ab} \equiv d\eta_{ab} - \omega^c{}_a\, \eta_{cb} - \omega^c{}_b\, \eta_{ac} = 0\ . \tag{1.310}$$

Since we are taking the components of η_{ab} to be literally constants, it follows from this equation, which is known as the equation of *metric compatibility*, that

$$\omega_{ab} = -\omega_{ba}\ , \tag{1.311}$$

where ω_{ab} is, by definition, $\omega^a{}_b$ with the upper index lowered using η_{ab}: $\omega_{ab} \equiv \eta_{ac}\, \omega^c{}_b$. With this imposed, it is now the case that we can take covariant exterior derivatives of products, and freely move the local-Lorentz metric tensor η_{ab} through the derivative. This means that we get the same answer if we differentiate the product and then contract some indices, or if instead we contract the indices and then differentiate. This is the analogue of our requirement that $\nabla_i g_{jk} = 0$ in the previous coordinate-basis discussion of the covariant derivative.

In addition to the requirement of metric compatibiilty we usually also choose a *torsion-free* spin-connection, meaning that we demand that the torsion 2-forms T^a defined by (1.292) vanish. In fact equation (1.292), together with the metric-compatibility condition (1.311), now determine $\omega^a{}_b$ uniquely. In other words, the two conditions

$$de^a = -\omega^a{}_b \wedge e^b \ , \qquad \omega_{ab} = -\omega_{ba} \tag{1.312}$$

have a unique solution. It can be given as follows. Let us say that, as a definition of the coefficients $c_{bc}{}^a$, the exterior derivatives of the vielbeins e^a are given by

$$de^a = -\tfrac{1}{2} c_{bc}{}^a \, e^b \wedge e^c \ , \tag{1.313}$$

where the *structure functions* $c_{bc}{}^a$ are, by definition, antisymmetric in bc. Then the solution for ω_{ab} is given by

$$\omega_{ab} = \tfrac{1}{2}(c_{abc} + c_{acb} - c_{bca})\, e^c \ , \tag{1.314}$$

where $c_{abc} \equiv \eta_{cd}\, c_{ab}{}^d$. It is easy to check by direct substitution that this indeed solves the two conditions (1.312).

The procedure, then, for calculating the curvature 2-forms for a metric g_{ij} with vielebeins e^a is the following. We write down a choice of vielbein, and by taking the exterior derivative we read off the coefficients $c_{bc}{}^a$ in (1.313). Using these, we calculate the spin connection using (1.314). Then, we substitute into (1.306), to calculate the curvature 2-forms.

Each curvature 2-form $\Theta^a{}_b$ has, as its components, a tensor that is antisymmetric in two coordinate indices. This is in fact the Riemann tensor, defined by

$$\Theta^a{}_b = \tfrac{1}{2} R^a{}_{bij}\, dx^i \wedge dx^j \ . \tag{1.315}$$

We may always use the vielbein e_i^a, which is a non-degenerate $n \times n$ matrix in n dimensions, to convert between coordinate indices i and tangent-space indices a. For this purpose we also need the inverse of the vielbein, denoted by E_a^i, and satisfying the defining properties

$$E_a^i\, e_j^a = \delta_j^i \ , \qquad E_a^i\, e_i^b = \delta_b^a \ . \tag{1.316}$$

Then we may define Riemann tensor components entirely within the tangent-frame basis, as follows:

$$R^a{}_{bcd} \equiv E^i_c\, E^j_d\, R^a{}_{bij} \ . \tag{1.317}$$

In terms of $R^a{}_{bcd}$, it is easily seen from the various definitions that we have

$$\Theta^a{}_b = \tfrac{1}{2} R^a{}_{bcd}\, e^c \wedge e^d \ . \tag{1.318}$$

From the Riemann tensor R_{abcd} two further tensors can be defined, as we did in the earlier coordinate-basis discussion, namely the Ricci tensor R_{ab} and the Ricci scalar R:

$$R_{ab} = R^c{}_{acb} \ , \qquad R = \eta^{ab}\, R_{ab} \ . \tag{1.319}$$

We again find that the Riemann tensor and Ricci tensor have the following symmetries, which can be proved straightforwardly from the definitions above:

$$\begin{aligned} &R_{abcd} = -R_{bacd} = -R_{abdc} = R_{cdab} \ , \\ &R_{abcd} + R_{acdb} + R_{adbc} = 0 \ , \\ &R_{ab} = R_{ba} \ . \end{aligned} \tag{1.320}$$

1.15.1 Relation to the coordinate-basis connection and curvature

As we mentioned above, the spin connection $\omega^a{}_b$ and the curvature 2-forms $\Theta^a{}_b$ are really giving an equivalent description of the connection and curvature that we introduced in the eralier coordinate-basis discussion. To make this more precise, we may define a covariant derivative D_i that is covariant with respect to both general coordinate transformations and local-Lorentz transformations. Acting on the vielbein, for example, we shall have

$$D_i e^a_j \equiv \partial_i e^a_j + \omega^a{}_{bi}\, e^b_j - \Gamma^k{}_{ij}\, e^a_k \, . \tag{1.321}$$

The extension of this definition to arbitrary Lorentz-valued general coordinate tensors should be obvious; it is just the appropriate combination of $\omega^a{}_b$ terms to covariantise each local-Lorentz index as in (1.300), and $\Gamma^i{}_{jk}$ terms to covariantise each coordinate index, as in (1.104).

The vielbein and its inverse can be used to map invertibly between coordinate indices and local-Lorentz indices. We would therefore like to have the property that $D_i e^a_j = 0$, so that these mappings will commute with covariant differentiation. This is in fact possible, and by requiring that $D_i e^a_j = 0$ we can obtain a relation between the spin connection $\omega^a{}_b$ and the affine connection $\Gamma^i{}_{jk}$. Thus, from (1.321) we find that $D_i e^a_j = 0$ implies

$$\partial_i e^a_j + \omega^a{}_{bi}\, e^b_j - \Gamma^k_{ij}\, e^a_k = 0 \, . \tag{1.322}$$

Multiplying by $\eta_{ac}\, e^c_k$ and symmetrising in kj gives

$$\partial_i g_{jk} - \Gamma^\ell{}_{ij}\, g_{g\ell} - \Gamma^\ell{}_{ik}\, g_{j\ell} = 0\,, \tag{1.323}$$

which is the same as we saw from (1.106) when we required $\nabla_i g_{jk} = 0$. If we again multiply (1.322) by $\eta_{ac}\, e^c_k$, but this time antisymmetrise in ij, we obtain

$$T^i{}_{jk} = 2\Gamma^i{}_{[jk]}\,, \tag{1.324}$$

where $T^i{}_{jk}$ is the torsion tensor defined by (1.292), with the upper local-Lorentz index converted to a coordinate index using the inverse vielbein: $T^i{}_{jk} = E^i_a\, T^a{}_{jk}$. We see that (1.324) agrees with our previous coordinate-index result in (1.183).

Comparing the curvatures obtained by the two approaches is a slightly involved calculation. Multiplying (1.322) by an inverse vielbein, one can easily see that

$$\Gamma^k{}_{ij} = E^k_a\, \partial_i e^a_j + E^k_a\, \omega^a{}_{bi}\, e^b_j\,. \tag{1.325}$$

Substituting this into the expression (1.138) for the components of the Riemann tensor $R^i{}_{jk\ell}$ in a coordinate basis, and then converting the first two indices to local-Lorentz indices using $R^a{}_{bk\ell} = e^a_i\, E^j_b\, R^i{}_{jk\ell}$, one can, with some perseverance, show that is equal to the expression for $R^a{}_{bk\ell}$ that came from (1.307) and (1.315).

2 General Relativity; Einstein's Theory of Gravitation

2.1 The Equivalence Principle

Men occasionally stumble over the truth, but most of them pick themselves up and hurry off as if nothing ever happened — **Sir Winston Churchill**

Contrary to what one might have expected, Einstein's theory of General Relativity is not based on a yet-further abstraction of the already counter-intuitive theory of Special Relativity. In fact, perhaps remarkably, it has as its cornerstone an observation that is absolutely familiar and intuitively understandable in everyday life. So familiar, in fact, that it took someone with the genius of Einstein to see it for what it really was, and to extract from it a profoundly new way of understanding the world. Sadly, even though this happened ninety years ago, not everyone has yet caught up with the revolution in understanding that Einstein achieved. Nowhere is this more apparent than in the teaching of mechanics in a typical undergraduate physics course.

The cornerstone of Special Relativity is the observation that the speed of light is the same in all inertial frames. From this the consequences of Lorentz contraction, time dilation, and the covariant behaviour of the fundamental physical laws under Lorentz transformations all logically follow. The intuition for understanding Special Relativity is not profound, but it has to be acquired, since it is not the intuition of our everyday experience. In our everyday lives velocities are so small in comparison to the speed of light that we don't notice even a hint of special-relativistic effects, and so we have to train ourselves to imagine how things will behave when the velocities are large. Of course in the laboratory it is now a commonplace to encounter situations where special-relativisitic effects are crucially important.

The cornerstone of General Relativity is the *Principle of Equivalence*. There are many ways of stating this, but perhaps the simplest is the assertion that gravitational mass and inertial mass are the same.

In the framework of Newtonian gravity, the *gravitational mass* of an object is the constant of proportionality M_{grav} in the equation describing the force on an object in the Earth's gravitational field $\vec{g}$:

$$\vec{F} = M_{\text{grav}}\,\vec{g} = \frac{G M_{\text{earth}}\,M_{\text{grav}}\,\vec{r}}{r^3}\,, \tag{2.1}$$

where $\vec{r}$ is the position vector of a point on the surface of the Earth.

More generally, if Φ is the Newtonian gravitational potential then an object with gravitational mass M_{grav} experiences a gravitational force given by

$$\vec{F} = -M_{\text{grav}}\,\vec{\nabla}\Phi\,. \tag{2.2}$$

The *inertial mass* M_{inertial} of an object is the constant of proportionality in Newton's second law, describing the force it experiences if it has an acceleration $\vec{a}$ relative to an inertial frame:

$$\vec{F} = M_{\text{inertial}}\,\vec{a}\,. \tag{2.3}$$

It is a matter of everyday observation, and is confirmed to high precision in the laboratory in experiments such as the Eötvös experiment, that

$$M_{\text{grav}} = M_{\text{inertial}}\,. \tag{2.4}$$

It is an immediate consequence of (2.1) and (2.3) that an object placed in the Earth's gravitational field, with no other forces acting, will have an acceleration (relative to the surface of the Earth) given by

$$\vec{a} = \frac{M_{\text{grav}}}{M_{\text{inertial}}}\,\vec{g}\,. \tag{2.5}$$

From (2.4), we therefore have the famous result

$$\vec{a} = \vec{g}\,, \tag{2.6}$$

which says that all objects fall at the same rate. This was supposedly demonstrated by Galileo in Pisa.

More generally, if the object is placed in a Newtonian gravitational potential Φ then from (2.2) and (2.3) it will suffer an acceleration given by

$$\vec{a} = -\frac{M_{\rm grav}}{M_{\rm inertial}}\,\vec{\nabla}\Phi = -\vec{\nabla}\Phi\,, \tag{2.7}$$

with the second equality holding if the inertial and gravitational masses of the object are equal.

In Newtonian mechanics, this equality of gravitational and inertial mass is noted, the two quantities are set equal and called simply M, and then one moves on to other things. There is nothing in Newtonian mechanics that *requires* one to equate $M_{\rm grav}$ and $M_{\rm inertial}$. If experiments had shown that the ratio $M_{\rm grav}/M_{\rm inertial}$ were different for different objects, that would be fine too; one would simply make sure to use the right type of mass in the right place. For a Newtonian physicist the equality of gravitational and inertial mass is little more than an amusing coincidence, which allows one to use one symbol instead of two, and which therefore makes some equations a little simpler.

The big failing of the Newtonian approach is that it fails to ask *why is the gravitational mass equal to the inertial mass?* Or, perhaps a better and more scientific way to express the question is *what symmetry in the laws of nature forces the gravitational and inertial masses to be equal?* The more we probe the fundamental laws of nature, the more we find that fundamental "coincidences" just don't happen; if two concepts that *a priori* look to be totally different turn out to be the same, nature is trying to tell us something. This, in turn, should be reflected in the fundamental laws of nature.

Einstein's genius was to recognise that the equality of gravitational and inertial mass is much more than just an amusing coincidence; nature is telling us something very profound about gravity. In particular, it is telling us that *we cannot distinguish, at least by a local experiment, between the "force of gravity," and the force that an object experiences when it accelerates relative to an inertial frame.* For example, an observer in a small closed box cannot tell whether he is sitting on the surface of the Earth, or instead is in outer space in a rocket accelerating at 32 ft. per second per second.

The Newtonian physicist responds to this by going through all manner of circumlocutions, and talks about "fictitious forces" acting on the rocket traveller, etc. Einstein, by contrast, recognises a fundamental truth of nature, and declares that, by definition, *gravity is the force experienced by an object that is accelerated relative to an inertial frame.* Winston Churchill's observation, reproduced under the heading of this chapter, rather accurately describes the reaction of the average teacher of Newtonian physics.

Einstein's message is: If it looks like gravity, smells like gravity and feels like gravity, then it *is* gravity!

Once this point is recognised, all kinds of muddles and confusions in Newtonian physics disappear. The observer in the closed box does not have to sneak a look outside before he is allowed to say whether he is experiencing a gravitational force or not. An observer in free fall, such as an astronaut orbiting the Earth, or a person who has fallen out of a window, is genuinely weightless because, by definition, he is in a free-fall frame and thus there is no gravity, locally at least, in his frame of reference. A child sitting on a rotating roundabout (or merry-go-round) in a playground is experiencing an *outward* gravitational force, which can unashamedly be called a centrifugal force (with no need for the quotation marks and the F-word "fictitious" that is so beloved of 218 lecturers!). Swept away completely is the muddling notion of the fictitious "force that dare not speak its name."

Notice that in the new order, there is a radical change of viewpoint about what constitutes an inertial frame. If we neglect any effects due to the Earth's rotation, a Newtonian physicist would say that a person standing in a laboratory is in an inertial frame. By contrast, in general relativity we say that a person who has jumped out of the laboratory window is (temporarily!) in an inertial frame. A person standing in the laboratory is accelerating relative to the inertial frame; indeed, that is why he is experiencing the force of gravity.

To be precise, the concept that one introduces in general relativity is that of the *local inertial frame.* This is a free-fall frame, such as that of the person who jumped out of the laboratory, or of the astronaut orbiting the Earth. We must, in general, insist on the word "local," because, as we shall see later, if there is curvature present then one can only define a free-fall frame in a small local region. For example, an observer falling out of a window in College Station is accelerating relative to an observer falling out of a window in Cambridge, since they are moving, with increasing velocities, along lines that are converging on the centre of the Earth. In a small enough region, however, the concept of the free-fall inertial frame makes sense.

Having recognised the equivalence of gravity and acceleration relative to a local inertial frame, it becomes evident that we can formulate the laws of gravity, and indeed *all* the fundamental laws of physics, in a completely frame-independent manner. To be more precise, we can formulate the fundamental laws of physics in such a way that they take the same form in all frames, whether or not they are locally inertial. In fact, another way of stating the equivalence principle is to assert that the fundamental laws of physics take the same form in all frames, i.e. in all coordinate systems. Not surprisingly, perhaps, the way to make this manifest is to use the formalism of general tensor calculus that we have been studying.

2.2 A Newtonian Interlude

Before proceeding with the main development, it is perhaps worthwhile to pause and consider in more detail how the Einsteinian way of thinking provides a superior framework for solving problems even in Newtonian gravity. Perhaps the best way to do this is by considering a couple of "218 level problems," to see how simply they can be solved by adopting the framework of the general relativist.

2.2.1 The helium balloon

(1) *Consider a helium balloon held on a string, which is in a car accelerating uniformly with acceleration a. What angle does the string make with the vertical?*

In the frame of the car, gravity has two components, namely the Earth's gravity g acting downwards, and a component a directed backwards, which is due to the car's acceleration. The vector sum of these two gives a net gravity of strength $\tilde{g} = \sqrt{g^2 + a^2}$, which is directed downwards and backwards, making an angle $\theta = \arctan(a/g)$ to the vertical. The balloon rises in the gravitational field such that the string is parallel to the direction of gravity, and so it therefore tilts *forwards* by the angle $\theta = \arctan(a/g)$.

It is a while since I have solved such problems by the "traditional" method of the 218 class, but I suspect one would have to draw force diagrams with bouyant forces, reaction forces, etc., etc. I think it is clear that Einstein's method is simpler.

2.2.2 The grandfather clock

(2) *For unclear reasons, a grandfather clock is to be set up in a truck that is accelerating with acceleration a. At what angle to the vertical must the clock be set, in order that it will function properly, and at what rate will it run?*

For the general relativist, this is essentially the same problem as the previous one. For a grandfather clock to function properly, it must be oriented vertically in the gravitational field. Therefore it should be tilted forwards by an angle $\theta = \arctan(a/g)$. The period of its pendulum (of length ℓ, say) will be $\widetilde{T} = 2\pi\sqrt{\ell/\tilde{g}}$, since $\tilde{g} = \sqrt{g^2 + a^2}$ is the strength of the gravitational field. This compares with the period $T = 2\pi\sqrt{\ell/g}$ when the truck is not accelerating. Thus the period of the pendulum is smaller by the factor

$$\frac{\widetilde{T}}{T} = \sqrt{\frac{g}{\tilde{g}}} = \Big(1 + \frac{a^2}{g^2}\Big)^{-1/4} \tag{2.8}$$

when it is set up in the accelerating truck.

I don't even want to contemplate how one would solve this problem by the traditional 218 methods! It would, I think, be considerably harder than the general relativist's approach, by which one can solve these problems in one's head.

Not only does the general relativist's approach have the merit of simplicity, it also emphasises the essential *unity* of the subject. Using traditional "218 methods," each problem such as (1) or (2) above has to be solved by its own method, and the essential point that one is really solving the same problem each time is lost in a haze of force diagrams, bouyancy forces, reaction forces, etc.

After this little Newtonian interlude, let us now return to general relativity in all its glory.

2.3 The Geodesic Equation

Consider first a particle in Minkowski spacetime, on which no external forces are acting. We shall denote the spacetime coordinates by x^μ, where $0 \le \mu \le 3$, with $\mu = 0$ corresponding to the time coordinate. The Minkowski metric is then

$$ds^2 = \eta_{\mu\nu}\, dx^\mu\, dx^\nu \,, \tag{2.9}$$

where

$$\eta = \mathrm{diag}\,(-1, 1, 1, 1)\,. \tag{2.10}$$

(We are using units here where the speed of light is set equal to 1. For example, distance is measured in light-seconds.) The quantity ds^2 gives the squared interval between two neighbouring spacetime "events," at (x^0, x^1, x^2, x^3) and $(x^0{+}dx^0, x^1{+}dx^1, x^2{+}dx^2, x^3{+}dx^3)$. One also has the notion of the *proper-time interval* $d\tau$ between the events, where

$$d\tau^2 = -ds^2 = -\eta_{\mu\nu}\, dx^\mu\, dx^\nu \,. \tag{2.11}$$

If the two events are at the same spatial location, so that $dx^1 = dx^2 = dx^3 = 0$, then the proper-time interval is just equal to the coordinate time interval dt, where $t = x^0$. Thus, for example, proper time is the coordinate time in the rest frame of a particle.

Assuming that the particle is not massless (so that it has a rest frame), we can use the elapse of proper time to parameterise the motion of the particle. In other words, we can say that the particle's spacetime coordinates at proper time τ are given by $x^\mu(\tau)$. Clearly the particle will move in a straight line, since no external forces are acting, and we can characterise this by the equation

$$\frac{d^2 x^\mu(\tau)}{d\tau^2} = 0\,. \tag{2.12}$$

Actually, it will be convenient to suppose that we start out using a coordinate system x'^μ, so that the equation for the particle's motion is

$$\frac{d^2 x'^\mu(\tau)}{d\tau^2} = 0\,. \tag{2.13}$$

Now we transform to a completely arbitrary system of coordinates x^μ; these can be related to the x'^μ by any general coordinate transformation. Using the chain rule, we therefore have from (2.13) that

$$\frac{d^2 x^\mu}{d\tau^2} + \frac{\partial x^\mu}{\partial x'^\nu}\,\frac{\partial^2 x'^\nu}{\partial x^\rho \partial x^\sigma}\,\frac{dx^\rho}{d\tau}\,\frac{dx^\sigma}{d\tau} = 0\,. \tag{2.14}$$

We can also calculate the metric tensor $g_{\mu\nu}$ in the unprimed coordinate system, by using the chain rule:

$$ds^2 = \eta_{\alpha\beta}\, dx'^\alpha\, dx'^\beta = \frac{\partial x'^\alpha}{\partial x^\mu}\,\frac{\partial x'^\beta}{\partial x^\nu}\,\eta_{\alpha\beta}\, dx^\mu\, dx^\nu \equiv g_{\mu\nu}\, dx^\mu\, dx^\nu\,, \tag{2.15}$$

and so

$$g_{\mu\nu} = \frac{\partial x'^\alpha}{\partial x^\mu}\,\frac{\partial x'^\beta}{\partial x^\nu}\,\eta_{\alpha\beta}\,. \tag{2.16}$$

Similarly, the inverse metric is given by

$$g^{\mu\nu} = \frac{\partial x^\mu}{\partial x'^\alpha}\,\frac{\partial x^\nu}{\partial x'^\beta}\,\eta^{\alpha\beta}\,. \tag{2.17}$$

It is a straightforward, if somewhat tedious, exercise to verify from the definition (1.110) of the affine connection that the equation (2.14) is nothing but

$$\frac{d^2 x^\mu}{d\tau^2} + \Gamma^\mu{}_{\nu\rho}\,\frac{dx^\nu}{d\tau}\,\frac{dx^\rho}{d\tau} = 0\,. \tag{2.18}$$

We derived equation (2.18) by rewriting the equation for a particle in free straight-line motion in Minkowski spacetime in an arbitrary coordinate system. However, (2.18) is in fact a completely covariant equation, and we can adopt it as the *definition* of "straight-line motion" for a particle in any spacetime. More properly, we should not say straight-line motion, since that is a rather ambiguous notion. Instead, we call it *Geodesic motion*. Thus (2.18) is the *Geodesic Equation*, which describes the motion of a massive particle in free-fall motion in any spacetime, flat or curved.

One can easily see, from the definition of the covariant derivative, that (2.18) can be written as

$$\frac{dx^\nu}{d\tau}\,\nabla_\nu \frac{dx^\mu}{d\tau} = 0\,. \tag{2.19}$$

This makes it manifest that the geodesic equation is covariant, since clearly $dx^\mu/d\tau$ are the components of a vector. In fact, we could have got directly from (2.13) to (2.18) for a particle in Minkowski spacetime rewritten in the unprimed coordinate system, simply by noting that (2.13) can be written as

$$\frac{dx'^\nu}{d\tau}\,\partial'_\nu \frac{dx'^\mu}{d\tau} = 0\,, \tag{2.20}$$

and then noting that under a change of coordinates, this must become an equation that is covariant with respect to general coordinate transformations. Thus it must become (2.19), since there is no other possible covariant equation one could write down.

The geodesic equation (2.18) is the analogue in general relativity of Newton's second law applied to the case of a particle in a gravitational field. To see this, it is useful to consider the geodesic equation in the Newtonian limit, where the gravitational field is very weak and stationary, and the particle is moving slowly. It will be convenient to split the spacetime coordinate index μ into $\mu = (0, i)$, where i ranges only over the spatial index values, $1 \le i \le 3$. Saying that the velocity is small (compared with the speed of light) means that

$$\Big|\frac{dx^i}{dt}\Big| << 1\,. \tag{2.21}$$

From (2.11) it follows that coordinate time t and proper time τ are essentially the same, and thus we also have

$$\frac{dx^0}{d\tau} \approx 1\,. \tag{2.22}$$

Consider now the spatial components of the geodesic equation (2.18). In this Newtonian limit, it therefore approximates to

$$\frac{d^2x^i}{dt^2} + \Gamma^i{}_{00} = 0\,. \tag{2.23}$$

Furthermore, since we are assuming the gravitational field is weak, we can assume that the metric is nearly flat, in which case we can choose a coordinate system in which it is approximated by small deviations from the Minkowski metric:

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}\,, \tag{2.24}$$

where the deviations $h_{\mu\nu}$ are very small compared to 1. From the expression (1.110) for the Christoffel connection, we therefore have, with the stationarity assumption $\partial h_{\mu\nu}/\partial t = 0$, that

$$\Gamma^i{}_{00} \approx -\tfrac{1}{2}\partial_i h_{00}\,. \tag{2.25}$$

Thus the geodesic equation reduces in the Newtonian limit to

$$\frac{d^2x^i}{dt^2} = \tfrac{1}{2}\partial_i h_{00}\,. \tag{2.26}$$

We now compare this with the Newtonian equation for a particle moving in a gravitational field. If the Newtonian potential is Φ, then the equation of motion following from Newton's second law (assuming that the gravitational and inertial masses are equal!) is

$$\frac{d^2x^i}{dt^2} = -\partial_i\Phi\,. \tag{2.27}$$

Comparing with (2.26), we see that

$$h_{00} = -2\Phi\,. \tag{2.28}$$

(We can take the constant of integration to be zero, since at large distance, where the Newtonian potential vanishes, the metric should reduce to exactly the Minkowski metric.)

Notice that in general relativity the equality of gravitational and inertial mass is built in from the outset; the geodesic equation (2.18) makes no reference to the mass of the particle.

Another important point is to note that in the geodesic equation (2.18), the Christoffel connection $\Gamma^\mu{}_{\nu\rho}$ is playing the rôle of the "gravitational force," since it is this term that describes the deviation from "linear motion" $d^2x^\mu/d\tau^2 = 0$. The fact that the gravitational force is described by a connection, and not by a tensor, is just as one would hope and expect. The point is that gravity can come or go, depending on what system of coordinates one uses. In particular, if one chooses a free-fall frame, in which the metric at any given point can be taken to be the Minkowski metric, and the first derivatives can also be taken to vanish at the point, then the Christoffel connection vanishes at the point also. Thus indeed, we have the vanishing of gravity (weightlessness) in a local free-fall frame.

2.4 The Einstein Field Equation

So far, we have seen how matter responds to gravity, namely, the geodesic equation which shows how matter moves under the influence of the gravitational field. The other side of the coin is to see how gravity is determined by matter. The equation which controls this is the Einstein field equation. This is the analogue of the Newtonian equation

$$\nabla^2 \Phi = 4\pi G \rho \,, \tag{2.29}$$

which governs the Newtonian gravitational potential Φ in the presence of a mass density ρ. Here G is Newton's constant.

The required field equation in general relativity can be expected, like Newton's field equation, to be of order 2 in derivatives. Again we can proceed by considering first the Newtonian limit of general relativity. Since, as we have seen, the deviation h_{00} of the metric component g_{00} from its Minkowskian value -1 is equal to -2Φ in the Newtonian limit, we are led to expect that the Einstein field equation should involve second derivatives of the metric. We also expect that it should be a tensorial equation, since we would like it to have the same form in all coordinate frames. Luckily, there exist candidate tensors constructed from the metric, since, as we saw earlier, the Riemann tensor, and its contractions to the Ricci tensor and Ricci scalar, involve second derivatives of the metric. Some appropriate construct built from the curvature will therefore form the "left-hand side" of the Einstein equation.

There remains the question of what will sit on the right-hand side, generalising the mass density ρ. There is again a natural tensor generalisation, namely the *energy-momentum tensor*, or *stress tensor*, $T_{\mu\nu}$. This is a symmetric tensor that describes the distribution of mass (or energy) density, momentum flux density, and stresses in a matter system. Specifically, if we decompose the four-dimensional spacetime index μ as $\mu = (0, i)$ as before, then T_{00} describes the mass density, T_{0i} describes the 3-momentum flux, and T_{ij} describes the stresses within the matter system.

A very important feature of the energy-momentum tensor for a closed system is that it is *conserved*, meaning that

$$\nabla^\mu T_{\mu\nu} = 0 \,. \tag{2.30}$$

This is analogous to the conservation law $\nabla^\mu J_\mu = 0$ for the 4-vector current density in electromagnetism. In that case, the conservation law ensures that charge is conserved, and by integrating J_0 over a closed spatial 3-volume and taking a time derivative, one shows that the rate of change of total charge within the 3-volume is balanced by the flux of electric current out of the 3-volume. Analogously, (2.30) ensures that the rate of change of total energy within a closed 3-volume is balanced by the momentum flux out of the region.

If we are to build a field equation whose right-hand side is a constant multiple of $T_{\mu\nu}$, it follows, therefore, that the left-hand side must also satisfy a conservation condition. There is precisely one symmetric 2-index tensor built from the curvature that has this property, namely the *Einstein tensor*

$$G_{\mu\nu} \equiv R_{\mu\nu} - \tfrac{1}{2} R\, g_{\mu\nu}\,, \tag{2.31}$$

which we met in equation (1.163). Thus our candidate field equation is $G_{\mu\nu} = \lambda\, T_{\mu\nu}$, i.e.

$$R_{\mu\nu} - \tfrac{1}{2} R\, g_{\mu\nu} = \lambda\, T_{\mu\nu}\,, \tag{2.32}$$

for some universal constant λ, which we may determine by requiring that we obtain the correct weak-field Newtonian limit.

In a situation where the matter system has low velocities, its energy-momentum tensor will be dominated by the T_{00} component, which describes the mass density ρ. Thus to find the Newtonian limit of (2.32), we should examine the 00 component. To do this, it is useful first to take the trace of (2.32), by multiplying by $g^{\mu\nu}$. This gives

$$-R = \lambda\, g^{\mu\nu}\, T_{\mu\nu}\,. \tag{2.33}$$

Since $T_{\mu\nu}$ is dominated by $T_{00} = \rho$, and the metric is nearly the Minkowski metric (so $g^{00} \approx -1$), we see that

$$R \approx \lambda\, \rho \tag{2.34}$$

in the Newtonian limit. Thus, (2.32) reduces to

$$R_{00} \approx \tfrac{1}{2} \lambda \rho\,. \tag{2.35}$$

It is easily seen from the expression (1.138) for the Riemann tensor, and the definition (1.160) for the Ricci tensor, that from (2.25) the component R_{00} is dominated by

$$R_{00} \approx \partial_i \Gamma^i{}_{00} \approx -\tfrac{1}{2} \partial_i \partial^i\, h_{00}\,. \tag{2.36}$$

From (2.28) we therefore have that $R_{00} \approx \nabla^2\, \Phi$ in the Newtonian limit, and hence, from (2.35), we obtain the result

$$\nabla^2\, \Phi \approx \tfrac{1}{2} \lambda\, \rho\,. \tag{2.37}$$

It remains only to compare this with Newton's equation (2.29), thus determining that $\lambda = 8\pi G$.

In summary, therefore, we have shown that the Einstein field equation

$$R_{\mu\nu} - \tfrac{1}{2} R\, g_{\mu\nu} = 8\pi\, G\, T_{\mu\nu} \tag{2.38}$$

has the correct Newtonian limit.

Note that the Einstein equation is a gravitational analogue of the field equation in Maxwell's theory of electromagnetism. Let us consider Maxwell's equations in Minkowski spacetime for simplicity. One introduces the antisymmetric Maxwell tensor $F_{\mu\nu}$, whose components are given in terms of the 3-vector electric field and magnetic field by

$$F_{0i} = -F_{i0} = -E_i\,, \qquad F_{ij} = \epsilon_{ijk}\, B_k\,. \tag{2.39}$$

The Maxwell field-strength tensor $F = \frac{1}{2} F_{\mu\nu}\, dx^\mu \wedge dx^\nu$ (which is a 2-form) can be expressed in terms of the exterior derivative of the 1-form gauge potential $A = A_\mu\, dx^\mu$, namely $F = dA$, or in terms of components,

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu\,. \tag{2.40}$$

The Maxwell field equation then reads

$$\partial_\mu\, F^{\mu\nu} = -4\pi\, J^\nu\,, \tag{2.41}$$

where J^μ is the 4-vector current density, with components

$$J^0 = \rho_e\,, \qquad J^i = j^i\,. \tag{2.42}$$

Here ρ_e is the electric charge density and the 3-vector $\vec{j}$, with components j^i, is the 3-vector current density. By considering the cases $\nu = 0$ and $\nu = i$ in the Maxwell equation (2.41), one recovers the familiar 3-vector Maxwell equations

$$\vec{\nabla} \cdot \vec{E} = 4\pi\, \rho_e\,, \qquad \vec{\nabla} \times \vec{B} - \frac{\partial \vec{E}}{\partial t} = 4\pi\, \vec{j} \tag{2.43}$$

respectively.[8] In terms of the gauge potential A introduced in (2.40), the Maxwell equation (2.41) is of second-order in derivatives.

[8]We have chosen units where the permitivity and permeability of free space are set to 1. Note that all the dimensionful so-called "fundamental constants," such as the speed of light, the permitivity of free space, Newton's constant, etc., are in actuality not really fundamental at all, but merely reflect the fact that we sometimes choose, for no logically necessary reason, to use different systems of units for measuring quantities that could perfectly well be measured in the same units. An example, which we met already, is that one can set the speed of light equal to 1 if one measures distance in light-seconds.

Note that with F written, locally at least, as $F = dA$, we have $dF = 0$, which reads in component language

$$\partial_{[\mu}\, F_{\nu\rho]} = 0\,. \tag{2.44}$$

Taking the two distinct cases $(\mu\nu\rho) = (0, i, j)$ and $(\mu\nu\rho) = (ijk)$, this implies

$$\vec\nabla \times \vec E + \frac{\partial \vec B}{\partial t} = 0\,, \qquad \vec\nabla \cdot \vec B = 0 \tag{2.45}$$

respectively. These are the remaining two Maxwell equations in terms of 3-vector notation. These are really identities, rather than field equations. In fact there is a close analogy with the notion of curvature in differential geometry, with F being the curvature of the connection A, and the equation $dF = 0$ being a Bianchi identity.

The Einstein equation (2.38) and the Maxwell equation (2.41) have considerable similarities, namely a left-hand side that is a curvature built from the fundamental field of the theory, and a right-hand side that is a source term, built from quantities such as mass or charge densities, currents, etc.

The discussion of the Maxwell equations that we gave above was in the case of a Minkowski spacetime background. It is, however, almost a triviality to generalise this to the case of an arbitrary spacetime background. We need to find a generally-covariant generalisation of the Minkowski-spacetime equation (2.41). The answer is trivially easy; the only generally-covariant equation, with the same number of derivatives, that has the property of reducing to (2.41) in Minkowski spacetime is

$$\nabla_\mu\, F^{\mu\nu} = -4\pi\, J^\nu\,, \tag{2.46}$$

and so this is what the Maxwell field equation in a general spacetime must be. The other "half" of the Maxwell equations, namely the Bianchi identity (2.44), requires no modification at all, since, as we well know, (2.44) is already a generally-covariant equation.

We have described above the form of the Maxwell equations in a general curved spacetime. In order to complete the discussion of the Einstein-Maxwell system, we need to consider the Einstein equation. The energy-momentum tensor for the electromagnetic field is given by

$$T_{\mu\nu} = \frac{1}{4\pi}\,(F_{\mu\rho}\, F_\nu{}^\rho - \tfrac14 F^2\, g_{\mu\nu})\,, \tag{2.47}$$

where $F^2 = F^{\mu\nu}\, F_{\mu\nu}$. We then substitute this into the Einstein equation (2.38). Setting Newton's constant $G = 1$ for convenience, the complete system of equations for gravity and electromagnetism, known collectively as the Einstein-Maxwell equations, is therefore

$$R_{\mu\nu} - \tfrac12 R\, g_{\mu\nu} \;\; = \;\; 2(F_{\mu\rho}\, F_\nu{}^\rho - \tfrac14 F^2\, g_{\mu\nu})\,,$$

$$\begin{aligned}\nabla_\mu F^{\mu\nu} &= 0\,,\\ \partial_{[\mu}\, F_{\nu\rho]} &= 0\,.\end{aligned} \tag{2.48}$$

(Note that we have taken the 4-vector current J^μ to be zero here. In other words, we the equations we have written are for the pure Einstein-Maxwell system, with no additional charged matter present.)

2.5 The Schwarzschild Solution

Although the Einstein equation (2.38) and the Maxwell equation (2.41) have quite a lot in common, there is, however, a very important difference, which reflects itself in the difficulty of solving the equation, and in the richness of the solutions.

The left-hand side of the Maxwell equation is *linear* in the basic gauge field A, whereas the basic field in general relativity, i.e. the metric tensor $g_{\mu\nu}$, appears highly non-linearly in the left-hand side of the Einstein equation. As a result, when one looks for solutions to the field equations in general relativity, one is faced with the problem of solving non-linear, rather than linear, differential equations.

On account of the non-linearity of the Einstein equation, it came as a considerable surprise to everybody in 1916 (one year after Einstein published his general theory of relativity) when Karl Schwarzschild succeeded in obtaining the exact solution for a spherically symmetric mass distribution. It was more or less his final achievement; he died in Russia a few months later, having enlisted to fight in the First World War.

The Schwarzschild solution is arguably the most important solution in general relativity. It is the analogue of the solution for the electric field outside a spherical charge distribution in Maxwell's theory, but it is enormously more subtle and intriguing. In fact, it was really only in the 1960's that it was properly understood, and taken seriously in its own right. It is the solution describing a spherically-symmetric black hole.

The derivation of the Schwarzschild solution is rather straightforward, and in view of its simplicity, we shall present it here. After a sequence of arguments, based on symmetry considerations together with the fact that one can choose coordinates arbitrarily in general relativity, it can be established that with a convenient choice of coordinate system, the metric for a static and spherically-symmetric geometry can be written in the form

$$ds^2 = -B(r)\, dt^2 + A(r)\, dr^2 + r^2\, (d\theta^2 + \sin^2\theta\, d\phi^2)\,, \tag{2.49}$$

where $A(r)$ and $B(r)$ are as-yet arbitrary functions of the radial variable r. They will be determined by solving the Einstein equation. Note that if we were to set $A(r) = B(r) = 1$, we would just get the metric

$$ds^2 = -dt^2 + dr^2 + r^2\,(d\theta^2 + \sin^2\theta\,d\phi^2)\,, \tag{2.50}$$

which is nothing but Minkowski spacetime with the spatial Euclidean 3-metric written in spherical polar coordinates.

Our interest is in the case of a source-free static spherically-symmetric solution, which is the gravitatiional analogue of the point charge in electrodynamics. Thus we wish to solve the Einstein equation (2.38) with $T_{\mu\nu} = 0$. By taking the trace (i.e. by multiplying by $g^{\mu\nu}$), we immediately see that the Ricci scalar must vanish, and hence the vacuum Einstein equation reduces to the Ricci-flat condition

$$R_{\mu\nu} = 0\,. \tag{2.51}$$

To solve this, we take the assumed metric form (2.49), and then calculate the Christoffel connection, the Riemann tensor, and finally the Ricci tensor. Demanding that this be zero will then give us some non-linear equations for the metric functions $A(r)$ and $B(r)$. Taking the coordinate indices to be

$$x^0 = t\,,\quad x^1 = r\,,\quad x^2 = \theta\,,\quad x^3 = \phi\,, \tag{2.52}$$

it is not hard to see from (1.110) that the non-vanishing components of the Christoffel connection $\Gamma^\mu{}_{\nu\rho}$ are given by

$$\begin{aligned}
&\Gamma^0{}_{01} = \frac{B'}{2B}\,,\\
&\Gamma^1{}_{00} = \frac{B'}{2A}\,,\quad \Gamma^1{}_{11} = \frac{A'}{2A}\,,\quad \Gamma^1{}_{22} = -\frac{r}{A}\,,\quad \Gamma^1{}_{33} = -\frac{r\sin^2\theta}{A}\,,\\
&\Gamma^2{}_{12} = \frac{1}{r}\,,\quad \Gamma^2{}_{33} = -\sin\theta\cos\theta\,,\\
&\Gamma^3{}_{13} = \frac{1}{r}\,,\quad \Gamma^3{}_{23} = \cot\theta\,.
\end{aligned} \tag{2.53}$$

(Of course, as always the symmetry in the lower two indices is understood, so we do not need to list the further components that are implied by this.) The notation here is that $A' = dA/dr$ and $B' = dB/dr$.

Plugging into the definition of the Rieman tensor, and then contracting to get the Ricci tensor, one then finds that the non-vanishing components are given by

$$\begin{aligned}
R_{00} &= \frac{B''}{2A} - \frac{B'}{4A}\Big(\frac{A'}{A} + \frac{B'}{B}\Big) + \frac{B'}{rA}\,,\\
R_{11} &= -\frac{B''}{2B} + \frac{B'}{4B}\Big(\frac{A'}{A} + \frac{B'}{B}\Big) + \frac{A'}{rA}\,,
\end{aligned}$$

$$
\begin{aligned}
R_{22} &= 1 + \frac{r}{2A}\Big(\frac{A'}{A} - \frac{B'}{B}\Big) - \frac{1}{A}\,, \\
R_{33} &= R_{22}\,\sin^2\theta\,.
\end{aligned} \tag{2.54}
$$

To solve the Ricci-flatness condition (2.51) we first note that setting $AR_{00} + BR_{11} = 0$ gives

$$\frac{1}{r}\Big(B' + \frac{A'\,B}{A}\Big) = 0\,, \tag{2.55}$$

which implies $(AB)' = 0$. Thus we have

$$AB = \text{constant}\,. \tag{2.56}$$

Now at large distance, we expect the metric to approach Minkowski spacetime, and so it should approach (2.50). This determines that $A(r)$ and $B(r)$ should both approach 1 at large distance, and hence we see that the constant in the solution (2.56) should be 1, and so $A = 1/B$.

From the condition $R_{22} = 0$, we then obtain the equation

$$1 - rB' - B = 0\,, \tag{2.57}$$

which can be written as

$$(rB)' = 1\,. \tag{2.58}$$

The solution to this, with the requirement that $B(r)$ approach 1 at large r, is given by

$$B = 1 + \frac{a}{r}\,, \tag{2.59}$$

where a is a constant. It is straightforward to verify that all the Einstein equations implied by $R_{\mu\nu} = 0$ are now satisfied.

Recalling that we showed previously that in the weak-field Newtonian limit, the metric $g_{\mu\nu}$ is approximately of the form $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$ with $h_{00} = -2\Phi$, where Φ is the Newtonian gravitational potential (see equation (2.28)), it follows that the constant a in (2.59) can be determined, by considering the Newtonian limit. Thus we shall have $a = -2GM$, where G is Newton's constant. Usually, in general relativity we choose units where $G = 1$, and so we arrive at the Schwarzschild solution

$$ds^2 = -\Big(1 - \frac{2M}{r}\Big)\,dt^2 + \Big(1 - \frac{2M}{r}\Big)^{-1}\,dr^2 + r^2\,(d\theta^2 + \sin^2\theta\,d\phi^2)\,. \tag{2.60}$$

This describes the gravitational field outside a spherically-symmetric static mass M.

As expected, the solution approaches Minkowski spacetime at large radius. It is clear that something rather drastic happens to the metric when r approaches $2M$. This radius, known as the *Schwarzschild Radius*, was thought for many years to correspond to some singularity of the solution. It was really only in the 1950's that it was first understood that the apparent singularity is merely a result of using a system of coordinates that becomes ill-behaved there. There is nothing singular about the solution as such. For example, the curvature is perfectly finite there, and in fact the only place where there is a curvature singularity is at $r = 0$.

There is, nevertheless, something intrinsically special about the Schwarzschild radius. The coordinate system in which the metric (2.60) is written is one that it well-adapted to describing the large-radius, or asymptotically flat, region. The coefficient of dt^2 passes through 0 and becomes positive for $r < 2M$, which means that t has then become a spatial coordinate. At the same time, the coefficient of dr^2 becomes negative, and so inside the region $r < 2M$ the coordinate r has become timelike.

What is happening is that the *light cones* (i.e. the trajectories followed by the light-fronts coming from a flash of light) are tilting over more and more as one approaches $r = 2M$ from the outside. This means that it is getting harder and harder for anything to escape out to infinity. Eventually, by the time one reaches $r = 2M$, not even light is able to escape to the future. This is what is called an *event horizon*. Nothing can escape to the outside from within the event horizon, and it is for this reason that John Wheeler coined the term *black hole* to describe the object.

The full global structure of black holes was finally understood in the 1960's. At that stage, they were still thought to be rather abstract and physically unrealistic objects, although work by Hawking, Penrose and others had by then established that it was inevitable that stars beyond a certain mass would eventually inevitably collapse to black holes, once the nuclear reactions supporting them against gravitational collapse were exhausted.

In much more recent times it has been understood that there is a giant black hole sitting at the centre of virtually every galaxy, including our own.

It should also be emphasised that the Schwarzschild solution can be used to describe the geometry outside any spherically-symmetric and stationary mass distribution, such as a non-rotating star or planet. In that case, the Schwarzschild solution itself would apply only down to the radius of the surface of the object (i.e. only in the exterior region where there is no matter). Inside the object, one would match on the metric that arises as the solution of the Einstein equation with a $T_{\mu\nu}$ source term, where $T_{\mu\nu}$ is the energy-momentum tensor for the matter of which the star or planet is composed. The solution is then precisely analogous to solving Maxwell's equations for the field outside a spherically-symmetric charge distribution. Outside the charged object, one is just solving the vacuum Maxwell equation, which leads to a potential Q/r where Q is the total charge of the configuration. Inside the object, one has to solve the Maxwell equation with the detailed charge distribution as the source. The two are matched at the surface of the charged object.

Thus, for example, the Schwarzschild metric (2.60) describes the exact gravitational field outside the sun (assuming that we neglect effects of rotation), with M being the mass of the sun. Of course in a case such as this, the solution (2.60) applies down to $r = r_{\rm sun}$, the radius of the sun. This radius is very much greater than the Schwarzschild radius $2M$ for a black hole whose mass is that of the sun (which would be about 1km).

2.6 Orbits Around a Star or Black Hole

In section 2.3, we derived the geodesic equation (2.18), which describes how a test particle will move in an arbitrary gravitational field. We can now use this equation to study the orbits of particles moving in the Schwarzschild geometry. This allows us to study, for example, planetary orbits around the sun. In particular, we can then investigate the deviation from the usual Kepler laws of planetary orbits implied by general relativity. We can also consider orbits in the more extreme situation of a black hole.

In order to study the geodesic equation in detail, it is useful first to show how it can be derived from a variational principle. This works as follows. We consider the Lagrangian

$$\mathcal{L} = \tfrac{1}{2} g_{\mu\nu}\, \dot x^\mu\, \dot x^\nu\,, \tag{2.61}$$

where a dot denotes a derivative with respect to the proper time τ that parameterises the path $x^\mu(\tau)$ of the particle. (We are assuming here that the particle is not massless, so that we can use the proper time to parameterise its path.) The geodesic equation (2.18) can then be derived by requiring that the action

$$I = \int \mathcal{L}\, d\tau \tag{2.62}$$

be stationary with respect to variations of the path $x^\mu(\tau)$. To show this, we perform the following manipulations:

$$\begin{aligned}
\delta I &= \int \delta\mathcal{L}\, d\tau = \tfrac12 \int (g_{\mu\nu}\, \dot x^\mu\, \delta\dot x^\nu + \tfrac12 \partial_\rho g_{\mu\nu}\, \dot x^\mu\, \dot x^\nu\, \delta x^\rho)\, d\tau\\
&= \int \Big(-\frac{d}{d\tau}(g_{\mu\rho}\dot x^\mu) + \tfrac12 \partial_\rho g_{\mu\nu}\, \dot x^\mu\, \dot x^\nu\Big)\, \delta x^\rho\, d\tau\\
&= \int \Big(- g_{\mu\rho}\, \ddot x^\mu - \partial_\nu g_{\mu\rho}\, \dot x^\nu\, \dot x^\mu + \tfrac12 \partial_\rho g_{\mu\nu}\, \dot x^\mu\, \dot x^\nu\Big)\, \delta x^\rho\, d\tau\,.
\end{aligned} \tag{2.63}$$

In the first line, we have simply varied the path $x^\nu(\tau)$, with the second term taking into account that $g_{\mu\nu}$ itself depends on x^ρ. In the second line, we have integrated the first term by parts, throwing the derivative off the δx^ν and onto its cofactor. We also, for convenenience, relabelled the dummy index ν in the first term as ρ, so that we could pull out a factor δx^ρ overall. In the third line we distributed the $d/d\tau$ in the first term, using the chain rule to differentiate $g_{\mu\rho}$ which depends on x^ν which depends on τ.

Demanding that the action be stationary under this infinitesimal variation of the path amounts to requiring that $\delta I = 0$ for any δx^ρ, which therefore means that

$$g_{\mu\rho}\, \ddot{x}^\mu + \partial_\nu g_{\mu\rho}\, \dot{x}^\nu\, \dot{x}^\mu - \tfrac{1}{2}\partial_\rho g_{\mu\nu}\, \dot{x}^\mu\, \dot{x}^\nu = 0\,. \tag{2.64}$$

Note that all we have been doing here is deriving the Euler-Lagrange equations

$$\frac{d}{d\tau}\Big(\frac{\partial \mathcal{L}}{\partial \dot{x}^\mu}\Big) - \frac{\partial \mathcal{L}}{\partial x^\mu} = 0\,, \tag{2.65}$$

for the Lagrangian (2.61), where $\mathcal{L}$ depends on a set of coordinates x^μ and their velocities $\dot{x}^\mu$.

In view of the symmetry of $\dot{x}^\mu\, \dot{x}^\nu$ under the interchange of μ and ν, we can rewrite (2.64) as

$$g_{\mu\rho}\, \ddot{x}^\mu + \tfrac{1}{2}(\partial_\nu g_{\mu\rho} + \partial_\mu g_{\nu\rho} - \partial_\rho g_{\mu\nu})\, \dot{x}^\mu\, \dot{x}^\nu = 0\,. \tag{2.66}$$

Finally, multiplying by $g^{\sigma\rho}$ we obtain

$$\ddot{x}^\sigma + \tfrac{1}{2} g^{\sigma\rho}(\partial_\nu g_{\mu\rho} + \partial_\mu g_{\nu\rho} - \partial_\rho g_{\mu\nu})\, \dot{x}^\mu\, \dot{x}^\nu = 0\,, \tag{2.67}$$

and then from the definition (1.110) we can recognise this as precisely the geodesic equation (2.18), namely (after an index relabelling)

$$\ddot{x}^\mu + \Gamma^\mu{}_{\nu\rho}\, \dot{x}^\nu\, \dot{x}^\rho = 0\,. \tag{2.68}$$

Before proceeding, it is worth pausing to note, as an aside, that we can use the result above as a convenient way to calculate the Christoffel connection components in any metric. We just write down the Lagrangian (2.61), derive the Euler-Lagrange equations (2.65) in the standard way, and organise the resulting equation (after raising the free index ν using $g^{\mu\nu}$, for each value of μ, in the form (2.68). We can then simply read off the Christoffel connection components. The nice thing about this calculation is that each Euler-Lagrange equation (i.e. for each value of μ) provides the results for all the connection components $\Gamma^\mu{}_{\nu\rho}$ for all ν and ρ in one go.

Here's an example, for the 2-sphere metric $ds^2 = d\theta^2 + \sin^2\theta d\phi^2$ that we studied earlier. The Lagrangian is therefore

$$\mathcal{L} = \tfrac{1}{2}\dot\theta^2 + \tfrac{1}{2}\sin^2\theta\,\dot\phi^2\,, \tag{2.69}$$

leading to the Euler-Lagrange equations

$$\ddot\theta + \sin\theta\cos\theta\,\dot\phi^2 = 0\,, \qquad \sin^2\theta\,\ddot\phi + 2\sin\theta\cos\theta\,\dot\theta\dot\phi = 0\,. \tag{2.70}$$

Thus we read off, taking $x^1 = \theta$, $x^2 = \phi$, that

$$\Gamma^1{}_{22} = \sin\theta\cos\theta\,, \qquad \Gamma^2{}_{12} = \cot\theta\,. \tag{2.71}$$

(Take care about the factor of $\frac{1}{2}$ when reading off a component such as $\Gamma^2{}_{12}$!)

The upshot of this somewhat lengthy diversion is that we have a rather simple way of obtaining the geodesic equation without the necessity of slogging out the expressions for all the components of the Christoffel connection. There is one further thing we should do, and that is to note that for the actual path followed by the particle, the Lagrangian (2.61) is equal to $-\frac{1}{2}$. This is easily seen; we just note that it is given by

$$\mathcal{L} = \tfrac{1}{2}g_{\mu\nu}\,\dot x^\mu\,\dot x^\nu = \tfrac{1}{2}g_{\mu\nu}\,\frac{dx^\mu}{d\tau}\,\frac{dx^\nu}{d\tau} = \tfrac{1}{2}\frac{g_{\mu\nu}\,dx^\mu\,dx^\nu}{d\tau^2} = -\tfrac{1}{2}\frac{d\tau^2}{d\tau^2} = -\tfrac{1}{2}\,. \tag{2.72}$$

Let us now apply the above discussion to the case of geodesics in the geometry of the Schwarzschild metric (2.60). We therefore consider the Lagrangian

$$\mathcal{L} = -\tfrac{1}{2}B\,\dot t^2 + \tfrac{1}{2}B^{-1}\,\dot r^2 + \tfrac{1}{2}r^2(\dot\theta^2 + \sin^2\theta\,\dot\phi^2)\,, \tag{2.73}$$

where as before

$$B = 1 - \frac{2M}{r}\,. \tag{2.74}$$

As in any Lagrangian problem, if $\mathcal{L}$ does not depend on a particular coordinate q (i.e. it is what is called an "ignorable coordinate"), then one has an associated first integral, since its Euler-Lagrange equation

$$\frac{d}{d\tau}\Big(\frac{\partial\mathcal{L}}{\partial\dot q}\Big) - \frac{\partial\mathcal{L}}{\partial q} = 0 \tag{2.75}$$

reduces to

$$\frac{d}{d\tau}\Big(\frac{\partial\mathcal{L}}{\partial\dot q}\Big) = 0\,, \tag{2.76}$$

which can be integrated to give

$$\frac{\partial\mathcal{L}}{\partial\dot q} = \text{constant}\,. \tag{2.77}$$

In our case, t and ϕ are ignorable coordinates, and so we have the two first integrals

$$B\,\dot t = E\,, \qquad r^2\sin^2\theta\,\dot\phi = J\,, \tag{2.78}$$

for integration constants E and J. The first of these is associated with energy conservation, and the second with angular-momentum conservation. We also have (2.72), which is like another first integral, giving

$$B\dot{t}^2 - B^{-1}\,\dot{r}^2 - r^2\dot{\theta}^2 - r^2\,\sin^2\,\dot{\phi}^2 = 1\,. \tag{2.79}$$

Of course one can plug (2.78) into (2.79).

It is easy to see, because of the symmetries of the problem, that just as in Newtonian mechanics, planetary orbits will lie in a plane. Because of the symmetries, we can, without loss of generality, take this to be the equatorial plane, $\theta = \frac{1}{2}\pi$. (The test of the assertion that the motion lies in a plane is to verify that the Euler-Lagrange equation for θ implies that $\ddot{\theta} = 0$ if we set $\theta = \frac{1}{2}\pi$ and $\dot{\theta} = 0$. In other words, if one starts the particle off with motion in the equatorial plane, it stays in the equatorial plane. We leave this as an exercise.)

If we proceed by taking $\theta = \frac{1}{2}\pi$ we have three first integrals for the three coordinates t, ϕ and r, and so the Euler-Lagrange equation for r is superfluous (since we already know its first integral). From (2.78) and (2.79) we therefore have

$$\Big(1 - \frac{2M}{r}\Big)\,\dot{t} = E\,, \qquad r^2\,\dot{\phi} = J\,, \qquad \dot{r}^2 = E^2 - \Big(1 + \frac{J^2}{r^2}\Big)\Big(1 - \frac{2M}{r}\Big)\,. \tag{2.80}$$

Note that the third equation has been obtained by substituting the first two into (2.79), and using also (2.74).

If we rewrite the third equation in (2.80) as

$$\dot{r}^2 + V(r) = E^2\,, \tag{2.81}$$

where

$$V(r) = \Big(1 + \frac{J^2}{r^2}\Big)\Big(1 - \frac{2M}{r}\Big)\,, \tag{2.82}$$

then it can be recognised as the equation for the one-dimensional motion of a particle of mass $m = 2$ in the effective potential $V(r)$. It is worth remarking that if we were instead solving the problem of planetary orbits in Newtonian mechanics, we would have $V(r) = J^2/r^2 - 2M/r$. The extra term 1 in the general relativistic expression (2.82) is just a shift in the zero point of the total energy E^2, corresponding to the rest mass of the particle. The important difference in general relativity is the extra term $-2MJ^2/r^3$ that comes from multiplying out the factors in (2.82). As we shall see, this term implies that the major axis of an elliptical planetary orbit will precess, rather than remaining fixed as it does in the Newtonian case. This is a testable prediction of general relativity, that has indeed been verified.

The nature of the orbits is determined by the shape of the effective potential $V(r)$ in equation (2.82). In particular, the crucial question is whether it has any critical points (where the derivative vanishes). From (2.82) we have

$$\frac{dV}{dr} = -\frac{2J^2}{r^3} + \frac{2M}{r^2} + \frac{6MJ^2}{r^4}\,, \tag{2.83}$$

and so $dV/dr = 0$ if

$$r = \frac{J^2 \pm J\sqrt{J^2 - 12M^2}}{2M}\,. \tag{2.84}$$

If $J^2 < 12M^2$ there are therefore no critical points, and the effective potential just plunges from $V = 1$ at $r = \infty$ to $V = -\infty$ as r goes to zero. There are no orbits possible in this case.

If $J^2 > 12M^2$, the effective potential $V(r)$ has two critical points, at radii $r_\pm$ given by

$$r_\pm = \frac{J^2 \pm J\sqrt{J}^2 - 12M^2}{2M}\,. \tag{2.85}$$

The effective potential attains a maximum at $r = r_-$, and a local minimum at $r = r_+$. There is a potential well in the region $r_0 \le r \le \infty$, where $V(r_0) = 1$ and r_0 occurs at some value greater than r_- and less than r_+. If the integration constant E (related to the energy of the particle) is appropriately chosen, we can then obtain orbits in which r oscillates between turning points that lie within the region $r_0 \le r \le \infty$.

The simplest case to consider is a circular orbit, achieved when $r = r_+$ so that we are sitting at the local minimum at the bottom of the potential well. This will be achieved if

$$E^2 = V(r_+)\,, \tag{2.86}$$

since then, as can be seen from (2.81), we shall have $\dot{r} = 0$ and so $r = r_+$ for all τ.

To analyse the orbits in general, it is useful, as in the Newtonian case, to introduce a new variable u instead of r, defined by

$$u = \frac{M}{r}\,. \tag{2.87}$$

We also define a rescaled angular momentum parameter $\widetilde{J}$, defined by

$$\widetilde{J} = \frac{J}{M}\,. \tag{2.88}$$

Since r and ϕ are both functions of τ it is then convenient to consider r, or the new variable u, as a function of ϕ. Elementary algebra shows that (2.81) gives rise to

$$\Big(\frac{du}{d\phi}\Big)^2 + (1-2u)(u^2+\widetilde{J}^{-2}) = E^2\,\widetilde{J}^{-2}\,. \tag{2.89}$$

In deriving this, we have used that $du/d\phi = \dot u/\dot\phi$, and we have substituted for $\dot\phi$ from (2.80).

The circular orbit discussed above corresponds, of course, to $du/d\phi = 0$, and so if we say this occurs at $u = u_0$, with energy given by E_0, we shall have

$$\widetilde{J}^{-2} = u_0(1-3u_0)\,, \tag{2.90}$$

coming from the condition that $dV/dr = 0$ at $r = r_0 = M/u_0$, and also

$$(1-2u_0)(u_0^2+\widetilde{J}^{-2}) = E_0^2\widetilde{J}^{-2}\,, \tag{2.91}$$

coming from (2.89) with $du/d\pi = 0$. Plugging (2.90) into (2.91), we can rewrite (2.91) as

$$E_0^2 = \frac{(1-2u_0)^2}{1-3u_0}\,. \tag{2.92}$$

Thus we have $\widetilde{J}$ and E_0 expressed in terms of the rescaled inverse radius u_0 of the circular orbit.

Having established the description of the circular orbit, we now consider an elliptical orbit. A convenient way to describe this is to think of keeping $\widetilde{J}$ the same, and u_0 the same, but changing to a different energy E. Simple algebra shows that (2.89) can then be rewritten as

$$\Big(\frac{du}{d\phi}\Big)^2 + (1-6u_0)(u-u_0)^2 - 2(u-u_0)^3 = (E^2-E_0^2)\,\widetilde{J}^{-2}\,. \tag{2.93}$$

Written in this way, it is manifest that we revert to the circular orbit with $u = u_0$ if we take the energy to be $E = E_0$.

The equation (2.93) is not easily solved analytically in terms of elementary functions. However, for our purposes it suffices to obtain an approximate solution. To do this we consider a slightly elliptical orbit, which can be described by writing

$$u = u_0\,(1+\epsilon\,\cos\omega\phi)\,. \tag{2.94}$$

Here ϵ is the eccentricity, and we are going to take $|\epsilon| << 1$. Plugging into (2.93), and working only up to order ϵ^2, we find

$$u_0^2\omega^2\sin^2\omega\phi + (1-6u_0)\epsilon^2\,\cos^2\omega\phi = (E^2-E_0^2)\widetilde{J}^{-2}\,. \tag{2.95}$$

Thus our trial solution does indeed work, up to order ϵ^2, if we have

$$\omega^2 = 1 - 6u_0\,, \qquad E^2 = E_0^2 + \widetilde{J}^2 u_0^2\,(1-6u_0)\,\epsilon^2\,. \tag{2.96}$$

The important equation here is the first one. From the form of the trial solution (2.94), we see that to go from one perihelion i.e. closest approach to the sun) to the next, the ϕ coordinate should advance through an angle $\Delta\phi$, where

$$\omega\,\Delta\phi = 2\pi\,. \tag{2.97}$$

Thus the azimuthal angle should advance by

$$\Delta\phi = \frac{2\pi}{\sqrt{1-6u_0}}\,. \tag{2.98}$$

If $\Delta\phi$ had been equal to 2π, the orbit would be a standard ellipse, returning to its perihelion after exactly a 2π rotation. Instead, we have the situation that $\Delta\phi$ is bigger than 2π, and so the azimuthal angle must advance by a bit more than 2π before the next perihelion. Thus the perihelion advances by an angle $\delta\phi$ per orbit, where

$$\delta\phi = \Delta\phi - 2\pi\,. \tag{2.99}$$

Now, we already noted that for a star such as the sun, the radius at its surface is hugely greater than the Schwarzschild radius for an object of the mass of the sun. Therefore since planetary orbits are certainly outside the sun (!), we have $r_0 >> M$, and so, from (2.87), we have $u_0 << 1$. We can therefore use a binomial approximation for $(1-6u_0)^{-1/2} = 1+3u_0+\cdots$ in (2.98), implying from (2.99) that the advance of the perihelion is approximated by

$$\delta\phi \approx 6\pi u_0 = \frac{6\pi M}{r_0}\,. \tag{2.100}$$

Clearly the effect will be largest for the planet whose orbital radius r_0 is smallest. This can be understood intuitively since it is experiencing the greatest gravitational attraction (it is deepest in the sun's gravitational potential), and so it experiences the greatest deviation from Newtonian gravity. In our solar system, it is therefore the planet Mercury that will exhibit the largest perihelion advance.

We can easily restore the dimensionful constants G and c in any formula at any time, just by appealing to dimensional analysis, i.e. noting that Newton's constant and the speed of light have dimensions

$$[G] = M^{-1}\,L^3\,T^{-2}\,, \qquad [c] = LT^{-1}\,. \tag{2.101}$$

Thus equation (2.100) becomes

$$\delta\phi \approx \frac{6\pi GM}{c^2 r_0}\,. \tag{2.102}$$

Putting in the numbers, this amounts to about 43 seconds of arc per century, for the advance of the perihelion of Mercury. Tiny though it is, this prediction has indeed been confirmed by observation, providing a striking vindication for Einstein's theory of general relativity.

3 Lie Groups and Algebras

3.1 Definition of a Group

Let us begin by defining a group. A group is a set A with the following additional structure:

1. A *law of composition* such that for each pair of elements a_1 and a_2, we get a third element denoted by $a_1 \circ a_2$.

2. The law of composition must be *associative*, i.e.

$$a_1 \circ (a_2 \circ a_3) = (a_1 \circ a_2) \circ a_3\,. \tag{3.1}$$

3. There must exist a *unit element* e, such that for any element a we have

$$e \circ a = a \circ e = a\,. \tag{3.2}$$

4. For every element a in A, there must exist an *inverse element* a^{-1} such that

$$a \circ a^{-1} = a^{-1} \circ a = e\,. \tag{3.3}$$

Some examples illustrating cases where there is a group structure, and where there isn't, are the following:

(a) The set of integers, Z, with addition as the law of composition, form a group. The identity element is 0, and the inverse of the integer n is the integer $-n$:

$$\begin{aligned} n + 0 &= 0 + n = n\,, \\ n + (-n) &= (-n) + n = 0\,. \end{aligned} \tag{3.4}$$

(b) The set of integers, with multiplication as the law of composition, do *not* form a group. An identity element exists ((i.e. 1), but the inverse of the integer n is $1/n$, which is not a member of the set of integers Z.

(c) The two integers $\{1, -1\}$ form a group under multiplication. This is called the group Z_2.

(d) The set $\mathbb{R}$ of all real numbers $-\infty < r < \infty$ forms a group under addition.

(e) The set $\mathbb{R}$ does *not* form a group under multiplication, since although the identity element exists (i.e. 1), not every element of $\mathbb{R}$ has an inverse; the inverse of 0 does not exist.

(f) The set $\mathbb{R}^+$ of all positive real numbers $0 < r < \infty$ forms a group under multiplication.

In all the examples (a), (c), (d) and (f) of groups, we have the feature that $a \circ b = b \circ a$ for any elements a and b. If all group elements satisfy this commutativity property, the group is said to be *abelian.* If there exist group elements for which $a \circ b \neq b \circ a$, the group is said to be *non-abelian.*

An example of a non-abelian group is the set of all real $n \times n$ matrices with non-vanishing determinant, where the law of composition is matrix multiplication. The condition of non-vanishing determinant ensures that every group element a has an inverse (the usual matrix inverse a^{-1}). However, matrix multiplication is non-commutative, and so in general $ab \neq ba$.

In our examples above, we have included discrete groups, where the number of elements is finite (as in case (c), where the group Z_2 has two elements) or infinite (as in case (a), where the group Z has a countable infinity of elements). We have also given an example of continuous groups, namely $\mathbb{R}$ in case (d), and $\mathbb{R}^+$ in case (f).

A finite group is said to be of order n if it has n elements. For example Z_2 is of order 2, while the group Z of integers under addition is of (countable) infinite order. All continuous groups are of uncountable infinite order. A useful way of characterising the "size" of a continuous group is by means of its *dimension.* The dimension of a continuous group is the number of independent continuous functions, or coordinates, that are needed in order to parameterise all the group elements. For example, for the group $\mathbb{R}$ of real numbers under addition, we need the single real parameter x, where $-\infty < x < \infty$.

One can form higher-dimension groups by taking tensor products of lower-dimension groups. For example, $\mathbb{R}^n$ (the n-fold tensor product of $\mathbb{R}$) is a group of dimension n, since we need n real parameters x_i, one for each copy of $\mathbb{R}$.

Note that we can also have groups for fields other than just the real numbers. For example, consider $\mathbb{C}$, the group of complex numbers under addition. To parameterise a point in $\mathbb{C}$ we need one complex number z, which we can write as $z = x + \mathrm{i}\, y$ in terms of two real numbers x and y. Thus we would say that $\mathbb{C}$ has complex dimension 1, and hence real dimension 2.

In this course, we shall be principally interested in continuous groups. In fact, we shall be interested in continuous groups with some extra structure, which are known as *Lie groups*.

3.2 Lie Groups

A *Lie group* of real dimension n is a set G that

1. Is a continuous group

2. Is an n-dimensional differentiable manifold

In other words, a Lie group is a continuous group in which the elements g in some patch can be parameterised by a set of n real numbers, or coordinates. In the overlap region between two patches, the first set of coordinates must be differentiable functions of the second set, and vice versa. This is exactly the notion of a differentiable manifold as we encountered earlier in these lectures.

The group combination law, and the taking of the inverse, should be smooth operations, i.e.

(a) The coordinates of the product $g'' = g g'$ of two group elements g and g' should be differentiable functions of the coordinates of g and g', provided that all three elements g, g' and g'' lie in a patch where a common set of coordinates can be used.

(b) The coordinates of g^{-1} should be differentiable functions of the coordinates of g, whenever g and g^{-1} are covered by the same coordinate patch.

As in our earlier discussion of differentiable manifolds, we will encounter examples of Lie groups where more than one coordinate patch is needed in order to cover the whole group. In fact, this is the case in general; only in exceptional cases, such as $\mathbb{R}^n$, can one use a single coordinate patch to cover the entire group.

A simple example of a Lie group where more than one coordinate patch is required is provided by the group $U(1)$ of all unit-modulus complex numbers. Obviously, such numbers g form a group under multiplication (since if g_1 and g_2 have unit modulus, then so does $g_1 g_2$). We can view the elements g as points on the unit circle $x^2 + y^2 = 1$ in the complex plane, where $z = x + \mathrm{i}\, y$. This shows that the group $U(1)$ of unit-modulus complex numbers is *isomorphic* to the circle, S^1. That is to say, there exists a 1-1 map between elements of $U(1)$ and elements of S^1, which preserves the group combination law.

Locally, therefore, we can parameterise $U(1)$ by means of a coordinate θ, by writing group elements g as

$$g = e^{\mathrm{i}\,\theta}\,, \qquad 0 \le \theta < 2\pi\,. \tag{3.5}$$

We now get into all the familiar issues that we encoutered in our earlier discussion of manifolds; we cannot use θ to cover all of S^1, since it suffers a discontinuous jump from 2π to 0 as one crosses the point $(x,y)=(1,0)$ on the circle. As in section 1.3.1, we can introduce a second coordinate $\tilde\theta$ that starts from $\tilde\theta = 0$ at $(x,y)=(-1,0)$, and cover S^1 in patches using θ for all points except $(x,y)=(1,0)$, and $\tilde\theta$ for all points except $(x,y)=(-1,0)$. Since, as in (1.4), we have $\tilde\theta = \theta+\pi$ in the upper semicircular overlap $(x>0)$, and $\tilde\theta=\theta-\pi$ in the lower semicircular overlap $(x<0)$, it follows that we have

$$e^{\mathrm{i}\,\tilde\theta} = -e^{\mathrm{i}\,\theta} \tag{3.6}$$

in the entire overlap region. One easily verifies that all the conditions of differentiablity, *etc.*, are satisfied.

It will be useful at this stage to enumerate examples of some of the most common groups that one encounters in physics and mathematics. Before doing so, we give one further definition:

A *subgroup* H of a group G is a subset of G for which the following properties hold:

1. The identity element e of G is contained in H

2. If h_1 and h_2 are any elements of H, then $h_1\circ h_2$ is an element of H, where $\circ$ is the group composition law of G.

3. If h belongs to H, then so does h^{-1}, where h^{-1} means the inverse of h according to the group inverse law of G.

If H is a subgroup of G, this is denoted by

$$H \subset G\,. \tag{3.7}$$

3.2.1 General linear group, $GL(n,\mathbb{R})$

Let $M(n,\mathbb{R})$ denote the set of all real $n\times n$ matrices with non-vanishing determinant. As we have already remarked, these matrices form a group under multiplication, which is called the *General linear group.* The requirement of non-vanishing determinant ensures that each matrix A has an inverse, A^{-1}. Clearly, the requirement of non-vanishing determinant is compatible with the group combination law, since if $\det A \neq 0$ and $\det B \neq 0$ then

$$\det(AB) = (\det A)(\det B) \neq 0\,. \tag{3.8}$$

The dimension of $GL(n,\mathbb{R})$ is equal to the number of independent components of a general $n\times n$ real matrix, namely n^2. Obviously, the requirement of non-vanishing determinant places a restriction on the parameters, but since it is in the form of an inequality ($\det A \neq 0$) rather than an equality, it does not reduce the number of parameters needed to characterise a general such matrix.

One can also consider the complex analogue, $GL(n,\mathbb{C})$, of $n\times n$ complex matrices of non-vanishing determinant. Now, we need n^2 complex parameters to specify a general $GL(n,\mathbb{C})$ matrix, and so this group has complex dimension n^2, implying real dimension $2n^2$.

3.2.2 Special linear group, $SL(n,\mathbb{R})$

Many of the groups that arise in physics and mathematics are subgroups of $GL(n,\mathbb{R})$ or $GL(n,\mathbb{C})$. The simplest example is the *Special linear group*, $SL(n,\mathbb{R})$. This is defined to be the set of all real $n\times n$ matrices A with unit determinant, $\det A = 1$. Obviously this is a subgroup of $GL(n,\mathbb{R})$. It is also obvious that the requirement $\det A = 1$ is compatible with the group combination law (matrix multiplication), since if A and B are any two real matrices with unit determinant, we have

$$\det(AB) = (\det A)(\det B) = 1\,. \tag{3.9}$$

The condition $\det A = 1$ imposes 1 real equation on the n^2 parameters of a $GL(,\mathbb{R})$ matrix, and so we have

$$\dim SL(n,\mathbb{R}) = n^2 - 1\,. \tag{3.10}$$

In a similar manner, we can define $SL(n,\mathbb{C})$, as the subgroup of $GL(n,\mathbb{C})$ comprising all $n\times n$ complex matrices with unit determinant. This will have real dimension

$$\dim SL(n,\mathbb{C}) = 2n^2 - 2\,, \tag{3.11}$$

since the condition $\det A = 1$ now imposes one complex equation, or in other words 2 real equations, on the $2n^2$ real parameters of $GL(n,\mathbb{C})$.

3.2.3 Orthogonal group, $O(n, \mathbb{R})$

These groups are very important in physics, since, amongst other things, they describe rotations in n-dimensional Euclidean space. $O(n, \mathbb{R})$ is defined as the subgroup of $GL(n, \mathbb{R})$ comprising all real $n \times n$ matrices A for which

$$A\, A^T = \mathbb{1}\,, \tag{3.12}$$

where A^T denotes the transpose of the matrix A. Obviously these have non-vanishing determinant, since

$$\det(AA^T) = (\det A)(\det A^T) = (\det A)^2 = \det \mathbb{1} = 1\,, \tag{3.13}$$

and hence $\det A = \pm 1$. Furthermore, it is obvious that the orthogonality condition (3.12) is compatible with the group multiplication law, since if A and B are orthogonal matrices, then so is (AB):

$$(AB)(AB)^T = ABB^T A^T = AA^T = \mathbb{1}\,. \tag{3.14}$$

Furthermore, if A is orthogonal then so is A^{-1}, and so the inverse also belongs to the subset.

Usually, unless specified otherwise, it is assumed that the orthogonal groups are composed of *real* orthogonal matrices, and so $O(n, \mathbb{R})$ is commonly written simply as $O(n)$.

The dimension of $O(n, \mathbb{R})$ can be calculated by counting the number of independent equations that the orthogonality condition

$$AA^T - \mathbb{1} = 0 \tag{3.15}$$

imposes on a general $n \times n$ real matrix. Since AA^T is a symmetric matrix,

$$(AA^T)^T = (A^T)^T A^T = AA^T\,, \tag{3.16}$$

it follows that (3.15) contains the same number of independent equations as there are in an $n \times n$ symmetric matrix, namely $\frac{1}{2}n(n+1)$. Therefore we have

$$\dim O(n, \mathbb{R}) = n^2 - \tfrac{1}{2}n(n+1) = \tfrac{1}{2}n(n-1)\,. \tag{3.17}$$

Note that we can also consider the subgroup $SO(n, \mathbb{R})$ of $O(n, \mathbb{R})$ comprising all $n \times n$ orthogonal matrices with unit determinant. We saw above that the orthogonality condition implied $\det A = \pm 1$, and so now we are restricting to the subset of orthogonal matrices A for which $\det A = +1$. Obviously this is compatible with the group multiplication law, and the group inverse. Since there are no additional continuous equations involved in imposing the restriction $\det A = +1$, the dimension of $SO(n,\mathbb{R})$ is the same as the dimension of $O(n,\mathbb{R})$:

$$\dim SO(n,\mathbb{R}) = \tfrac{1}{2}n(n-1)\,. \tag{3.18}$$

Note that $SO(n,\mathbb{R})$ is a subgroup of $SL(n,\mathbb{R})$, but $O(n,\mathbb{R})$ is not.

3.2.4 Unitary group, $U(n)$

The unitary group $U(n)$ is defined as the subgroup of $GL(n,\mathbb{C})$ comprising all complex $n\times n$ matrices A that are unitary:

$$AA^\dagger = \mathbb{1}\,, \tag{3.19}$$

where $A^\dagger \equiv (A^T)^*$ is the hermitean conjugate of A (i.e. the complex conjugate of the transpose). Again, one easily checks that the unitary condition is conpatible with the matrix multiplication law of group combination, and with the inverse. By counting the number of independent equations implied by the restriction (3.19), one straightforwardly sees that the real dimension of $U(n)$ is given by

$$\dim U(n) = n^2\,. \tag{3.20}$$

Note that the case $n=1$ corresponds to complex numbers of unit modulus; we already met the group $U(1)$ in our earlier discussion.

3.2.5 Special unitary group, $SU(n)$

$U(n)$ matrices A satisfy $AA^\dagger = \mathbb{1}$, and so

$$\det(AA^\dagger) = (\det A)(\det A^T)^* = (\det A)(\det A)^* = |\det A|^2 = 1\,, \tag{3.21}$$

meaning that $\det A$ is a complex number of unit modulus. If we impose the further restriction

$$\det A = 1\,, \tag{3.22}$$

this says that the *phase* of the complex number is 0, and therefore it imposes 1 further real condition on the components of the $U(n)$ matrix. Since the condition $\det A = 1$ is obviously compatible with the law of multiplication and the group inverse, we see that the group of special unitary $n\times n$ matrices, denoted by $SU(n)$, is a subgroup of $U(n)$ with real dimension given by

$$\dim SU(n) = n^2 - 1\,. \tag{3.23}$$

3.2.6 Some properties of $SU(2)$

We have already seen in detail for the ablian group $U(1)$ how it is isomorphic to the circle, S^1. The general $U(1)$ group element g is written as $g = e^{i\theta}$, where θ is the coordinate on S^1, and all the usual caveats about needing to cover S^1 in patches apply.

Now, let us look at a slightly more complicated example, namely the non-abelian group $SU(2)$. For many purposes $SU(2)$ is a very useful example to study, because it encapsulates many of the generic features of any non-abelian Lie group. For now, we shall focus in particular on the global structure of the $SU(2)$ group manifold. As we shall see, it is isomorphic to the 3-sphere S^3.

To begin, consider the group $U(2)$ of unitary 2×2 matrices, whose elements we may write as

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \tag{3.24}$$

where

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \bar{a} & \bar{c} \\ \bar{b} & \bar{d} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{3.25}$$

In other words the complex numbers (a, b, c, d) are subject to the conditions

$$a\bar{a} + b\bar{b} = 1\,, \qquad c\bar{c} + d\bar{d} = 1\,, \qquad a\bar{c} + b\bar{d} = 0\,. \tag{3.26}$$

The first two equations are real, and so each imposes 1 real condition on the 4 complex numbers. The third equation is complex, and so it imposes 2 further real conditions, making 4 real conditions in total. Thus we are leaft with $8 - 4 = 4$ real numbers characterising the general $U(2)$ matrix, in accordance with our earlier counting.

Now we impose the further condition $\det A = 1$, in order to restrict to the subgroup $SU(2)$. This implies the further condition

$$ad - bc = 1\,. \tag{3.27}$$

(This is only one additional real condition, since the previous $U(2)$ conditions already ensured that $\det A$ must have unit modulus.) Thus $SU(2)$ has dimension $8 - 4 - 1 = 3$. Multiplying (3.27) by $\bar{c}$, and using (3.26), we can easily see that

$$c = -\bar{b}\,, \qquad d = \bar{a}\,, \tag{3.28}$$

and in fact that these two equations, together with (3.27), imply the three equations in (3.26). The upshot, therefore, is that we have parameterised the most general $SU(2)$ matrix in the form

$$A = \begin{pmatrix} a & b \\ -\bar{b} & \bar{a} \end{pmatrix}, \tag{3.29}$$

where

$$a\bar{a} + b\bar{b} = 1\,. \tag{3.30}$$

Thus A is written in terms of the two complex numbers a and b, subject to the single real constraint (3.30).

If we now write $a = x_1 + \mathrm{i}\, x_2$ and $b = x_3 + \mathrm{i}\, x_4$ in terms of the four real numbers (x_1, x_2, x_3, x_4), we see that the constraint (3.30) is

$$x_1^2 + x_2^2 + x_3^2 + x_4^2 = 1\,. \tag{3.31}$$

This can be interpreted as the the restriction of the coordinates (x_1, x_2, x_3, x_4) on $\mathbb{R}^4$ to the unit 3-sphere. Since we have established a 1-1 mapping between points in S^3 and points in $SU(2)$, and the mapping is clearly compatible with the group combination rule (matrix multiplication), we have therefore shown that $SU(2)$ and S^3 are isomorphic,

$$SU(2) \cong S^3\,. \tag{3.32}$$

Having seen the isomorphisms $U(1) \cong S^1$ and $SU(2) \cong S^3$, one might wonder whether any of the other Lie groups are isomorphic to spheres. In fact one can show that S^1 and S^3 are the *only* spheres that are isomorphic to group manifolds. We shall return to this point later.

3.3 The Classical Groups

At this point, it is appropriate to give a complete description of all the so-called *Classical* Lie groups. To do so, recall from section (1.4.1) that we introduced the notion of a set of basis vectors E_i on a vector space. (We shall use indices $i, j, \ldots$ here to label the basis vectors, rather than $a, b, \ldots$ as in section (1.4.1).) We may now define the various classical groups in terms of transformations between bases for an n-dimensional vector space V, together with some possible additional structure imposed on the vector space.

3.3.1 The General Linear Group

This group requires the least structure, and is defined purely in terms of transformations of the vector space itself. Thus we may define a new basis E'_i, related to E_i by

$$E'_i = A_i{}^j\, E_j\,, \tag{3.33}$$

for some set of n^2 quantities $A_i{}^j$, which may be thought of as the components of an $n \times n$ matrix A with rows labelled by i and columns labelled by j.

In order that the change of basis be non-singular, so that we can invert to get E_i expressed in terms of E'_j, it must be that $\det A \neq 0$. Thus we define $GL(n, \mathbb{R})$ as the group of all real $n \times n$ matrices with non-vanishing determinant; it is the group of arbitrary non-singular changes of basis for a real n-dimensional vector space.

In a similar fashion, we can define $GL(n,\mathbb{C})$, comprising $n \times n$ complex matrices A with $\det A \neq 0$, as the group of arbitrary non-singular transformations of an n-dimensional complex vector space.

3.3.2 The Special Linear Group

To define the special linear groups, we form the n-fold antisymmetric tensor product of the vector space V. To do this we proceed in a manner analogous to our discussion of differential p-forms, and take the tensor product of p copies of V, and then perform a total antisymmetrisation over the indices labelling the basis vectors:

$$E_{i_1} \wedge \cdots \wedge E_{i_p} \equiv E_{i_1} \otimes \cdots \otimes E_{i_p} + \text{even permutations} - \text{odd permutations}\,. \tag{3.34}$$

Then, we define Ω as the n-fold antisymmetric product:

$$\Omega \equiv E_1 \wedge E_2 \wedge \cdots \wedge E_n\,. \tag{3.35}$$

This is called the *volume element* of the vector space. Clearly we can write this as

$$\Omega = \frac{1}{n!}\, \varepsilon^{i_1 i_2 \cdots i_n}\, E_{i_1} \wedge E_{i_2} \wedge \cdots \wedge E_{i_n}\,, \tag{3.36}$$

where we define $\varepsilon^{12\cdots n} = +1$, and conversely

$$E_{i_1} \wedge E_{i_2} \wedge \cdots \wedge E_{i_n} = \varepsilon_{i_1 i_2 \cdots i_n}\, \Omega\,, \tag{3.37}$$

where we also define $\varepsilon_{12\cdots n} = +1$.

If we again perform a linear transformation to a new basis E'_i, given by

$$E'_i = A_i{}^j\, E_j\,, \tag{3.38}$$

then the quantity Ω transforms to

$$\begin{aligned}
\Omega' &= \frac{1}{n!}\, \varepsilon^{i_1 i_2 \cdots i_n}\, E'_{i_1} \wedge E'_{i_2} \wedge \cdots \wedge E'_{i_n} \\
&= \frac{1}{n!}\, \varepsilon^{i_1 i_2 \cdots i_n}\, A_{i_1}{}^{j_1}\, A_{i_2}{}^{j_2} \cdots A_{i_n}{}^{j_j}\, E_{j_1} \wedge E_{j_2} \wedge \cdots \wedge E_{j_n} \\
&= \frac{1}{n!}\, \varepsilon^{i_1 i_2 \cdots i_n}\, A_{i_1}{}^{j_1}\, A_{i_2}{}^{j_2} \cdots A_{i_n}{}^{j_j}\, \varepsilon_{j_1 j_2 \cdots j_n}\, \Omega\,, \\
&= (\det A)\, \Omega\,.
\end{aligned} \tag{3.39}$$

We may therefore define the subsets of $GL(n,\mathbb{R})$ or $GL(n,\mathbb{C})$ matrices that preserve the volume element Ω, i.e. for which

$$\Omega' = (\det A)\,\Omega = \Omega\,, \tag{3.40}$$

by imposing the requirement that $\det A = 1$. Thus we have the groups $SL(n,\mathbb{R})$ and $SL(n,\mathbb{C})$ of volume-preserving linear transformations on the n-dimensional real or complex vectore space.

3.3.3 Metrics on Vector Spaces

The remaining classical groups are defined by introducing an additional structure on the vector space V, namely a *metric*. This is closely analogous to our discussion of metrics in differential geometry, with the main difference here being that we do not necessarily insist on having a *symmetric* metric.

We define a metric on the vector space V as a function on V which provides a rule for associating a number f to each pair of vectors u and v in V:

$$(u,v) = f\,. \tag{3.41}$$

If V is a real vector space then f is real, whilst if v is a complex vector space then f is in general complex.

The metric is required to satisfy the following properties:

$$\begin{aligned}
(u, v+w) &= (u,v) + (u,w)\,,\\
(u+v, w) &= (u,w) + (v,w)\,,\\
(u, \lambda\, v) &= \lambda\,(u,v)\,,
\end{aligned} \tag{3.42}$$

for any vectors (u,v,w), and for any number λ. In the case of a real vector space, λ is real, whilst for a complex vector space λ is complex. There is one further condition, which takes one of two possible forms. We have either *Bilinear Metrics* or *Sesquilinear Metrics*, which satisfy one or other of the following two conditions:

$$\textbf{Bilinear metrics:} \qquad (\lambda\, u, v) = \lambda\,(u,v)\,, \tag{3.43}$$

$$\textbf{Sesquilinear metrics:} \qquad (\lambda\, u, v) = \bar\lambda\,(u,v)\,. \tag{3.44}$$

Note that the possibility of a sesquilinear metric arises only in the case of a complex vector space, whilst bilinear metrics can arise either for real or complex vector spaces.

The components of the metric, with respect to a basis E_i, are defined by

$$g_{ij} \equiv (E_i, E_j)\,. \tag{3.45}$$

For any pair of vector u and v, expanded in terms of components as $u = u^i E_i$, $v = v^i E_i$, we have

$$\begin{aligned} \textbf{Bilinear:} \qquad & (u,v) = (u^i E_i, v^j E_j) = u^i v^j (E_i, E_j) = g_{ij} u^i v^j\,, && (3.46)\\ \textbf{Sesquilinear:} \qquad & (u,v) = (u^i E_i, v^j E_j) = \bar u^i v^j (E_i, E_j) = g_{ij} \bar u^i v^j\,, && (3.47)\\ & && (3.48) \end{aligned}$$

Under a change of basis $E_i' = A_i{}^j\, E_j$ we have $g'_{ij} = (E_i', E_j') = (A_i{}^k E_k, A_j{}^\ell E_\ell)$ and hence

$$\begin{aligned} \textbf{Bilinear:} \qquad & g'_{ij} = A_i{}^k\, A_j{}^\ell\, g_{k\ell}\,, && (3.49)\\ \textbf{Sesquilinear:} \qquad & g'_{ij} = \bar A_i{}^k\, A_j{}^\ell\, g_{k\ell}\,. && (3.50) \end{aligned}$$

We can now define subgroups of $GL(n,{\rm I\!R})$ or $GL(n,{\rm C\!\!\!I})$ matrices by choosing a metric structure on the vector space V, and requiring that the $GL(n)$ matrices leave the metric g_{ij} invariant.[9] Thus we have metric-preserving subgroups if

$$\begin{aligned} \textbf{Bilinear:} \qquad & A_i{}^k\, A_j{}^\ell\, g_{k\ell} = g_{ij}\,, && (3.51)\\ \textbf{Sesquilinear:} \qquad & \bar A_i{}^k\, A_j{}^\ell\, g_{k\ell} = g_{ij}\,. && (3.52) \end{aligned}$$

We must verify that $GL(n)$ matrices subject to these conditions do indeed form a group; namely that products of such matrices also satisfy the metric-preserving condition, and that the inverse of any such matrix also satisfies the condition. For example, for the bilinear case, if we suppose that A and B satisfy (3.51), then we shall have

$$\begin{aligned} (AB)_i{}^k\, (AB)_j{}^\ell\, g_{k\ell} &= A_i{}^m\, B_m{}^k\, A_j{}^n\, B_n{}^\ell\, g_{k\ell}\\ &= A_i{}^m\, A_j{}^n\, g_{mn}\\ &= g_{ij}\,, \end{aligned} \tag{3.53}$$

which proves that (AB) satisfies (3.51) too. Multiplying (3.51) by $(A^{-1})_m{}^i\, (A^{-1})_n{}^j$ gives

$$g_{mn} = (A^{-1})_m{}^i\, (A^{-1})_n{}^j\, g_{ij}\,, \tag{3.54}$$

which shows that A^{-1} also satisfies (3.51). The proofs for the sequilinear case are almost identical.

[9] We shall adopt the convention that when we refer simply to $GL(n)$, we mean in general that this could be $GL(n,{\rm I\!R})$ or $GL(n,{\rm C\!\!\!I})$.

We can now classify all the possible metric-preserving groups by classifying all the possible *canonical forms* for non-singular metrics g_{ij}. In other words, we want to enumerate all the genuinely inequivalent possible choices for g_{ij}, modding out by equivalences such as mere relabellings of indices, or whatever.

3.3.4 Canonical Forms for Bilinear Metrics

In the bilinear case we can write an arbitrary metric as a sum of its symmetric and anti-symmetric parts:

$$g_{ij} = g_{(ij)} + g_{[ij]}\,, \tag{3.55}$$

where, as usual, we use the notation

$$g_{(ij)} \equiv \tfrac{1}{2}(g_{ij} + g_{ji})\,, \qquad g_{[ij]} \equiv \tfrac{1}{2}(g_{ij} - g_{ji})\,. \tag{3.56}$$

We can then discuss the symmetric and antisymmetric parts separately.

For the symmetric part, we note that under a change of basis $E_i' = S_i{}^j\, E_j$, the metric transforms to

$$g_{ij}' = S_i{}^k\, S_j{}^\ell\, g_{k\ell}\,, \tag{3.57}$$

and for a symmetric g_{ij} one can always find a choice of $S_i{}^j$ that diagonalises g_{ij}'. By rescaling the new basis vectors, we can then make these diagonal entries equal to $+1$ or -1. Thus in general we can assume that we have

$$g_{(ij)} = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & -1 & & \\ & & & & \ddots & \\ & & & & & -1 \end{pmatrix}, \tag{3.58}$$

where there are p entries $+1$ on the upper part of the diagonal, and q entries -1 on the remaining lower part of the diagonal, where $p + q = n$. For much of the time, we shall be concerned with the case where $p = n$ and $q = 0$, so that $g_{ij} = \delta_{ij}$.

If g_{ij} is antisymmetric, then in order to be non-singular it must be that the dimension n is *even*. To see this, suppose $g^T = -g$, and take the determinant:

$$\det g^T = \det(-g) = \det(-\mathbb{1})\, \det g = (-1)^n\, \det g\,. \tag{3.59}$$

But from the properties of the determinant we have $\det g^T = \det g$, and thus we conclude that if $\det g \neq 0$ we must have $(-1)^n = 1$, and hence n must be even, $n = 2m$. By an appropriate change of basis the antisymmetric matrix can be cast into a block-diagonal form:

$$g_{[ij]} = \begin{pmatrix} 0 & \lambda_1 & & & & & \\ -\lambda_1 & 0 & & & & & \\ & & 0 & \lambda_2 & & & \\ & & -\lambda_2 & 0 & & & \\ & & & & \ddots & & \\ & & & & & 0 & \lambda_m \\ & & & & & -\lambda_m & 0 \end{pmatrix} . \tag{3.60}$$

By rescaling the new basis vectors, we can choose $\lambda_i = 1$ for all i, giving

$$g_{[ij]} = \begin{pmatrix} 0 & 1 & & & & & \\ -1 & 0 & & & & & \\ & & 0 & 1 & & & \\ & & -1 & 0 & & & \\ & & & & \ddots & & \\ & & & & & 0 & 1 \\ & & & & & -1 & 0 \end{pmatrix} . \tag{3.61}$$

Thus $g_{[ij]}$ has m eigenvalues $+1$, and m eigenvalues -1.

Alternatively, by permuting the basis elements the antisymmetric metric can be cast into the anti-diagonal form

$$g_{[ij]} = \begin{pmatrix} & & & & & 1 \\ & & & & \iddots & \\ & & & 1 & & \\ & & -1 & & & \\ & \iddots & & & & \\ -1 & & & & & \end{pmatrix} . \tag{3.62}$$

3.3.5 Canonical Forms for Sesquilinear Metrics

These arise only for complex vector spaces. We can write a general sesquilinear metric as a sum of its Hermitean and anti-Hermitean parts:

$$g_{ij} = g_{ij}^{(H)} + g_{ij}^{(AH)} , \tag{3.63}$$

where

$$g_{ij}^{(H)} \equiv \tfrac{1}{2}(g_{ij} + \bar{g}_{ji}) , \qquad g_{ij}^{(AH)} \equiv \tfrac{1}{2}(g_{ij} - \bar{g}_{ji}) . \tag{3.64}$$

However, since multiplication of an anti-Hermitean matrix by i gives an Hermitean matrix, there is effectively no distinction to be made between the two cases, and therefore we may assume without loss of generality that the metric is Hermitean. By change of basis, such a matrix can be diagonalised, and after rescaling its diagonal entries can be made to be ± 1. Thus we may take the sesquilinear metric g_{ij} to have the canonical form

$$g_{ij} = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & -1 & & \\ & & & & \ddots & \\ & & & & & -1 \end{pmatrix} , \tag{3.65}$$

where there are p entries $+1$ in the upper diagonal, and q entries -1 in the remaining lower part of the diagonal, where $p + q = n$. Commonly, we shall be concerned with the case where $q = 0$, and hence $g_{ij} = \delta_{ij}$.

3.3.6 The Metric-Preserving Groups

Having established the canonical forms of the possible non-degenerate metrics on the vector spaces, we see that there are three classes of examples of metric-preserving groups. These are:

Metric	**Group**	**Notation**
Bilinear Symmetric $n = p + q$	Orthogonal	$O(p, q; \mathbb{R})$, $O(p, q; \mathbb{C})$
Bilinear Antisymmetric $n = 2m$	Symplectic	$Sp(2m; \mathbb{R})$, $Sp(2m; \mathbb{C})$
Sesquilinear Symmetric $n = p + q$	Unitary	$U(p, q; \mathbb{C})$

We shall now describe each of these three classes of metric-preserving groups.

Orthogonal Groups:

The bilinear symmetric metric has the form (3.58). If we consider the case $p = n$, $q = 0$, for which $g_{ij} = \delta_{ij}$, the metric-preserving condition is

$$A_i{}^k \, A_j{}^\ell \, \delta_{k\ell} = \delta_{ij} \, , \tag{3.66}$$

which in matrix language reads

$$AA^T = \mathbb{1}\,. \tag{3.67}$$

This is just the condition for orthogonal matrices that we discussed previously. When the matrices are real, we abbreviate the general notation $O(n,0;\mathbb{R})$ to simply $O(n)$. The group $O(n)$ is called the the *compact form* of the orthogonal group in n dimensions. This means, as we shall discuss later, that the group manifold has finite volume. The various possibilities $O(p,q;\mathbb{R})$ with p and q both non-zero correspond to different non-compact forms of the orthogonal group in $n = p+q$ dimensions. Again, when we are talking about the real case we usually omit the $\mathbb{R}$, and just call it $O(p,q)$. The non-compact forms have group manifolds of infinite volume.

As well as the $q = 0$ compact form, for which $O(n)$ is just the rotation group in n dimensions, the case when $p = n-1$, $q = 1$ also arises commonly in physics; this is the *Lorentz group* in n dimensions, which is the group of symmetries of Minkowski spacetime in special relativity. Thus, the usual four-dimensional Lorentz group is $O(3,1)$.

We saw already, by counting the number of conditions implied by (3.67), that $O(n)$ has dimension $\frac{1}{2}n(n-1)$. The counting is identical for all the non-compact forms. For the complex case, there is just a doubling of the real dimension, since every component that was previously real can now be complex. Thus we have

$$\begin{aligned} \mathrm{Dim}(O(p,q;\mathbb{R})) &= \tfrac{1}{2}n(n-1)\,, \\ \mathrm{Dim}(O(p,q;\mathbb{C})) &= n(n-1)\,, \end{aligned} \tag{3.68}$$

where $n = p+q$.

For all the orthogonal groups one can see by taking the determinant of the defining equation (3.51) that $\det A = \pm 1$ for all matrices. One can always impose the further condition $\det A = +1$, yielding the special orthogonal groups $SO(p,q;\mathbb{R})$ and $SO(p,q;\mathbb{C})$ as subgroups of $O(p,q;\mathbb{R})$ and $O(p,q;\mathbb{C})$ respectively. They have the same dimensions as the orthogonal groups, since no continuous parameters are lost when one imposes the sign choice $\det A = +1$.

Symplectic Groups:

For these, the canonical form of the metric is given by (3.62), with the matrices satisfying

$$A_i{}^k\, A_j{}^\ell\, g_{k\ell} = g_{ij}\,. \tag{3.69}$$

Since the left-hand side is automatically antisymmetric for any A (and so, of course, is the right-hand side), it follows that this equation imposes $\frac{1}{2}n(n-1)$ constraints on the n^2

components of an arbitrary matrix A. Thus we have the real dimensions

$$\begin{aligned} \mathrm{Dim}(Sp(n;\mathbb{R})) &= \tfrac{1}{2}n(n+1)\,, \\ \mathrm{Dim}(Sp(n;\mathbb{C})) &= n(n+1)\,, \end{aligned} \tag{3.70}$$

where $n = 2m$. The symplectic groups as defined here are all non-compact.

One can also impose a unit-determinant condition, giving subgroups $SSp(n;\mathbb{R})$ and $SSp(n;\mathbb{C})$ of $Sp(n;\mathbb{R})$ and $Sp(n;\mathbb{C})$ respectively. Again, since the $Sp(n;\mathbb{R})$ and $Sp(n;\mathbb{C})$ matrices already satisfied $\det A = \pm 1$, the imposition of the unit-determinant condition implies no loss of continuous parameters, and so the dimensions of $SSp(n;\mathbb{R})$ and $SSp(n;\mathbb{C})$ are again $\frac{1}{2}n(n+1)$ and $n(n+1)$ respectively.

Unitary Groups:

The canonical form of the sesquilinear symmetric metric is given by (3.65). If we consider the case $p = n$, $q = 0$, then $g_{ij} = \delta_{ij}$, and the metric-preserving condition (3.52) just becomes

$$\bar{A}_i{}^k\, A_j{}^\ell\, \delta_{k\ell} = \delta_{ij}\,, \tag{3.71}$$

which in matrix notation reads $\bar{A}A^T = \mathbb{1}$. By complex conjugating, this becomes

$$AA^\dagger = \mathbb{1}\,, \tag{3.72}$$

which is just the unitary condition that we met previously when describing the matrices $U(n)$. This is the compact form of the unitary group; the more general possibilities $U(p,q;\mathbb{C})$ (which we usually just write as $U(p,q)$) with $p + q = n$ are non-compact forms of $U(n)$. They all have real dimension given by

$$\mathrm{Dim}(U(p,q)) = n^2\,, \tag{3.73}$$

where $n = p + q$, as we discussed previously for $U(n)$.

One can impose the unit-determinant condition, yielding the subgroup $SU(p,q)$ of $U(p,q)$, which has

$$\mathrm{Dim}(SU(p,q)) = n^2 - 1\,, \qquad n = p + q\,. \tag{3.74}$$

We close this section with a few further remarks:

(1) We have considered groups defined for vector spaces over the real numbers and the complex numbers. One can also consider vector spaces over the field of quaternionic numbers.[10] Some of the multiplication operations must be handled with care, since quaternion multiplication is itself non-commutative. Groups based on quaternion-valued matrices can be defined.

(**2**) One can consider matrices that are both unitary *and* symplectic. Thus we may define the so-called *Unitary-symplectic* group $USp(2m)$ of matrices that are simultaneously in $U(2m)$ and $Sp(2m;\mathbb{C})$:

$$USp(2m) = U(2m) \cap Sp(2m;\mathbb{C})\,. \tag{3.75}$$

(**3**) Some of the classical groups of low dimension are isomorphic, or homomorphic.[11] Some examples are:

$$\begin{aligned}
&\underline{\textbf{Dimension 3}}: \quad && SU(2) \cong SO(3) \cong USp(2)\\
& && SU(1,1) \cong SO(2,1) \cong Sp(2,;\mathbb{R}) \cong SL(2,\mathbb{R})\\
&\underline{\textbf{Dimension 6}}: \quad && SO(4) \cong SU(2) \times SU(2)\\
& && SO(3,1) \cong SL(2;\mathbb{C})\\
& && SO(2,2) \cong SL(2;\mathbb{R}) \times SL(2;\mathbb{R})\\
&\underline{\textbf{Dimension 10}}: \quad && SO(5) \cong USp(4)\\
& && SO(3,2) \cong Sp(4;\mathbb{R})\\
&\underline{\textbf{Dimension 15}}: \quad && SO(6) \cong SU(4)\\
& && SO(4,2) \cong SU(2,2)\\
& && SO(3,3) \cong SL(4,\mathbb{R})\,.
\end{aligned} \tag{3.76}$$

[10] These are ordered pairs of complex numbers, generalising the description of complex numbers as ordered pairs of real numbers. See my lecture notes for 615 Mathematical Methods, for a detailed discussion of the four division algebras; real numbers, complex numbers, quaternions and octonions.

[11] Two groups are homomorphic if there is a mapping between them that preserves the group combination law, but the mapping is not 1-1.

3.4 Lie Algebras

3.4.1 Introduction

So far, we have been looking at the structure of the entire set of matrices that form a group under multiplication. For many purposes, it is not necessry to study the entire group–it is sufficient to look at the elements in the neighbourhood of the identity.

The local structure of the group can be probed by looking at elements of the form

$$g = \mathbb{1} + \epsilon\, X\,, \tag{3.77}$$

where $|\epsilon| << 1$, and so we can work just to order ϵ. The object X is called a *generator* of the group. The local structure in the neighbourhood of the identity is called the *Lie Algebra*.

Commonly, we denote a Lie group by the symbol G, and its associated Lie algebra by $\mathcal{G}$.

The elements of the full group can be obtained by exponentiating the generators of the Lie algebra. For a compact group (where the group manifold has a finite volume), one usually takes the generators to be Hermitean matrices,

$$X_a = X_a^\dagger\,, \qquad a = 1, \ldots, \dim G\,. \tag{3.78}$$

The group elements can then be obtained by exponentiation:

$$g = \exp(\mathrm{i}\, \alpha^a\, X_a)\,. \tag{3.79}$$

Here, the quantities α^a are parameters, which can be thought of as coordinates on the group manifold.

Let us consider the example of the group $SU(2)$. The three algebra generators can be taken to be the Pauli matrices,

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\,, \qquad \sigma_2 = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}\,, \qquad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\,. \tag{3.80}$$

We can write the $SU(2)$ group elements as

$$g = \exp(\mathrm{i}\, \alpha^a\, \sigma_a)\,. \tag{3.81}$$

Writing $\alpha^a = \alpha\, n^a$, where n^a is a unit 3-vector, $n^a\, n^a = 1$, we can use the multiplication algebra of the Pauli matrices,

$$\sigma_a\, \sigma_b = \delta_{ab}\, \mathbb{1} + \mathrm{i}\, \epsilon_{abc}\, \sigma_c \tag{3.82}$$

to show that

$$\begin{aligned}(\alpha^a\,\sigma_a)^2 &= \alpha^2\,(n^a\,\sigma_a)^2 = \alpha^2\,n^a\,n^b\,(\delta_{ab}\,\mathbb{1} + \mathrm{i}\,\epsilon_{abc}\,\sigma_c)\\ &= \alpha^2\,n^a\,n^b\,\delta_{ab}\,\mathbb{1} = \alpha^2\,\mathbb{1}\,, \end{aligned}\tag{3.83}$$

and hence, using the definition of the exponential

$$\exp X = \sum_{m=0}^{\infty} \frac{1}{m!}\,X^m\,, \tag{3.84}$$

we get

$$g = \exp(\mathrm{i}\,\alpha^a\,\sigma_a) = \mathbb{1}\,\cos\alpha + \mathrm{i}\,n^a\,\sigma_a\,\sin\alpha\,. \tag{3.85}$$

Comparing with our previous parameterisation of $SU(2)$ matrices in equation (3.29), where the complex numbers a and b were subject to the constraint $|a|^2 + |b|^2 = 1$ in (3.30), we have

$$a = \cos\alpha + \mathrm{i}\,n_3\,\sin\alpha\,, \qquad b = (n_2 + \mathrm{i}\,n_1)\,\sin\alpha\,. \tag{3.86}$$

3.4.2 Structure Constants

Consider the group elements

$$A = e^{\mathrm{i}\,\lambda\,X_a}\,, \qquad B = e^{\mathrm{i}\,\lambda\,X_b}\,, \tag{3.87}$$

obtained by exponentiating the a'th and b'th generators with parameter λ. Then

$$\begin{aligned} ABA^{-1}\,B^{-1} &= e^{\mathrm{i}\,\lambda X_a}\,\epsilon^{\mathrm{i}\,\lambda X_b} e^{-\mathrm{i}\,\lambda X_a}\,\epsilon^{-\mathrm{i}\,\lambda X_b}\\ &= (1 + \mathrm{i}\,\lambda X_a - \tfrac{1}{2}\lambda^2 X_a^2 + \cdots)(1 + \mathrm{i}\,\lambda X_b - \tfrac{1}{2}\lambda^2 X_b^2 + \cdots) \times\\ &\quad\; (1 - \mathrm{i}\,\lambda X_a - \tfrac{1}{2}\lambda^2 X_a^2 + \cdots)(1 - \mathrm{i}\,\lambda X_b - \tfrac{1}{2}\lambda^2 X_b^2 + \cdots)\\ &= 1 - \lambda^2\,[X_a, X_b] + \mathcal{O}(\lambda^3)\,, \end{aligned}\tag{3.88}$$

where $[X_a, X_b]$ is the commutator, $[X_a, X_b] = X_a X_b - X_b X_a$. Since $ABA^{-1}\,B^{-1}$ must also be an element of the group, we must be able to write it as

$$ABA^{-1}\,B^{-1} = e^{\mathrm{i}\,\beta^c\,X_c}\,, \tag{3.89}$$

for some constants β^c. If we take $\lambda \to 0$, we have

$$1 - \lambda^2\,[X_a, X_b] = 1 + \mathrm{i}\,\beta^c X_c\,, \tag{3.90}$$

and so β^c is of order λ^2. We may write

$$\beta^c = -f_{ab}{}^c\,\lambda^2\,, \tag{3.91}$$

since β^c denotes a set of constants that depend upon the choice of generators X_a and X_b. Thus we have

$$[X_a, X_b] = \mathrm{i}\, f_{ab}{}^c X_c\,. \tag{3.92}$$

The constants $f_{ab}{}^c$ are called the *structure constants* of the Lie algebra.

The structure constants $f_{ab}{}^c$ have the following important properties. Firstly, since $[X_a, X_b] = -[X_b, X_a]$, we must have antisymmetry on the first two indices,

$$f_{ab}{}^c = -f_{ba}{}^c\,. \tag{3.93}$$

Secondly, we may observe that the generators obey the *Jacobi Identity*:

$$[X_a, [X_b, X_c]] + [X_b, [X_c, X_a]] + [X_c, [X_a, X_b]] = 0\,. \tag{3.94}$$

This is obvious from the matrix representation; one just has to write out the total of 12 terms, and see that there is a pairwise cancellation. It is also true from the abstract definition of generators, as a consequence of the associativity of the group multiplication law. Thus, one can derive the Jacobi identity from expanding out

$$e^{\mathrm{i}\lambda X_a}\,(e^{\mathrm{i}\lambda X_b}\, e^{\mathrm{i}\lambda X_c}) = (e^{\mathrm{i}\lambda X_a}\, e^{\mathrm{i}\lambda X_b})\, e^{\mathrm{i}\lambda X_c} \tag{3.95}$$

to order λ^3. Substituting $[X_a, X_b] = \mathrm{i}\, f_{ab}{}^c X_c$ into the Jacobi identity (3.94), we get

$$f_{bc}{}^d\, f_{ad}{}^e + f_{ca}{}^d\, f_{bd}{}^e + f_{ab}{}^d\, f_{cd}{}^e = 0\,. \tag{3.96}$$

This is also commonly referred to as the Jacobi identity.

Let us consider the example of the Lie Algebra of $SU(2)$, which, as we have seen, is generated by the the three Pauli matrices σ_a given in (3.80). Specifically, we shall choose a normalisation where we take as our $SU(2)$ generators

$$X_a = \tfrac{1}{2}\sigma_a\,. \tag{3.97}$$

From elementary computations, summarised in the multiplication rules (3.82), it follows that

$$[X_a, X_b] = \mathrm{i}\,\epsilon_{abc}\, X_c\,, \tag{3.98}$$

(There is no distinction between upstairs and downstairs indices in this case.) Comparing with (3.92), we see that the structure constants for $SU(2)$ are then given by

$$f_{ab}{}^c = \epsilon_{abc}\,. \tag{3.99}$$

In our $SU(2)$ example, we do not need to distinguish between upstairs and downstairs indices on the structure constants. In general, the indices are raised and lowered using the so-called *Cartan-Killing Metric*. It may be defined as follows:[12]

$$g_{ab} \equiv -\tfrac{1}{2} f_{ac}{}^d\, f_{bd}{}^c\,. \tag{3.100}$$

It is obviously symmetric in a and b. (Note that this is constructed from the structure constants in their "natural" up and down positions.) Upstairs indices may be lowered using the Cartan-Killing metric, and, assuming its inverse g^{ab} exists, indices may also be raised. Using the Cartan-Killing metric, we may lower the upstairs index on the structure constants $f_{ab}{}^c$, giving

$$f_{abc} = g_{cd}\, f_{ab}{}^d\,. \tag{3.101}$$

One can show, using the Jacobi identity (3.96), that f_{abc} is totally antisymmetric in its indices. From its definition it is clearly antisymmetric in ab, so it remains only to show it is antisymmetric in one other pair, say a and c. This is shown by the following calculation:

$$\begin{aligned}
-2f_{abc} &= -2f_{ab}{}^d\, g_{dc} = f_{ab}{}^d\, f_{de}{}^f\, f_{cf}{}^e\\
&= -f_{ab}{}^d\, f_{ed}{}^f\, f_{cf}{}^e\\
&= f_{be}{}^d\, f_{ad}{}^f\, f_{cf}{}^e + f_{ea}{}^d\, f_{bd}{}^f\, f_{cf}{}^e\\
&= f_{ad}{}^f\, f_{be}{}^d\, f_{cf}{}^e - f_{ae}{}^d\, f_{cf}{}^e\, f_{bd}{}^f\\
&= f_{ad}{}^f\, f_{be}{}^d\, f_{cf}{}^e - f_{ad}{}^f\, f_{ce}{}^d\, f_{bf}{}^e\\
&= -f_{ad}{}^f\, f_{fe}{}^d\, f_{bc}{}^e = f_{ad}{}^f\, f_{ef}{}^d\, f_{bc}{}^e\\
&= -2g_{ae}\, f_{bc}{}^e = -2f_{bca}\,.
\end{aligned} \tag{3.102}$$

Thus we see that $f_{abc} = -f_{cba}$, which was to be proved. We can, of course, express the total antisymmetry of the structure constants in the equation

$$f_{abc} = f_{[abc]}\,. \tag{3.103}$$

[12] Commonly, the Cartan-Killing metric is defined to be (-2) times the one defined here. This is only a matter of convention, and it is not important, as long as one is consistent in one's choice. The advantage of the convention we are choosing is that the metric is positive definite (all positive eigenvalues) for a compact group. The normalising factor is chosen so that for $SU(2)$, with $f_{ab}{}^c = \epsilon_{abc}$, we shall have $g_{ab} = -\frac{1}{2}\epsilon_{acd}\,\epsilon_{bdc} = \frac{1}{2}\epsilon_{acd}\,\epsilon_{bcd} = \delta_{ab}$, so that indeed we can, as stated above, avoid the distinction between up and down indices in this case. It should be emphasised also that the Cartan-Killing metric is completely distinct from the metrics on the vector spaces that we discussed previously when giving the classification of classical groups.

3.4.3 Simple and Semi-Simple Lie Algebras

First, we define the notion of an *Invariant Subalgebra.* Let Y be any generator in a Lie algebra $\mathcal{G}$. This has an invariant subalgebra $\mathcal{H}$ if, for every generator X in $\mathcal{H}$,

$$[X, Y] = X' \,, \tag{3.104}$$

where X' is another generator in $\mathcal{H}$, for any Y in $\mathcal{G}$. Note that X' can be zero. Obviously the entire Lie algebra $\mathcal{G}$ fulfils the requirements for being an invariant subalgebra, and so it is useful to define a *Proper Invariant Subalgebra* as an invariant subalgbra that is strictly smaller than $\mathcal{G}$ itself.

We may now define a *Simple Lie Algebra*, as being a Lie algbera that has no proper invariant subalgebras.

A special case of an invariant subalgebra is an *Abelian Invariant Subalgebra.* If X is an element of an abelian invariant subalgebra, and if Y is any element in the full algebra $\mathcal{G}$, then there is an abelian invariant subalgebra $\mathcal{H}$ if, for every generator X in $\mathcal{H}$,

$$[X, Y] = 0 \tag{3.105}$$

for *all* Y in $\mathcal{G}$. Each such generator X corresponds to a $U(1)$ factor (or in the non-compact case an $\mathbb{R}$ factor) in the Lie algebra $\mathcal{G}$. If X_a, for some given value of a, is such an abelian generator then it follows from $[X_a, X_b] = \mathrm{i}\, f_{ab}{}^c\, X_c$ that

$$f_{ab}{}^c = 0 \qquad \text{for all } b \text{ and } c\,. \tag{3.106}$$

In this case it follows that the Cartan-Killing metric

$$g_{ab} = -\tfrac{1}{2} f_{ac}{}^d\, f_{bd}{}^c \tag{3.107}$$

has a zero eigenvalue, since we shall have $g_{ab} = 0$ for all b. Thus if g_{ab} has p zero eigenvalues then there are p abelian invariant factors $U(1)$ or $\mathbb{R}$ in the Lie algebra. Note that if there are any such factors we shall have $\det(g_{ab}) = 0$, and so the metric is not invertible.

A Lie algebra with no abelian invariant subalgebras is called a *Semi-Simple Lie Algebra.*

At the level of the Lie group, we may say that a semi-simple Lie group has no $U(1)$ or $\mathbb{R}$ factors. A simple Lie group is not a product of subgroups.

A consequence of the above is that when discussing the classification of Lie groups we may concentrate on the simple Lie groups.

3.4.4 Properties of the Lie Algebra Generators

Here, we re-examine the defining conditions for our classification of Lie groups, but now at the infinitesimal level of the Lie algebra. Recall that we encountered three classes of metric-preserving classical groups, corresonding to having a bilinear symmetric, bilinear antisymmetric, or sequilinear symmetric metric on the vector space on which the matrices act.

To avoid the risk of confusion with the Cartan-Killing metric, let us for now use the symbol G_{ij} to denote the invariant metric on the vector space. Thus for the bilinear metrics, we had that the matrices $A_i{}^j$ acting on the vector space preserve G_{ij} according to

$$A_i{}^k\, A_j{}^\ell\, G_{k\ell} = G_{ij}\,. \tag{3.108}$$

In matrix notation, this reads

$$A\, G\, A^T = G\,. \tag{3.109}$$

For the Lie algebra, we can express A via exponentiation of the Lie algebra generators X_a, as

$$A = e^{\mathrm{i}\,\alpha^a\, X_a}\,. \tag{3.110}$$

For generators close to the identity we can take the coefficients α^a to be very small, and work only to linear order in α^a. Thus we may write

$$A = \mathbb{1} + \mathrm{i}\,\alpha^a\, X_a\,, \tag{3.111}$$

neglecting the higher-order terms from expanding the exponential in a Taylor series. The metric-preserving condition (3.109) becomes

$$(\mathbb{1} + \mathrm{i}\,\alpha^a\, X_a)\, G\, (\mathbb{1} + \mathrm{i}\,\alpha^b\, X_b) = G\,, \tag{3.112}$$

which, to linear order in α gives

$$G + \mathrm{i}\,\alpha^a\, (X_a\, G + G\, X_a^T) = G\,, \tag{3.113}$$

and hence

$$X_a\, G + G\, X_a^T = 0\,. \tag{3.114}$$

In the case of a bilinear symmetric metric, and choosing the compact form where it has all positive eigenvalues, the canonical form was just $G = \mathbb{1}$, and hence (3.114) becomes just

$$X_a = -X_a^T\,, \tag{3.115}$$

i.e. that X_a is antisymmetric. This, then, is the condition on the generators of $O(n)$ or $SO(n)$. Since we are taking the generators to be Hermitean, this means they are imaginary.

For bilinear antisymmetric metrics, the canonical form for G is given in (3.62). With this choice for G, the equations (3.114) give the conditions on the generators X_a for $Sp(2m)$.

Finally, for sequilinear symmetric metrics, the metric-preserving condition (3.71) reads, in matrix notation,

$$A\,G\,A^\dagger = G\,. \tag{3.116}$$

In the infinesimal form for generators $A = e^{\mathrm{i}\,\alpha^a\,X_a}$ close to the identity, this becomes

$$X_a\,G - G\,X_a^\dagger = 0\,. \tag{3.117}$$

The canonical form for G is is given in (3.65). For the compact case (i.e. $SU(n)$), we have $G = \mathbb{1}$, and then (3.117) becomes simply

$$X_a = X_a^\dagger\,, \tag{3.118}$$

i.e. X_a is Hermitean. (Recall that this is what we saw in our $SU(2)$ example discussed previously.)

3.5 Roots and Weights

3.5.1 Notation

We have been thinking of the generators X_a as being matrices, but we can instead think of them as linear operators acting on states (as in quantum mechanics). We can then consider the matrix elements $[X_a]_{ij}$ of the generators X_a, defined by

$$[X_a]_{ij} = \langle i|X_a|j\rangle \tag{3.119}$$

in a Dirac *Bra* and *Ket* notation, where the states are normalised such that

$$\langle i|j\rangle = \delta_{ij}\,. \tag{3.120}$$

An arbitrary state $|\Psi\rangle$ can be expressed as a linear combination of states $|i\rangle$:

$$|\Psi\rangle = \sum_i a_i\,|i\rangle\,. \tag{3.121}$$

The expansion coefficients a_i can be read off by multiplying by $\langle j|$:

$$\langle j|\Psi\rangle = \sum_i a_i\,\langle j|i\rangle = \sum_i a_i\,\delta_{ij} = a_j\,, \tag{3.122}$$

whence we have

$$|\Psi\rangle = \sum_i |i\rangle \langle i|\Psi\rangle \,. \tag{3.123}$$

Since this is true for any state $|\Psi\rangle$, we have the *Completeness Relation*

$$\sum_i |i\rangle \langle i| = \mathbb{1} \,. \tag{3.124}$$

We can now calculate the action of X_a on $|i\rangle$, obtaining

$$X_a|i\rangle = \sum_j |j\rangle \langle j|X_a|i\rangle = \sum_j |j\rangle \, [X_a]_{ji} \,. \tag{3.125}$$

This shows that the states $|j\rangle$ can be thought of as row vectors, with the matrix $[X_a]_{ji}$ associated with the linear operator X_a acting by matrix multiplication from the right.

3.5.2 The Example of $SU(2)$

Here, we shall review some basic results about the construction of the representations of the $SU(2)$ algebra. This will probably be very familiar from quantum mechanics. The purpose of doing this is that the procedures used for studying $SU(2)$ will generalise to any Lie algebra, as we shall see in subsequent sections.

We saw in 3.4.2 that the structure constants of $SU(2)$ are given by $f_{ab}{}^c = \epsilon_{abc}$. Thus if we call the generators J_a, with $a = 1, 2, 3$, then we shall have

$$[J_1, J_2] = \mathrm{i}\, J_3 \,, \qquad [J_2, J_3] = \mathrm{i}\, J_1 \,, \qquad [J_3, J_1] = \mathrm{i}\, J_2 \,. \tag{3.126}$$

These are, of course, just the familiar commutation relations of the angular momentum generators in quantum mechanics.

Suppose we have an N-dimensional irreducible representation of $SU(2)$. (This is what is known in quantum mechanics as a spin-$(2N+1)$ representation.) Since the operators J_a are Hermitean, we can choose a basis of states in the representation such that J_3 is diagonal.[13]

Since the number N of states in the representation is finite, and they are all, by construction, eigenstates of J_3, it follows that there must exist a state with the largest eigenvalue, say λ. Let us denote this state by $|\lambda, \alpha\rangle$, where we have introduced α as an additional index which will label distinct states having the same eigenvalue λ, in case it should turn out that there is a degeneracy. By definition, we shal have

$$J_3 \, |\lambda, \alpha\rangle = \lambda \, |\lambda, \alpha\rangle \,. \tag{3.127}$$

[13]We cannot, of course, simultaneously have J_1 or J_2 being diagonal, since J_1 and J_2 do not commute with J_3.

We can always orthonormalise these states so that

$$\langle \lambda, \alpha | \lambda, \beta \rangle = \delta_{\alpha\beta}\,. \tag{3.128}$$

We now define

$$J_\pm \equiv \frac{1}{\sqrt{2}}\,(J_1 \pm \mathrm{i}\,J_2)\,. \tag{3.129}$$

From (3.126), it follows that we shall have the commutation relations

$$[J_3, J_\pm] = \pm J_\pm\,, \qquad [J_+, J_-] = J_3\,. \tag{3.130}$$

On a state $|\mu\rangle$, with eigenvalue μ, i.e. $J_3\,|\mu\rangle = \mu\,|\mu\rangle$, we have

$$\begin{aligned} J_3\,J_\pm\,|\mu\rangle &= [J_3, J_\pm]\,|\mu\rangle + J_\pm\,|\mu\rangle \\ &= \pm J_\pm\,|\mu\rangle + \mu\,J_\pm\,|\mu\rangle \\ &= (\mu \pm 1)\,|\mu\rangle\,. \end{aligned} \tag{3.131}$$

As will be familiar from quantum mechanics, the operators $J_\pm$ are called raising and lowering operators, since they increase or decrease the J_3 eigenvalue.

Since we are assuming $|\lambda, \alpha\rangle$ has the highest possible eigenvalue in the N-dimensional representation, it follows that we must have

$$J_+\,|\lambda, \alpha\rangle = 0\,, \tag{3.132}$$

since if it were non-vanishing, it would by virtue of (3.131) have the larger eigenvalue $\lambda + 1$, which is impossible.

We know from (3.131) that $J_-\,|\lambda, \alpha\rangle$ is a state with J_3 eigenvalue $\lambda - 1$, and so we may write

$$J_-\,|\lambda, \alpha\rangle = N_\lambda(\alpha)\,|\lambda - 1, \alpha\rangle\,, \tag{3.133}$$

for some constant $N_\lambda(\alpha)$. The Hermitean conjugate of (3.133) is given by

$$\langle \lambda, \alpha|\,J_+ = \bar{N}_\lambda(\alpha)\,\langle \lambda - 1, \alpha|\,. \tag{3.134}$$

(Recall that J_1 and J_2 are Hermitean, so $J_-^\dagger = J_+$.) We therefore have

$$\begin{aligned} \bar{N}_\lambda(\beta) N_\lambda(\alpha)\,\langle \lambda - 1, \beta | \lambda - 1, \alpha\rangle &= \langle \lambda, \beta | J_+\,J_-\,|\lambda, \alpha\rangle \\ &= \langle \lambda, \beta | [J_+, J_-]\,|\lambda, \alpha\rangle \\ &= \langle \lambda, \beta | J_3\,|\lambda, \alpha\rangle \\ &= \lambda\,\langle \lambda, \beta | \lambda, \alpha\rangle \\ &= \lambda\,\delta_{\alpha\beta}\,. \end{aligned} \tag{3.135}$$

It then follows from (3.128) that we can choose

$$\langle\lambda-1,\beta|\lambda-1,\alpha\rangle=\delta_{\alpha\beta}\,, \tag{3.136}$$

and that we can choose $N_\lambda(\alpha)$ to be real, independent of α, and given by

$$N_\lambda(\alpha)=N_\lambda\equiv\sqrt{\lambda}\,. \tag{3.137}$$

Note that we shall also have

$$\begin{aligned} J_+\,|\lambda-1,\alpha\rangle &= \frac{1}{N_\lambda}\,J_+\,J_-\,|\lambda,\alpha\rangle=\frac{1}{N_\lambda}\,[J_+,J_-]\,|\lambda,\alpha\rangle \\ &= \frac{1}{N_\lambda}\,J_3\,|\lambda,\alpha\rangle=\frac{1}{N_\lambda}\,\lambda\,|\lambda,\alpha\rangle \\ &= N_\lambda\,|\lambda,\alpha\rangle\,. \end{aligned} \tag{3.138}$$

Note also that $J_\pm$ raises or lowers the J_3 eigenvalue without changing α.

Proceeding by considering $J_-\,|\lambda-1,\alpha\rangle$, the same argument as above shows that there are orthonormal states $|\lambda-1,\alpha\rangle$ satisfying

$$J_-\,|\lambda-1,\alpha\rangle=N_{\lambda-1}\,|\lambda-2,\alpha\rangle\,,\qquad J_+\,|\lambda-2,\alpha\rangle=N_{\lambda-1}\,|\lambda-1,\alpha\rangle\,, \tag{3.139}$$

for certain real constants $N_{\lambda-1}$. Continuing, we shall have

$$J_-\,|\lambda-k,\alpha\rangle=N_{\lambda-k}\,|\lambda-k-1,\alpha\rangle\,,\qquad J_+\,|\lambda-k-1,\alpha\rangle=N_{\lambda-k}\,|\lambda-k,\alpha\rangle\,. \tag{3.140}$$

The constants N_μ are determined as follows:

$$\begin{aligned} N^2_{\lambda-k} &= N^2_{\lambda-k}\,\langle\lambda-k,\alpha|\lambda-k,\beta\rangle \\ &= \langle\lambda-k,\alpha|\,J_+\,J_-\,|\lambda-k,\beta\rangle \\ &= \langle\lambda-k,\alpha|\,[J_+,J_-]\,|\lambda-k,\beta\rangle+\langle\lambda-k,\alpha|\,J_-\,J_+\,|\lambda-k,\beta\rangle \\ &= \langle\lambda-k,\alpha|\,J_3\,|\lambda-k,\beta\rangle+N^2_{\lambda-k+1}\,\langle\lambda-k+1,\alpha|\lambda-k+1,\beta\rangle \\ &= \lambda-k+N^2_{\lambda-k+1}\,. \end{aligned} \tag{3.141}$$

Thus we have

$$\begin{aligned} N^2_\lambda &= \lambda\,, \\ N^2_{\lambda-1}-N^2_\lambda &= \lambda-1\,, \\ N^2_{\lambda-2}-N^2_{\lambda-1} &= \lambda-2\,, \\ &\;\;\vdots \\ N^2_{\lambda-2}-N^2_{\lambda-1} &= \lambda-k\,. \end{aligned} \tag{3.142}$$

Thus by adding these, we get

$$
\begin{aligned}
N^2_{\lambda-k} &= (k+1)\,\lambda - \tfrac{1}{2}k(k+1) \\
&= \tfrac{1}{2}(k+1)(2\lambda - k)\,.
\end{aligned} \tag{3.143}
$$

Eventually, if we act sufficiently many times with J_-, we must reach a state (or states) in the representation with the lowest possible eigenvalue under J_3. (This must be the case, since we have assumed there are only a finite number of states in the representation.) Thus for some integer n, it must be that we have a state (or states) $|\lambda - n, \alpha\rangle$ such that

$$
J_- \, |\lambda - n, \alpha\rangle = 0\,, \tag{3.144}
$$

and therefore $N_{\lambda-n} = 0$. It follows from (3.143) that

$$
\lambda = \tfrac{1}{2}n\,. \tag{3.145}
$$

Since n is an integer, we see that the highest J_3 eigenvalue λ is an integer or half-integer, and that all the other states in the representation have J_3 eigenvalues μ that are likewise either all integers or all half-integers. These J_3 eigenvalues lie in integer steps between $+\frac{1}{2}n$ and $-\frac{1}{2}n$. The highest eigenvalue $\lambda = \frac{1}{2}n$ is usually called the *Spin*, and denoted by j. There are in total $(2j+1)$ eigenvalues in the range between j and $-j$.

Since the raising and lowering operators change the J_3 eigenvalue without changing α, it follows that different values of α correspond to disjoint and independent representations of $SU(2)$. For a so-called *Irreducible Representation*, there is just one α, and so we don't need to carry around the α label any more.

We normally denote the full set of states in an irreducible representation of $SU(2)$ by $|m, j\rangle$, where j is the spin that labels the representation, and m is the J_3 eigenvalue of each state;

$$
J_3 \, |m, j\rangle = m \, |m, j\rangle\,. \tag{3.146}
$$

We have seen that m can take integer-spaced values in the interval

$$
-j \le m \le j\,. \tag{3.147}
$$

The total number of states in the spin-j representation is therefore $(2j+1)$.

To use the terminology that we shall be using for the more general discussion of representations for arbitrary Lie algebras, we call $|j, j\rangle$ the *Highest Weight State* in the irreducible representation, and the state $|m, j\rangle$ is said to have *weight* m.

The states $|m, j\rangle$ can be shown to be orthogonal, in the sense that

$$\langle m, j|m', j'\rangle = 0\,, \qquad \text{unless } m = m' \text{ and } j = j'\,. \tag{3.148}$$

The orthogonality for different values of m is easy to see. We just sandwich J_3 in the inner product, and use the facts that

$$J_3\,|m', j'\rangle = m'\,|m', j'\rangle\,, \qquad \langle m, j|\,J_3 = m\,\langle m, j|\,. \tag{3.149}$$

(The second equation follows just from Hermitean conjugation of $J_3\,|m, j\rangle = m\,|m, j\rangle$, noting that J_3 itself is Hermitean.) Thus we can evaluate (3.148) two ways, depending on whether we act with J_3 to the left or the right, leading to

$$(m - m')\,\langle m, j|m', j'\rangle = 0\,. \tag{3.150}$$

This immediately shows that $\langle m, j|m', j'\rangle = 0$ if $m \ne m'$.

It remains to prove that $\langle m, j|m, j'\rangle = 0$ unless $j = j'$. Without loss of generaliity, we may assume $j' > j$, and consider

$$\langle j, j|\,J_-\,|j+1, j'\rangle = N_{\lambda-k}\,\langle j, j|j, j'\rangle\,, \tag{3.151}$$

where $N_{\lambda-k}$ can be read off from (3.143). Being careful about the meaning of the symbols, we see that $\lambda = j'$ and $\lambda - k = j + 1$, which then implies that

$$N_{\lambda-k}^2 = \tfrac{1}{2}(k+1)(2\lambda - k) = \tfrac{1}{2}(j' - j)(j' + j + 1)\,. \tag{3.152}$$

This is manifestly non-zero (since $J' > j$, and both j and j' are non-negative). On the other hand the left-hand side of (3.151) is clearly zero, since $\langle j, j|\,J_-$ is the Hermitean conjugate of $J_+\,|j, j\rangle$, which is obviously zero since $|j, j\rangle$ is the highest-weight state. Therefore we conclude from (3.151) that

$$\langle j, j|j, j'\rangle = 0\,. \tag{3.153}$$

By analogous calculations, making repeated applications of J_+ and J_- operators, we can similarly show that

$$\langle m, j|m, j'\rangle = 0 \tag{3.154}$$

for all values of m, which is what we wanted to establish.[14] Having established the orthogonality of the states, we can now normalise them so that they satisfy

$$\langle m, j|m', j'\rangle = \delta_{mm'}\,\delta_{jj'}\,. \tag{3.155}$$

[14]There is in fact a much simpler proof of the orthogonality (3.154), which follows by inserting the Hermitean operator $J^2 \equiv J_a J_a$. This operator (the "total angular momentum," in the language of quantum

3.5.3 Arbitrary Simple Lie Algebras

Consider an arbitrary compact simple Lie algebra $\mathcal{G}$. Suppose we have matrices X_a in some particular representation D of $\mathcal{G}$, which generate the algebra, $[X_a, X_b] = \mathrm{i}\, f_{ab}{}^c X_c$. We can normalise the generators so that

$$\mathrm{tr}(X_a^\dagger\, X_b) = \lambda\, \delta_{ab}\,, \tag{3.156}$$

for some positive constant λ.

We can divide the generators X_a into two categories:

1) A maximal set of m mutually-commuting Hermitean generators that can be simultaneously diagonalised. These are denoted by H_i, and they are called the generators of the *Cartan Subalgebra.* They are the generalisation of J_3 in the $SU(2)$ algebra.

2) The remaining generators are organised into raising and lowering operators denoted by $E_{\vec\alpha}$, where the subscript $\vec\alpha$ on a given such generator $E_{\vec\alpha}$ denotes an m-component vector label for that generator, whose significance will be explained below. The generators $E_{\vec\alpha}$ are the generalisation of $J_\pm$ in the $SU(2)$ case.

A decomposition of this type can always be made. Having organised the generators in this fashion, we will have the following structure of commutation relations:

$$\begin{aligned}
&[H_i, H_j] = 0\,, \qquad [H_i, E_{\vec\alpha}] = \alpha_i\, E_{\vec\alpha}\,,\\
&[E_{\vec\alpha}, E_{\vec\beta}] = \mathcal{N}_{\vec\alpha,\vec\beta}\, E_{\vec\alpha+\vec\beta}\,, \qquad \text{if } \vec\alpha \ne -\vec\beta\,,\\
&[E_{\vec\alpha}, E_{-\vec\alpha}] = \sum_i \alpha_i\, H_i\,.
\end{aligned} \tag{3.157}$$

Since the basis and normalisation of the Cartan generators is not yet specified, we can arrange things so that

$$\mathrm{tr}(H_i H_j) = k_D\, \delta_{ij}\,, \tag{3.158}$$

where k_D is some constant that depends upon the representation D.

mechanics) has the property $J^2\,|m,j\rangle = j(j+1)\,|m,j\rangle$, and hence we have

$$\langle m,j|\, J^2\, |m,j'\rangle = j(j+1)\,\langle m,j|m,j'\rangle = j'(j'+1)\,\langle m,j|m,j'\rangle$$

by acting either to the right or the left. Thus we have

$$(j'-j)(j'+j+1)\,\langle m,j|m,j'\rangle = 0\,,$$

which shows (3.154) when $j \ne j'$. The reason why we have not used ths proof is that it does not generalise to the case of arbitrary Lie groups, unlike the proof we have presented earlier.

The meaning of the vector subscript $\vec{\alpha}$ on a generator $E_{\vec{\alpha}}$ is now apparent. Because of the way we have organised them, the generators $E_{\vec{\alpha}}$ are all *eigenstates* with respect to all m of the Cartan generators H_i, in the sense given on the top line of (3.157): Each commutator $[H_i, E_{\vec{\alpha}}]$, for each value of the index i labelling the Cartan generators, gives a constant multiple of $E_{\vec{\alpha}}$ itself. That constant is called α_i, and the set of these eigenvalues, or *weights*, is assembled into an m-component vector $\vec{\alpha}$ that is used as a label for the particular generator $E_{\vec{\alpha}}$:

$$\vec{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_m)\,. \tag{3.159}$$

The total set of generators therefore comprise the m Cartan generators H_i, and the remaining ones $E_{\vec{\alpha}}$, where the vector label $\vec{\alpha}$ on each such generator indicates the eigenvalues, or weights, of that particular generator under the Cartan generators.

The meanings of the other commutation relations in (3.157) are as follows. Firstly, the relations $[H_i, H_j] = 0$ obviously just say that the Cartan generators commute amongst themselves. The commutation relation for $[E_{\vec{\alpha}}, E_{\vec{\beta}}]$ shows that if one picks any two of the raising and lowering operators, $E_{\vec{\alpha}}$ and $E_{\vec{\beta}}$, then their commutator will in general produce another generator whose eigenvalues under the Cartan generators are $\vec{\alpha} + \vec{\beta}$. This can be easily understood, by writing out the Jacobi identity:

$$\begin{aligned} 0 &= [H_i, [E_{\vec{\alpha}}, E_{\vec{\beta}}]] + [E_{\vec{\alpha}}, [E_{\vec{\beta}}, H_i]] + [E_{\vec{\beta}}, [H_i, E_{\vec{\alpha}}]] \\ &= [H_i, [E_{\vec{\alpha}}, E_{\vec{\beta}}]] - [E_{\vec{\alpha}}, [H_i, E_{\vec{\beta}}]] - [[H_i, E_{\vec{\alpha}}], E_{\vec{\beta}}] \\ &= [H_i, [E_{\vec{\alpha}}, E_{\vec{\beta}}]] - \beta_i\, [E_{\vec{\alpha}}, E_{\vec{\beta}}] - \alpha_i\, [E_{\vec{\alpha}}, E_{\vec{\beta}}]\,. \end{aligned} \tag{3.160}$$

Hence we have

$$[H_i, [E_{\vec{\alpha}}, E_{\vec{\beta}}]] = (\alpha_i + \beta_i)\, [E_{\vec{\alpha}}, E_{\vec{\beta}}]\,. \tag{3.161}$$

The constant $\mathcal{N}_{\vec{\alpha},\vec{\beta}}$ in the commutation relation (3.157) is dependent on how the generators are normalised. Note that it might be that for a given pair of generators $E_{\vec{\alpha}}$ and $E_{\vec{\beta}}$ that their commutator vanishes, in which case $\mathcal{N}_{\vec{\alpha},\vec{\beta}}$ will be zero. The calculation in (3.160) and (3.161) shows that if their commutator is non-vanishing, then the weights of the resulting generator $E_{\vec{\alpha}+\vec{\beta}}$ will be $\vec{\alpha} + \vec{\beta}$. This should make clear why it is that we can think of the generators $E_{\vec{\alpha}}$ as raising or lowering operators; when one commutes $E_{\vec{\alpha}}$ with any other generator $E_{\vec{\beta}}$, one gets another generator whose weight is the original $\vec{\beta}$ boosted by the addition of $\vec{\alpha}$. Whether we call a given $E_{\vec{\alpha}}$ a raising operator or a lowering operator will depend upon the way in which we classify the weights $\vec{\alpha}$ as being positive or negative. We shall explain this in detail later.

There is one exception to the above, and that is if one considers the commutator of generator $E_{\vec{\alpha}}$ with its "negative," namely $E_{-\vec{\alpha}}$. The calculation in (3.160) and (3.161) is still valid, and it now shows that $[E_{\vec{\alpha}}, E_{-\vec{\alpha}}]$ will have zero weights under all the Cartan generators. This means in fact that $[E_{\vec{\alpha}}, E_{-\vec{\alpha}}]$ will be a linear combination of the Cartan generators themselves. As it turns out, the coefficients in this sum over Cartan generators are as given in the final line of (3.157).

Before proceeding with the general discussion, let us return briefly to our earlier example of $SU(2)$, to see how it fits in this general framework. We organised the generators J_i, satisfying (3.126), into the combinations $J_\pm$ and J_3 as in (3.129), and found that they satisfied (3.130). Casting this into our general framework, we shall have

$$H_1 = J_3\,, \qquad E_1 = J_+\,, \qquad E_{-1} = J_-\,. \tag{3.162}$$

Note that since we have just one Cartan generator in this case, our vectors $\vec{\alpha}$ labelling the raising and lowering generators $E_{\vec{\alpha}}$ are just 1-component objects, i.e. numbers. The generators satisfy

$$[H_1, E_1] = E_1\,, \qquad [H_1, E_{-1}] = -E_{-1}\,, \qquad [E_1, E_{-1}] = H_1\,, \tag{3.163}$$

which can be compared with the general set-up in (3.157). One should be careful to understand the notation here. We call the Cartan generator H_1, with its "1" subscript simply indicating that it is the first (and only!) Cartan generator of $SU(2)$. The "1" subscript on E_1, on the other hand, denotes that E_1 has weight 1 under the Cartan generator. The lowering operator E_{-1} has subscript -1 because it has weight -1 under the Cartan generator. In this $SU(2)$ example the algebra is so small (only three generators in total) that we aren't seeing any of the $[E_{\vec{\alpha}}, E_{\vec{\beta}}] \sim E_{\vec{\alpha}+\vec{\beta}}$ type commutation relations, because we don't have enough generators to play with. As we proceed, we shall look at more complicated examples that have more "beef."

To proceed, it will be useful to look at how the generators act on states in a representation. Suppose we denote some representation of a Lie algebra $\mathcal{G}$ by D. Since the Cartan generators H_i commute, we can organise the states so that each one is simultaneously an eigenvector under each Cartan generator. Let us represent a state in the representation by $|\vec{\mu}, D\rangle$, satisfying

$$H_i\,|\vec{\mu}, D\rangle = \mu_i\,|\vec{\mu}, D\rangle\,. \tag{3.164}$$

What we have done here is to label the state by its weights $\vec{\mu}$ under the Cartan generators. The vector $\vec{\mu}$ is, not surprisingly, called the *weight vector* of the state.

First, we shall consider a very particular representation, called the *adjoint representation*. Every algebra has an adjoint representation; its dimension (i.e. the number of states in the representation) is just equal to the dimension of the algebra itself. In fact we can simply use the straucture constants $f_{ab}{}^c$ themselves to construct a matrices of the adjoint representation. Let Y_a be a matrix whose components are $(Y_a)_b{}^c$, where b labels rows and c labels columns, given by

$$(Y_a)^b{}_c = \mathrm{i}\, f_{ac}{}^b\,. \tag{3.165}$$

Evaluating the matrix commutator, we shall have

$$\begin{aligned}
[Y_a, Y_b]^c{}_d &= (Y_a)^c{}_e\,(Y_b)^e{}_d - (Y_b)^c{}_e\,(Y_a)^e{}_d \\
&= -f_{ae}{}^c\,f_{bd}{}^e + f_{be}{}^c\,f_{ad}{}^e \\
&= -f_{ae}{}^c\,f_{bd}{}^e - f_{be}{}^c\,f_{da}{}^e \\
&= f_{de}{}^c\,f_{ab}{}^e \\
&= -f_{ab}{}^e\,f_{ed}{}^c \\
&= \mathrm{i}\, f_{ab}{}^e\,(Y_e)^c{}_d\,,
\end{aligned} \tag{3.166}$$

and so we have

$$[Y_a, Y_b] = \mathrm{i}\, f_{ab}{}^c\, Y_c\,. \tag{3.167}$$

(We used the Jacobi identity (3.96) in getting from the 3'rd to the 4'th line above.)

We can look at this also at the level of the states. In the adjoint representation we can associate a state with each generator X_a of the Lie algebra $\mathcal{G}$, and denote it by $|X_a\rangle$, for $a = 1, \ldots,$dim $\mathcal{G}$. With the generators normalised as in (3.156), we define the states $|X_a\rangle$ such that

$$\langle X_a|X_b\rangle = \lambda^{-1}\,\mathrm{tr}(X_a^\dagger X_b) = \delta_{ab}\,. \tag{3.168}$$

From the discussion above, we see that the matrix elements of the generators will then be given by

$$\langle X_a|X_b|X_c\rangle = -\mathrm{i}\, f_{cb}{}^a\,, \tag{3.169}$$

where we have normalised the states so that $\langle X_a|X_b\rangle = \delta^a_b$. Using the completeness relation $|X_c\rangle\,\langle X_c| = \mathbb{1}$, we then have

$$\begin{aligned}
X_a\,|X_b\rangle &= |X_c\rangle\,\langle X_c|X_a|X_b\rangle \\
&= |X_c\rangle\,(Y_a)^c{}_b \\
&= |X_c\rangle\,(\mathrm{i}\, f_{ab}{}^c) \\
&= |[X_a, X_b]\rangle\,,
\end{aligned} \tag{3.170}$$

since $[X_a, X_b] = \mathrm{i}\, f_{ab}{}^c\, X_c$. Thus we have

$$X_a\, |X_b\rangle = |[X_a, X_b]\rangle\,. \tag{3.171}$$

Of the total of $n = \dim \mathcal{G}$ states in the adjoint representation, we know that $m = \mathrm{rank}\, \mathcal{G}$ of them, which we can denote by $|H_i\rangle$, will have zero weights:

$$H_i\, |H_j\rangle = |[H_i, H_j]\rangle = 0\,. \tag{3.172}$$

The remaining $n-m$ states will all be associated with the raising and lowering generators $E_{\vec\alpha}$, and so we denote these by $|E_{\vec\alpha}\rangle$. They therefore satisfy

$$H_i\, |E_{\vec\alpha}\rangle = |[H_i, E_{\vec\alpha}]\rangle = \alpha_i\, |E_{\vec\alpha}\rangle\,. \tag{3.173}$$

Note that the raising and lowering generators are not Hermitean, and in fact the Hermitean conjugate of a raising generator gives a lowering generator, and *vice versa*. (Recall that in the $SU(2)$ example we had $J_\pm^\dagger = J_\mp$, since $J_\pm = (J_1 \pm \mathrm{i}\, J_2)/\sqrt{2}$, and J_1 and J_2 themselves are Hermitean.) To see this, consider

$$\begin{aligned} [H_i, E_{\vec\alpha}]^\dagger &= (H_i\, E_{\vec\alpha})^\dagger - (E_{\vec\alpha}\, H_i)^\dagger \\ &= E_{\vec\alpha}^\dagger\, H_i^\dagger - H_i^\dagger\, E_{\vec\alpha}^\dagger \\ &= E_{\vec\alpha}^\dagger\, H_i - H_i\, E_{\vec\alpha}^\dagger \\ &= -[H_i, E_{\vec\alpha}^\dagger] \end{aligned} \tag{3.174}$$

and so we have from $[H_i, E_{\vec\alpha}] = \alpha_i\, E_{\vec\alpha}$ that

$$[H_i, E_{\vec\alpha}^\dagger] = -\alpha_i\, E_{\vec\alpha}^\dagger\,. \tag{3.175}$$

It is therefore natural to write

$$E_{\vec\alpha}^\dagger = E_{-\vec\alpha}\,. \tag{3.176}$$

As usual, we normalise the states to have unit length, and so we shall have

$$\begin{aligned} \langle E_{\vec\alpha} | E_{\vec\beta}\rangle &= \delta_{\vec\alpha,\vec\beta} \equiv \delta_{\alpha_1\beta_1}\, \delta_{\alpha_2\beta_2} \cdots \delta_{\alpha_m\beta_m}\,, \\ \langle H_i | H_j\rangle &= \delta_{ij}\,. \end{aligned} \tag{3.177}$$

We shall return in a moment to considering the states in an arbitrary representation D of the Lie algebra $\mathcal{G}$. Before doing so, let us just recapitulate that in the discussion above, we have considered specifically the states $|H_i\rangle$ and $|E_{\vec\alpha}\rangle$ of the n-dimensional *adjoint representation*. They satisfy the eigenvalue equations

$$H_i\, |H_j\rangle = 0\,, \qquad H_i\, |E_{\vec\alpha}\rangle = \alpha_i\, |E_{\vec\alpha}\rangle\,. \tag{3.178}$$

For a general representation, the eigenvalues of the various states with respect to the Cartan generators H_i are called the *weights* of the states. In the special case of the adjoint representation that we have been considering, the weights are called the *roots*. Thus we say that $|E_{\vec\alpha}\rangle$ has the *root vector* $\vec\alpha$.

3.5.4 Arbitrary irreducible representation

Now let us return to considering an arbitrary irreducible representation D of the Lie algebra $\mathcal{G}$. The state $|\vec\mu, D\rangle$ satisfies

$$H_i\,|\vec\mu, D\rangle = \mu_i\,|\vec\mu, D\rangle\,, \tag{3.179}$$

and $\vec\mu$ is called the weight vector of the state. By a standard manipulation that is precisely analogous to the one we performed for $SU(2)$, we see that the generator $E_{\vec\alpha}$ acts as a raising or lowering operator on this state:

$$\begin{aligned} H_i\,E_{\vec\alpha}\,|\vec\mu, D\rangle &= [H_i, E_{\vec\alpha}]\,|\vec\mu, D\rangle + E_{\vec\alpha}\,H_i\,|\vec\mu, D\rangle \\ &= \alpha_i\,E_{\vec\alpha}\,|\vec\mu, D\rangle + \mu_i\,E_{\vec\alpha}\,|\vec\mu, D\rangle \\ &= (\mu_i + \alpha_i)\,E_{\vec\alpha}\,|\vec\mu, D\rangle\,. \end{aligned} \tag{3.180}$$

Of course we shall also have

$$H_i\,E_{-\vec\alpha}\,|\vec\mu, D\rangle = (\mu_i - \alpha_i)\,E_{-\vec\alpha}\,|\vec\mu, D\rangle\,. \tag{3.181}$$

Thus, as with $SU(2)$ we then define $|\vec\mu \pm \vec\alpha, D\rangle$ as the states with weights $(\vec\mu \pm \vec\alpha)$, and write

$$E_{\pm\vec\alpha}\,|\vec\mu, D\rangle = N_{\pm\vec\alpha,\vec\mu}\,|\vec\mu \pm \vec\alpha, D\rangle\,. \tag{3.182}$$

The $N_{\pm\vec\alpha,\vec\mu}$ are constants to be determined. As usual, the states will all be normalised to unit length.

Now, in the adjoint representation, the state $|E_{\vec\alpha}\rangle$ has weight $\vec\alpha$, i.e. $H_i\,|E_{\vec\alpha}\rangle = \alpha_i\,|E_{\vec\alpha}\rangle$. Therefore $E_{-\vec\alpha}\,|E_{\vec\alpha}\rangle$ has weight zero, and so it must be a linear combination of the zero-weight states $|H_i\,\rangle$:

$$E_{-\vec\alpha}\,|E_{\vec\alpha}\rangle = c_i\,|H_i\rangle\,. \tag{3.183}$$

We can determine the constants c_i by noting that

$$\langle H_j|\,E_{-\vec\alpha}\,|E_{\vec\alpha}\rangle = c_i\,\langle H_j|H_i\rangle = c_i\,\delta_{ij} = c_j\,. \tag{3.184}$$

Thus we have

$$\begin{aligned} c_j &= \langle H_j|\, E_{-\vec{\alpha}}\, |E_{\vec{\alpha}}\rangle = \langle E_{\vec{\alpha}}|\, E_{\vec{\alpha}}\, |H_j\rangle \\ &= \langle E_{\vec{\alpha}}|\, [E_{\vec{\alpha}}, H_j]\rangle = -\langle E_{\vec{\alpha}}|\, [H_j, E_{\vec{\alpha}}]\rangle \\ &= -\alpha_j\, \langle E_{\vec{\alpha}}|E_{\vec{\alpha}}\rangle = -\alpha_j\,. \end{aligned} \tag{3.185}$$

Since by definition $E_{-\vec{\alpha}}\, |E_{\vec{\alpha}}\rangle = |\, [E_{-\vec{\alpha}}, E_{\vec{\alpha}}]\rangle$, we have proved that

$$|\, [E_{-\vec{\alpha}}, E_{\vec{\alpha}}]\rangle = -\alpha_i\, |H_i\rangle\,, \tag{3.186}$$

and hence that

$$[E_{\vec{\alpha}}, E_{-\vec{\alpha}}] = \alpha_i\, H_i\,, \tag{3.187}$$

which we had asserted previously in (3.157).

Now, let us return again to the consideration of an arbitrary representation D of the Lie algebra $\mathcal{G}$. Consider

$$\begin{aligned} \langle \vec{\mu}, D|\, [E_{\vec{\alpha}}, E_{-\vec{\alpha}}]\, |\vec{\mu}, D\rangle &= \alpha_i\, \langle \vec{\mu}, D|H_i|\vec{\mu}, D\rangle \\ &= \alpha_i\, \mu_i\, \langle \vec{\mu}, D|\vec{\mu}, D\rangle \\ &= \vec{\alpha}\cdot\vec{\mu}\,. \end{aligned} \tag{3.188}$$

On the other hand, we have

$$\begin{aligned} \langle \vec{\mu}, D|\, [E_{\vec{\alpha}}, E_{-\vec{\alpha}}]\, |\vec{\mu}, D\rangle &= \langle \vec{\mu}, D|\, E_{\vec{\alpha}}\, E_{-\vec{\alpha}}\, |\vec{\mu}, D\rangle - \langle \vec{\mu}, D|\, E_{-\vec{\alpha}}\, E_{\vec{\alpha}}\, |\vec{\mu}, D\rangle \\ &= |N_{-\vec{\alpha},\vec{\mu}}|^2 - |N_{\vec{\alpha},\vec{\mu}}|^2\,, \end{aligned} \tag{3.189}$$

(see (3.182)), and so we have

$$\vec{\alpha}\cdot\vec{\mu} = |N_{-\vec{\alpha},\vec{\mu}}|^2 - |N_{\vec{\alpha},\vec{\mu}}|^2\,. \tag{3.190}$$

We also have

$$\begin{aligned} N_{-\vec{\alpha},\vec{\mu}} &= \langle \vec{\mu}-\vec{\alpha}, D|E_{-\vec{\alpha}}|\vec{\mu}, D\rangle \\ &= \langle \vec{\mu}-\vec{\alpha}, D|E_{\vec{\alpha}}^\dagger|\vec{\mu}, D\rangle \\ &= \langle \vec{\mu}, D|E_{\vec{\alpha}}|\vec{\mu}-\vec{\alpha}, D\rangle^* \\ &= \bar{N}_{\vec{\alpha},\vec{\mu}-\vec{\alpha}}\,, \end{aligned} \tag{3.191}$$

and so (3.190) gives

$$|N_{\vec{\alpha},\vec{\mu}-\vec{\alpha}}|^2 - |N_{\vec{\alpha},\vec{\mu}}|^2 = \vec{\alpha}\cdot\vec{\mu}\,. \tag{3.192}$$

Since we are assuming that the representation D is finite dimensional, it must be that if we apply $E_{\vec{\alpha}}$ or $E_{-\vec{\alpha}}$ repeatedly we must eventually get zero, since each application adds or subtracts $\vec{\alpha}$ to the weight $\vec{\mu}$. (This is the direct analogue of the argument for $SU(2)$ that repeated application of J_+ or J_- on a finite-dimensional state must eventually give zero.) Suppose, then, that for some non-negative integers p and q, we have that

$$|\vec{\mu}+p\,\vec{\alpha},D\rangle \neq 0\,,\qquad |\vec{\mu}-q\,\vec{\alpha},D\rangle \neq 0\,, \tag{3.193}$$

but that

$$E_{\vec{\alpha}}\,|\vec{\mu}+(p+1)\,\vec{\alpha},D\rangle = 0\,,\qquad E_{-\vec{\alpha}}\,|\vec{\mu}-(q+1)\,\vec{\alpha},D\rangle = 0\,. \tag{3.194}$$

It therefore follows from (3.182) that

$$N_{\vec{\alpha},\vec{\mu}+p\,\vec{\alpha}} = 0\,,\qquad N_{-\vec{\alpha},\vec{\mu}-q\,\vec{\alpha}} = 0\,, \tag{3.195}$$

and then using (3.191) the second of these equations implies

$$\bar{N}_{\vec{\alpha},\vec{\mu}-(q+1)\,\vec{\alpha}} = 0\,. \tag{3.196}$$

Now we can solve for the coefficients $N_{\vec{\alpha},\vec{\mu}}$, by following a strategy that is again precisely analogous to the one we used for $SU(2)$. From (3.192) we can write

$$\begin{aligned}
|N_{\vec{\alpha},\vec{\mu}+(p-1)\,\vec{\alpha}}|^2 - 0 &= \vec{\alpha}\cdot(\vec{\mu}+p\,\vec{\alpha})\,,\\
|N_{\vec{\alpha},\vec{\mu}+(p-2)\,\vec{\alpha}}|^2 - |N_{\vec{\alpha},\vec{\mu}+(p-1)\,\vec{\alpha}}|^2 &= \vec{\alpha}\cdot(\vec{\mu}+(p-1)\,\vec{\alpha})\,,\\
&\;\vdots\\
|N_{\vec{\alpha},\vec{\mu}}|^2 - |N_{\vec{\alpha},\vec{\mu}+\vec{\alpha}}|^2 &= \vec{\alpha}\cdot(\vec{\mu}+\,\vec{\alpha})\,,\\
|N_{\vec{\alpha},\vec{\mu}-\vec{\alpha}}|^2 - |N_{\vec{\alpha},\vec{\mu}}|^2 &= \vec{\alpha}\cdot\vec{\mu}\,,\\
&\;\vdots\\
|N_{\vec{\alpha},\vec{\mu}-q\,\vec{\alpha}}|^2 - |N_{\vec{\alpha},\vec{\mu}-(q-1)\,\vec{\alpha}}|^2 &= \vec{\alpha}\cdot(\vec{\mu}-(q-1)\,\vec{\alpha})\,,\\
0 - |N_{\vec{\alpha},\vec{\mu}-q\,\vec{\alpha}}|^2 &= \vec{\alpha}\cdot(\vec{\mu}-q\,\vec{\alpha})\,.
\end{aligned} \tag{3.197}$$

Adding up all these equations gives

$$\begin{aligned}
0 &= (p+q+1)\,\vec{\alpha}\cdot\vec{\mu} + \tfrac12\vec{\alpha}^2\,[p(p+1)-q(q+1)]\\
&= (p+q+1)\,[\vec{\alpha}\cdot\vec{\mu} + \tfrac12\vec{\alpha}^2\,(p-q)]\,,
\end{aligned} \tag{3.198}$$

and hence we conclude that

$$\frac{2\vec{\alpha}\cdot\vec{\mu}}{\vec{\alpha}^2} = -(p-q)\,. \tag{3.199}$$

In particular, note that the right-hand side is an integer.

One can straightfowardly obtain explicit expressions for all the $N_{\vec{\alpha},\vec{\mu}}$ from the above equations, but actually we shall not need them. The main result, which will be of very great importance, is (3.199).

First, we apply the general result (3.199) to the special case of the adjoint representation. This is especially important because the weights $\vec{\mu}$ are the roots of the algebra. We shall in general denote roots by early letters in the Greek alphabet, usually $\vec{\alpha}$ and $\vec{\beta}$. Since $H_i\,|E_{\vec{\beta}}\rangle = \beta_i\,|E_{\vec{\beta}}\rangle$, it follows from (3.199) that

$$\frac{2\vec{\alpha}\cdot\vec{\beta}}{\vec{\alpha}^2} = -(p-q) = m\,, \tag{3.200}$$

where we have simply defined the integer $m = q - p$. On the other hand, we could equally well have applied $E_{\pm\vec{\beta}}$ repeatedly to $|E_{\vec{\alpha}}\rangle$, rather than applying $E_{\pm\vec{\alpha}}$ repeatedly to $|E_{\vec{\beta}}\rangle$, and so we must also have

$$\frac{2\vec{\alpha}\cdot\vec{\beta}}{\vec{\beta}^2} = -(p'-q') = m'\,. \tag{3.201}$$

Multiplying (3.200) and (3.201) gives

$$\cos^2\theta = \tfrac{1}{4} m m'\,, \tag{3.202}$$

where

$$\cos\theta \equiv \frac{\vec{\alpha}\cdot\vec{\beta}}{|\vec{\alpha}|\,|\vec{\beta}|}\,, \tag{3.203}$$

and θ is the angle between the root vectors $\vec{\alpha}$ and $\vec{\beta}$. Since m and m' are integers, we have the very important result that only certain very special angles are possible:

mm'	θ
0	90^o
1	60^o, 120^o
2	45^o, 135^o
3	30^o, 150^o
4	0^o, 180^o

An implicit assumption in the discussion above was that for each root vector $\vec{\alpha}$ there is a unique operator $E_{\vec{\alpha}}$. In other words, we have been implicitly assuming that every generator has a different root vector. This is easily proved, by supposing that there could exist two independent generators $E_{\vec{\alpha}}$ and $E'_{\vec{\alpha}}$ with the same root vector $\vec{\alpha}$, and deriving a contradiction. Thus we begin by supposing that there exist independent states $|E_{\vec\alpha}\rangle$ and $|E'_{\vec\alpha}\rangle$ satisfying

$$H_i\,|E_{\vec\alpha}\rangle = \alpha_i\,|E_{\vec\alpha}\rangle\,,\qquad H_i\,|E'_{\vec\alpha}\rangle = \alpha_i\,|E'_{\vec\alpha}\rangle\,. \tag{3.204}$$

As usual, we can always choose our two hypothetically independent states so that

$$\langle E'_{\vec\alpha}|E_{\vec\alpha}\rangle = 0\,. \tag{3.205}$$

(If they were not orthogonal, we could define an orthogonal pair by the standard procedure of taking linear combinations – this is sometimes called Gram-Schmidt orthogonalisation.) Repeated application of $E_{\pm\vec\alpha}$ to $|E'_{\vec\alpha}\rangle$ shows, using (3.199), that

$$\frac{2\vec\alpha\cdot\vec\alpha}{\vec\alpha^2} = -(p-q) = 2\,. \tag{3.206}$$

However, we can furthermore show that $E_{-\vec\alpha}\,|E'_{\vec\alpha}\rangle = 0$, and hence that $q=0$. To see this, we note that it must be a zero-weight state (since $\vec\alpha-\vec\alpha=0$), and so

$$E_{-\vec\alpha}\,|E'_{\vec\alpha}\rangle = c_i\,|H_i\rangle \tag{3.207}$$

for some constants c_i. Therefore we have

$$\begin{aligned} c_i &= \langle H_i|\,E_{-\vec\alpha}\,|E'_{\vec\alpha}\rangle \\ &= \langle E'_{\vec\alpha}|\,E_{\vec\alpha}\,|H_i\rangle \\ &= \langle E'_{\vec\alpha}|\,[E_{\vec\alpha},H_i]\rangle \\ &= -\alpha_i\,\langle E'_{\vec\alpha}|E_{\vec\alpha}\rangle \\ &= 0\,, \end{aligned} \tag{3.208}$$

where the last step follows from the orthogonality (3.205). Thus we have proved that $q=0$, and so (3.206) gives

$$2 = -p\,, \tag{3.209}$$

which is a contradiction since by definition p is non-negative. Hence we conclude that there is only one generator with any given value for its root vector.

3.5.5 $SU(3)$ as an example

It will be helpful at this stage to consider an example in detail. After $SU(2)$, which, as we have seen, is not complicated enough to illustrate all the features of the general situation, the next simplest example is $SU(3)$. The algebra has dimension 8, and it can be represented by 3×3 Hermitean traceless matrices. A convenient basis is the set of so-called Gell-Mann matrices λ_a, which are an $SU(3)$ generalisation of the Pauli matrices of $SU(2)$. They are given by

$$\begin{aligned}
\lambda_1 &= \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \lambda_2 = \begin{pmatrix} 0 & -\mathrm{i} & 0 \\ \mathrm{i} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \\
\lambda_4 &= \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \qquad \lambda_5 = \begin{pmatrix} 0 & 0 & -\mathrm{i} \\ 0 & 0 & 0 \\ \mathrm{i} & 0 & 0 \end{pmatrix}, \\
\lambda_6 &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad \lambda_7 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\mathrm{i} \\ 0 & \mathrm{i} & 0 \end{pmatrix}, \\
\lambda_3 &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \lambda_8 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}.
\end{aligned} \tag{3.210}$$

By inspection, we can see that these provide a basis of Hermitian traceless 3×3 matrices. The two written on the bottom line, λ_3 and λ_8, are diagonal, and so they obviously commute with each other. In fact these are the maximal set of mutually-commuting Hermitian matrices, and so we can take them to define the Cartan subalgebra.

Defining generators $T_a = \frac{1}{2}\lambda_a$, we obtain the $SU(3)$ algebra with a canonical normalisation of the structure constants,

$$[T_a, T_b] = \mathrm{i}\, f_{ab}{}^c\, T_c\,. \tag{3.211}$$

Obviously, we can work out all the $f_{ab}{}^c$ if we wish, simply by slogging out the evaluation of all the commutators. Note that the T_a have been normalised so that

$$\mathrm{tr}(T_a T_b) = \tfrac{1}{2}\delta_{ab}\,. \tag{3.212}$$

It should also be noted that T_1, T_2 and T_3 generate an $SU(2)$ subalgebra. This is obvious from the fact that λ_1, λ_2 and λ_3 are just of the form

$$\lambda_a = \begin{pmatrix} \sigma_a & 0 \\ 0 & 1 \end{pmatrix}, \tag{3.213}$$

where σ_a are the 2×2 Pauli matrices. Note also that the pair (λ_4, λ_5) are very similar to the pair (λ_1, λ_2), except for their non-zero entries being in the 13 and 31 positions in the matrix, rather than 12 and 21. Likewise the pair (λ_5, λ_6) are also similar, excet that they have their non-zero entries instead in the 23 and 32 positions in the matrix.

As we already indicated, we shall take the Cartan subalgebra generators H_i to be

$$H_1 = T_3\,, \qquad H_2 = T_8\,. \tag{3.214}$$

Since there are two of them, $SU(3)$ has rank 2.

The Gell-Mann matrices provide a 3-dimensional representation of $SU(3)$. As we know from our earlier discussion, we can just think of the group $SU(3)$ in terms of 3×3 special unitary matrices acting on a 3-dimensional vector space, which, at the infinitesimal level, becomes Hermitian traceless matrices acting on the vector space. We can also view the vectors in the vector space as states.

A convenient basis of vectors for the 3-dimensional representation of $SU(3)$ is therefore simply

$$V_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}\,, \qquad V_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}\,, \qquad V_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}\,. \tag{3.215}$$

Their eigenvalues under H_1 and H_2 can be read off by inspection, since H_1 and H_2 are diagonal. Thus the eigenvalues under (H_1, H_2) for the vectors V_1, V_2 and V_3 are

$$V_1: \quad (\tfrac{1}{2}, \tfrac{1}{2\sqrt{3}})\,, \qquad V_2: \quad (-\tfrac{1}{2}, \tfrac{1}{2\sqrt{3}})\,, \qquad V_3: \quad (0, -\tfrac{1}{\sqrt{3}})\,. \tag{3.216}$$

We can write the states corresponding to V_1, V_2 and V_3 as

$$V_1 = |\tfrac{1}{2}, \tfrac{1}{2\sqrt{3}}\rangle\,, \qquad V_2 = |-\tfrac{1}{2}, \tfrac{1}{2\sqrt{3}}\rangle\,, \qquad V_3 = |0, -\tfrac{1}{2\sqrt{3}}\rangle\,, \tag{3.217}$$

where we are labelling the states by their weights under (H_1, H_2). This three-dimensional representation is called the **3** of $SU(3)$. It can also be called the *Defining Representation*, since it is the basic representation arising from the definition of the $SU(3)$ algebra in terms of Hermitean traceless matrices acting on a three-dimensional complex vector space.

We can plot the weights of the states in the **3** of $SU(3)$ on the plane, with axes corresponding to the weights under H_1 and H_2 respectively. The result, called the *Weight Diagram* for the **3** representation, is depicted in Figure 4 below.

We have already seen from the general discussion that the raising and lowering operators $E_{\vec\alpha}$ must take us between states in a representation. Thus we are led to define

$$\begin{aligned} E_{1,0} &= \tfrac{1}{\sqrt{2}}(T_1 + \mathrm{i}\, T_2) = \tfrac{1}{\sqrt{2}} \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}\,, \\ E_{-1,0} &= \tfrac{1}{\sqrt{2}}(T_1 - \mathrm{i}\, T_2) = \tfrac{1}{\sqrt{2}} \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}\,, \end{aligned}$$

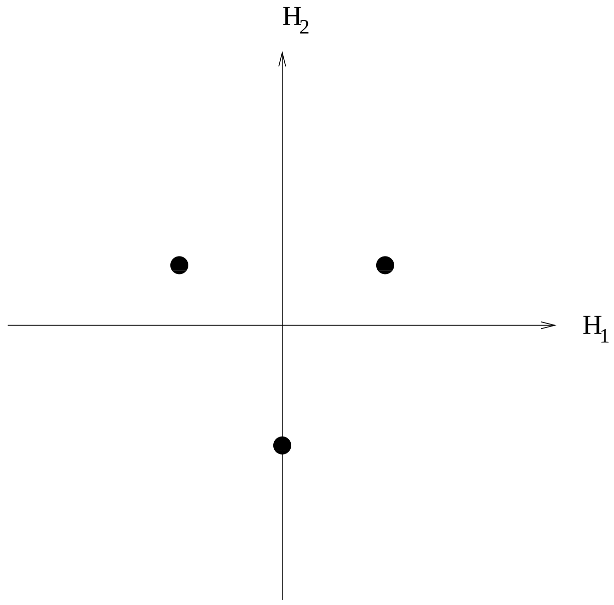

Figure 4: The weight diagram for the **3** representation of $SU(3)$

$$
\begin{aligned}
E_{\frac{1}{2},\frac{\sqrt{3}}{2}} &= \tfrac{1}{\sqrt{2}}(T_3 + \mathrm{i}\,T_4) = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \\
E_{-\frac{1}{2},-\frac{\sqrt{3}}{2}} &= \tfrac{1}{\sqrt{2}}(T_3 - \mathrm{i}\,T_4) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \\
E_{-\frac{1}{2},\frac{\sqrt{3}}{2}} &= \tfrac{1}{\sqrt{2}}(T_5 + \mathrm{i}\,T_6) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \\
E_{\frac{1}{2},-\frac{\sqrt{3}}{2}} &= \tfrac{1}{\sqrt{2}}(T_5 - \mathrm{i}\,T_6) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.
\end{aligned}
\tag{3.218}
$$

The action of these matrices on the three basis vectors V_i defined in (3.215) is easily seen by inspection. For example, we have

$$E_{-1,0}\, V_1 = \tfrac{1}{\sqrt{2}}\, V_2\,. \tag{3.219}$$

The reason for writing it with the $(-1,0)$ 2-vector subscript is therefore clear; it has taken a state with weight $\vec\mu = (\frac{1}{2}, \frac{1}{2\sqrt{3}})$ into a state with weight $(-\frac{1}{2}, \frac{1}{2\sqrt{3}})$. We know in general that

$$E_{\vec\alpha}\, |\vec\mu\rangle = N_{\vec\alpha,\vec\mu}\, |\vec\mu + \vec\alpha\rangle\,, \tag{3.220}$$

and so in this case we can deduce that the operator $E_{-1,0}$ has weight $\vec{\alpha} = (-\frac{1}{2}, \frac{1}{2\sqrt{3}}) - (\frac{1}{2}, \frac{1}{2\sqrt{3}}) = (-1, 0)$. One can similarly check for all the other combinations defined in (3.218) that the subscript label is simply the root vector $\vec{\alpha}$ associated with that particular raising or lowering operator.

What we have now achieved is to reorganise the original 8 generators T_a of $SU(3)$ into two Cartan generators H_i, and the 6 raising and lowering operator combinations $E_{\vec{\alpha}}$ in (3.218). One can also directly verify by slogging out the commutators that these satisfy

$$[H_i, E_{\vec{\alpha}}] = \alpha_i\, E_{\vec{\alpha}}\,. \tag{3.221}$$

The six vectors $\vec{\alpha}$ are the six root vectors of the $SU(3)$ algebra. They are given by

$$\vec{\alpha} = (1, 0)\,, \quad (-1, 0)\,, \quad (\tfrac{1}{2}, \tfrac{1}{2}\sqrt{3})\,, \qquad (-\tfrac{1}{2}, -\tfrac{1}{2}\sqrt{3})\,, \quad (-\tfrac{1}{2}, \tfrac{1}{2}\sqrt{3})\,, \quad (\tfrac{1}{2}, -\tfrac{1}{2}\sqrt{3})\,. \tag{3.222}$$

They can be plotted in a weight diagram too. Since we are talking here of the adjoint representation, for which the weights of the $E_{\vec{\alpha}}$ are called the roots, the resulting weight diagram in this case is called the *Root Diagram* for $SU(3)$. It is depicted in Figure 5 below. As can be seen, the six roots lie at the vertices of a regular hexagon. Note, in particular, that the angles between adjacent roots are all 60^o. This is consistent with our findings from equation (3.202), which led to the discrete list of possible angles between root vectors given in the table below (3.202).

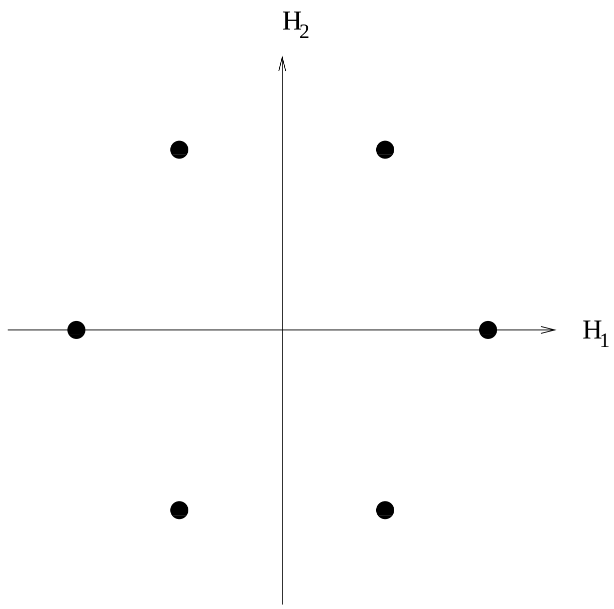

Figure 5: The root diagram for $SU(3)$

Note that the adjoint representation, being eight-dimensional, can also be denoted as the **8** of $SU(3)$. We could plot all eight weight vectors in a weight diagram too. It will, obviously, consist of the six points in the root diagram shown in Figure 5, together with two extra points sitting at the origin, corresponding to the zero-weight vectors of the two Cartan states $|H_1\rangle$ and $|H_2\rangle$.

3.5.6 Simple Roots

To begin, we introduce the notion of ordering weight vectors. To do this, we first define *positive* and *negative* weight vectors.

- A weight vector $\vec{\mu} = (\mu_1, \mu_2, \ldots, \mu_m)$ is said to be *positive* if its first non-zero component, working from the left, is positive. This is written $\vec{\mu} > 0$. Similarly, $\vec{\mu}$ is said to be *negative*, written as $\vec{\mu} < 0$, if its first non-zero component, working from the left, is negative. If all its components are zero, $\vec{\mu}$ has zero weight.

This definition might seem somewhat arbitrary, and indeed it is. For example, there is no pre-ordained or unique choice of what order to write the Cartan generators in. If we chose a different ordering for them, then this would amount to shuffling around the order of the components in all the weight vectors. What was a positive weight vector for one choice of labelling of the Cartan generators could be a negative weight vector for another choice of labelling. One could also perform redefinitions of the Cartan generators that amounted to more than just rr-ordering them. One could make any non-singular redefinition involving taking linear combinations of the original set of Cartan generators, with real coefficients, and get another equally valid set.

It is true, therefore, that the definition of positivity and negativity of weight vectors is in that sense arbitrary. The point is, though, that in the end it doesn't matter. Although the specifics of which weight vectors may be positive, and which negative, can change under a change of basis for the Cartan generators, all the statements and theorems we are going to use will work equally well for any choice. The only important thing is that one must fix on a basis and then stick with it.

If $\vec{\mu}$ and $\vec{\nu}$ are two weight vectors, then we say that $\vec{\mu} > \vec{\nu}$ if $\vec{\mu} - \vec{\nu}$ is positive. Note that if $\vec{\mu} > \vec{\nu}$, and $\vec{\nu} > \vec{\lambda}$, then it follows that $\vec{\mu} > \vec{\lambda}$. This is easily proven from the definition of positivity.

We are now in a position to define the *highest weight* in a finite-dimensional representation, as the weight that is greater than the weights of any of the other states in the representation. As in the case of $SU(2)$, there is in general a *unique* highest-weight state in the representation. Starting from this state, the entire set of states in the representation can be built up, by acting with the lowering and raising operators $E_{\vec\alpha}$.

Now that we have defined the notion of positivity and negativity of weights, we are final in a position to define which amongst the $E_{\vec\alpha}$ are raising operators, and which are lowering operators. Recalling that the roots vectors $\vec\alpha$ are just the weights of the non-zero weight states in the adjoint representation, we define:

- $E_{\vec\alpha}$ is a raising operator if $\vec\alpha$ is positive, and $E_{\vec\alpha}$ is a lowering operator if $\vec\alpha$ is negative.

Since, as we saw, $E_{\vec\alpha}^\dagger = E_{-\vec\alpha}$, and since obviously if $\vec\alpha$ is positive then $-\vec\alpha$ is negative, it follows that the full set of root vectors splits into equal-sized subsets of positive root vectors and negative root vectors. For every positive root vector $\vec\alpha$, there is an equal and opposite negative root vector $-\vec\alpha$.

Note that if we act with $E_{\vec\alpha}$ on the highest-weight state in a representation, then we shall necessarily get zero if $\vec\alpha$ is positive. This follows from the fact that, as we saw earlier, acting with $E_{\vec\alpha}$ on a state $|\vec\mu\rangle$ with weight $\vec\mu$ gives a state proportional to $|\vec\mu+\vec\alpha\rangle$ with weight $\vec\mu+\vec\alpha$. Therefore if $\alpha > 0$ it follows that $\vec\mu+\vec\alpha > \vec\mu$, so if $|\vec\mu\rangle$ was already the highest-weight state in the representation, then $|\vec\mu+\vec\alpha\rangle$ cannot exist.

Next, we define the notion of a *simple root*:

- A *simple root* is a positive root that cannot be written as the sum of two positive roots.

The simple roots determine the entire structure of the group.

An important theorem that we can easily prove is that if $\vec\alpha$ and $\vec\beta$ are ant two simple roots, then $\vec\beta-\vec\alpha$ is not a root. The proof is as follows:

If $\vec\beta-\vec\alpha$ were a positive root, then from the identity

$$\vec\beta = \vec\alpha + (\vec\beta-\vec\alpha) \tag{3.223}$$

we would have that $\vec\beta$ can be written as the sum of two positive roots, which contradicts the fact that $\vec\beta$ is a simple root. Conversely, if $\vec\beta-\vec\alpha$ were a negative root, then from the identity

$$\vec\alpha = \vec\beta + (\vec\alpha-\vec\beta) \tag{3.224}$$

we would have that $\vec\alpha$ can be written as the sum of two positive roots, which contradicts the fact that $\vec\alpha$ is a simple root. If $\vec\beta-\vec\alpha$ is neither a positive root nor a negative root, then it is not a root at all. This completes the theorem.

Having established that $\vec{\beta}-\vec{\alpha}$ is not a root if roots $\vec{\alpha}$ and $\vec{\beta}$ are simple roots, it follows that we must have

$$E_{-\vec{\alpha}}\,|E_{\vec{\beta}}\rangle = 0\,. \tag{3.225}$$

Now recall the master formula (3.199), i.e.

$$\frac{2\vec{\alpha}\cdot\vec{\mu}}{\vec{\alpha}^2} = -(p-q)\,. \tag{3.226}$$

where the states $|\mu+p\vec{\alpha}\rangle$ and $|\mu-q\vec{\alpha}\rangle$ exist but $|\mu+(p+1)\vec{\alpha}\rangle$ and $|\mu-(q+1)\vec{\alpha}\rangle$ do not, where p and q are non-negative integers. Applying this to the state $|E_{\vec{\beta}}\rangle$ in the adjoint representation, so $\vec{\mu}=\vec{\beta}$, we have from (3.225) that $q=0$, and hence

$$\frac{2\vec{\alpha}\cdot\vec{\beta}}{\vec{\alpha}\cdot\vec{\alpha}} = -p\,, \tag{3.227}$$

for any pair of simple roots $\vec{\alpha}$ and $\vec{\beta}$.

Knowing the integer p for each pair of simple roots $\vec{\alpha}$ and $\vec{\beta}$ determines the angles between all the simple roots, and the relative lengths of the simple roots. Recall once again that p is the integer that tells us how many times we can commute $E_{\vec{\alpha}}$ with $E_{\vec{\beta}}$ before we get zero.

We can, of course, interchange the roles of $\vec{\alpha}$ and $\vec{\beta}$ in the above discussion. If the state $|\vec{\beta}+p'\,\vec{\alpha}\rangle$ exists, but $|\vec{\beta}+(p'+1)\,\vec{\alpha}\rangle$ does not, then we shall have

$$\frac{2\vec{\alpha}\cdot\vec{\beta}}{\vec{\beta}^2} = -p'\,. \tag{3.228}$$

Multiplying (3.225) by (3.228), we find that

$$\cos\theta = -\tfrac{1}{2}\sqrt{p\,p'}\,,\qquad \frac{|\vec{\beta}|}{|\vec{\alpha}|} = \sqrt{\frac{p}{p'}}\,, \tag{3.229}$$

where θ is the angle between $\vec{\alpha}$ and $\vec{\beta}$. Note that we have

$$\frac{\pi}{2}\le\theta<\pi\,,\qquad \text{i.e.}\quad \vec{\alpha}\cdot\vec{\beta}\le 0\,. \tag{3.230}$$

(We cannot have $\theta=\pi$, because that would imply $\vec{\beta}$ was a positive multiple of $-\vec{\alpha}$, which is impossible if $\vec{\alpha}$ and $\vec{\beta}$ are both positive. In fact from (3.229), the allowed angles between a pair of simple roots are

$$\theta=\frac{\pi}{2}\,,\qquad \frac{2\pi}{3}\,,\qquad \frac{3\pi}{4}\,,\qquad \frac{5\pi}{6}\,, \tag{3.231}$$

or, in degrees,

$$\theta = 90^o\,,\qquad 120^o\,,\qquad 135^o\,,\qquad 150^o\,. \tag{3.232}$$

We can prove that the simple roots have the following properties:

(**1**) *The simple roots of a Lie algebra are linearly independent.*

To see this, label the simple roots by $\vec{\alpha}_a$. Note that a here is an index that labels each simple root; let us suppose there are N of them. Suppose that the simple roots were linearly dependent. Then, for some coefficients c_a, we would have the relation

$$\sum_{a=1}^{N} c_a\, \vec{\alpha}_a = 0\,. \tag{3.233}$$

Now, in general some of the constants c_a will be positive, and some will be negative. Divide the summation into two separate sums over these cases, and define

$$\vec{y} = \sum_{a \text{ with } c_a>0} c_a\, \vec{\alpha}_a\,, \qquad \vec{z} = \sum_{a \text{ with } c_a<0} (-c_a)\, \vec{\alpha}_a\,. \tag{3.234}$$

Equation (3.233) is now expressed as

$$\vec{y} = \vec{z}\,. \tag{3.235}$$

Now clearly, from the construction, $\vec{y}$ and $\vec{z}$ are both positive, since they are each the sum of positive (in fact simple) roots with positive coefficients. From (3.235) we get

$$\vec{y}^{\,2} = \vec{y}\cdot\vec{z}\,. \tag{3.236}$$

Since we have shown above that $\vec{\alpha}\cdot\vec{\beta} \le 0$ for any pair of simple roots, its follows that $\vec{y}\cdot\vec{z} \le 0$. Thus from (3.236) we obtain a contradiction, since $\vec{y}^{\,2} \ge 0$, with equality if and only if $\vec{y} = 0$, which it clearly cannot be. The conclusion is that the supposition (3.233) of a linear dependence among the simple roots is false. Hence, we have proved that the simple roots must all be linearly independent.

(**2**) *Any positive root $\vec{\gamma}$ can be written as a sum of simple roots $\vec{\alpha}_a$, with non-negative integer coefficients k_a,* i.e. $\vec{\gamma} = \sum_a k_a\, \vec{\alpha}_a$.

If $\vec{\gamma}$ is itself simple, the statement is obviously true. If $\vec{\gamma}$ is not simple, then it must be possible to split it as $\vec{\gamma} = \vec{\gamma}_1 + \vec{\gamma}_2$, where $\vec{\gamma}_1$ and $\vec{\gamma}_2$ are positive roots. (Recall that the simple roots are those positive roots that cannot be written as the sum of positive roots. Therefore, by definition, any positive root that is not simple *must* be expressible as a sum of positive roots.) If either $\vec{\gamma}_1$ or $\vec{\gamma}_2$ is not simple, then split them again. Continuing iteratively, one must eventually end up with $\vec{\gamma}$ decomposed as a sum over simple roots, with non-negative integer coefficients.

(**3**) *The number of simple roots is m =rank $\mathcal{G}$*

To prove this, we first note that since the simple roots are m-component vectors (the weights of the corresponding root generator under the m Cartan generators H_i), there can be at most m of them. This is an immediate consequence of Property 1 above, where we showed that the simple roots are all linearly independent.

Suppose, now, that there were less than m simple roots. We could then choose a basis for the Cartan generators so that all the simple roots had a 0 as their first component, i.e. $\vec{\alpha} = (0, \alpha_2, \ldots, \alpha_m)$ for every simple root $\vec{\alpha}$. It would then follow that the first component of *every* root vector would be zero, since they are all expressible as sums over simple roots. In other words, we would have that

$$[H_1, E_{\vec{\gamma}}] = 0 \tag{3.237}$$

for every raising or lowering operator $E_{\vec{\gamma}}$. Of course we also have $[H_1, H_i] = 0$ for all i. The conclusion would then be that H_1 would commute with *every* generator in the Lie algebra $\mathcal{G}$. In other words, H_1 would be a generator of an abelian invariant subalgebra. But at the outset of our discussion of the classification, we agreed to exclude such cases, and only classify the simple Lie algebras. Therefore, we conclude that for a simple Lie algebra $\mathcal{G}$, the number of simple roots is equal to rank $\mathcal{G}$, i.e. it is equal to the number of Cartan generators.

Let us see now in more detail how we express the positive roots as sums of simple roots. In the process, we shall see how one builds up the entire Lie algebra from the knowledge of the simple roots.

Our task, then, is to discover which vectors

$$\vec{\gamma} = \sum_{\vec{\alpha}} k_{\vec{\alpha}}\, \vec{\alpha} \tag{3.238}$$

are positive root vectors in the algebra, where $\vec{\alpha}$ are the simple roots, and $k_{\vec{\alpha}}$ are non-negative integers.[15] The of all the positive roots, together with their negatives, form what is called the *Root System* of the Lie algebra $\mathcal{G}$. It is useful to define

$$k = \sum_{\vec{\alpha}} k_{\vec{\alpha}}\,, \tag{3.239}$$

which is called the *level number* of the particular root $\vec{\gamma}$ that is constructed in (3.238). We can then in turn consider roots at level 1, level 2, level 3, and so on.

[15]Hopefully by now it will be completely clear what is meant by equation (3.238). We are using $\vec{\alpha}$ as a generic vector label to denote the set of all simple roots.

First, we note that the roots at level 1 are just the simple roots themselves, since it must be that all the $k_{\vec\alpha}$ are zero except for one of them, which equals 1.

At $k = 2$, there would appear to be two possibilties. We could satisfy (3.239) with $k = 2$ either by having all $k_{\vec\alpha} = 0$ except for one simple root, say $\vec\alpha_1$, for which $k_{\vec\alpha_1} = 2$. Or, we could satisfy it by having all $k_{\vec\alpha} = 0$ except for two different simple roots, say $\vec\alpha_1$ and $\vec\alpha_2$, with $k_{\vec\alpha_1} = k_{\vec\alpha_2} = 1$. The first of these possibilities does not arise; in other words, there is never a positive root given by $2\vec\alpha_1$, where $\vec\alpha_1$ is a simple root. The proof is easy. To get a state with weight $2\vec\alpha_1$ in the adjoint representation, we would build it as $E_{\vec\alpha_1}|E_{\vec\alpha_1}\rangle$. But

$$E_{\vec\alpha_1}|E_{\vec\alpha_1}\rangle = |[E_{\vec\alpha_1}, E_{\vec\alpha_1}]\rangle = 0\,, \tag{3.240}$$

since any generator commutes with itself.

At $k = 2$ we are left, therefore, with the possibility that we make a state $E_{\vec\alpha_1}|E_{\vec\alpha_2}\rangle$, where $\vec\alpha_1$ and $\vec\alpha_2$ are two different simple roots. We now use the master formula

$$\frac{2\vec\alpha_1\cdot\vec\alpha_2}{\vec\alpha_1^2} = -(p-q) \tag{3.241}$$

(see (3.200)). Recall that we showed previously that if $\vec\alpha_1$ and $\vec\alpha_2$ are simple roots, then $\vec\alpha_2 - \vec\alpha_1$ is not a root. Therefore $q = 0$ in (3.241). If $\vec\alpha_1\cdot\vec\alpha_2 < 0$ (i.e. it is strictly less then zero), then $\vec\alpha_1 + \vec\alpha_2$ must be a root, since (3.241) therefore implies that $p > 0$, and hence

$$p \ge 1\,. \tag{3.242}$$

Recall the significance of p; we know from the construction of the master formula that $\vec\alpha_2 + p\,\vec\alpha_1$ *is* a root, but $\vec\alpha_2 + (p+1)\vec\alpha_1$ is not. Without further knowledge about the details of the algebra we don't know in this case, where we are supposing that $\vec\alpha_1\cdot\vec\alpha_2 < 0$, whether $p = 1$ or $p > 1$, but we do know that p is at least 1, and so we know that $\vec\alpha_2 + \vec\alpha_1$ is a root. If in fact $p > 1$, then we would have that $\vec\alpha_2 + 2\vec\alpha_1$ is a root also. Of course the knowledge of the details of the algebra that we need in order to make a definite statement about the value of p is to know exactly what $\vec\alpha_1\cdot\vec\alpha_2$ is equal to, and what $\vec\alpha^2$ is equal to.

For $k \ge 3$ the process of building up the root system obviously gets more and more complicated. Suppose we have found all the roots up to and including level $k = n$. A vector $\vec\gamma + \vec\alpha$ at level $k = n+1$ is obtained by acting on a state at level $k = n$ having root vector $\vec\gamma$ with the simple root generator $E_{\vec\alpha}$. Is this new vector $\vec\gamma + \vec\alpha$ a root vector? Again, we use the master formula, to get

$$\frac{2\vec\alpha\cdot\vec\gamma}{\vec\alpha^2} = -(p-q)\,. \tag{3.243}$$

Unlike at level 2, we no longer know in general that $q = 0$, since $\vec\gamma$ is not simple. But we can determine q by looking at all the roots we have built up so far at levels $k < n$. Knowing q, and knowing the value of $\vec{\alpha}\cdot\vec{\gamma}$, we will therefore be able to calculate p. If $p>0$ (strictly), then we will know that $\vec{\gamma}+\vec{\alpha}$ is a root.

Proceeding in this way, we can build up all the roots at the level $k=n+1$, and then pass on to the next level, $k=n+2$. We continue this process until all the roots have been found. The endpoint of the process is when one has reached some level number at which one fails to find any further roots. The procedure terminates here, and teh task is complete.

It should be clear from this discussion that the key to everything is the master formula (3.200). The only information we need to know is the lengths of the simple roots, and the angles between the simple roots. Everything else then follows mechanically, and the entire root system can be worked out. Note that we do not even need to know the basis for the Cartan generators; i.e. we do not need to know the specific components of the simple root vectors. Only the lengths, and the angles, are important.

Let us return to our $SU(3)$ example at this point. We had the list (3.222) of the six root vectors, which we can write more succinctly as

$$\vec{\alpha}=(\pm 1,0)\,,\qquad \pm(\tfrac{1}{2},\tfrac{1}{2}\sqrt{3})\,,\qquad \pm(\tfrac{1}{2},-\tfrac{1}{2}\sqrt{3})\,. \tag{3.244}$$

Using the rule that a root is positive if its first non-zero component, working from the left, is positive, we see that the three positive roots are

$$(1,0)\,,\qquad (\tfrac{1}{2},\tfrac{1}{2}\sqrt{3})\,,\qquad (\tfrac{1}{2},-\tfrac{1}{2}\sqrt{3})\,. \tag{3.245}$$

We know that since $SU(3)$ has rank 2 (there are 2 Cartan generators), it must have 2 simple roots. This example is sufficiently elementary that we can spot the simple roots by inspection; they are

$$\vec{\alpha}_1=(\tfrac{1}{2},\tfrac{1}{2}\sqrt{3})\,,\qquad \vec{\alpha}_2=(\tfrac{1}{2},-\tfrac{1}{2}\sqrt{3})\,. \tag{3.246}$$

Clearly the third positive root is given by

$$\vec{\alpha}_1+\vec{\alpha}_2=(1,0)\,. \tag{3.247}$$

From (3.246) we see that

$$\vec{\alpha}_1^2=\vec{\alpha}_2^2=1\,,\qquad \vec{\alpha}_1\cdot\vec{\alpha}_2=-\tfrac{1}{2}\,. \tag{3.248}$$

This means that the angle between the two simple roots is 120^o.

Of course in this $SU(3)$ example we had the advantage of already having constructed the algebra, and so we already knew the entire root system. As a practice for what we shall be doing later, when we come at an algebra “from the other end” and start out knowing only the properties of its simple roots, let us pretend for $SU(3)$ that we know only the information given in (3.248).

We now try to build the entire $SU(3)$ root system, using the procedure outlined above. Thus at level $k=1$ we have the two simple roots $\vec{\alpha}_1$ and $\vec{\alpha}_2$. At level 2, we can only consider $\vec{\gamma}=\vec{\alpha}_1+\vec{\alpha}_2$. Is this a root? We plug into the master formula (3.241), knowing that $q=0$, and we get, using the results in (3.248), that

$$\frac{2\vec{\alpha}_1\cdot\vec{\alpha}_2}{\vec{\alpha}_1^2} = -1 = -p\,, \tag{3.249}$$

and hence $\vec{\alpha}_2+\vec{\alpha}_1$ is a root, but $\vec{\alpha}_2+2\vec{\alpha}_1$ is not. We could run the argument round the other way, exchanging the roles of $\vec{\alpha}_1$ and $\vec{\alpha}_2$, and thereby deduce that $\vec{\alpha}_1+2\vec{\alpha}_2$ is not a root either. Thus we have already learnt that for $SU(3)$ we have roots:

$$\begin{aligned}
&\text{Level } k=1: \qquad \vec{\alpha}_1\,, \quad \vec{\alpha}_2\\
&\text{Level } k=2: \qquad \vec{\alpha}_1+\vec{\alpha}_2\\
&\text{Level } k=3: \qquad \text{Nothing}
\end{aligned} \tag{3.250}$$

Once one has found no vectors at all at a given level, the process terminates; all the positive roots have been found.

3.5.7 Dynkin Diagrams

We have seen that once the lengths of the simple roots are known, and the angles between them, then the entire root system can be determined. Once the root system is known, the entire Lie algebra is known,

$$\begin{aligned}
&[H_i,H_j]=0\,, \qquad [H_i,E_{\vec{\alpha}}]=\alpha_i\,E_{\vec{\alpha}}\,,\\
&[E_{\vec{\alpha}},E_{-\vec{\alpha}}]=\alpha_i H_i\,, \qquad [E_{\vec{\alpha}},E_{\vec{\beta}}]=\mathcal{N}_{\vec{\alpha},\vec{\beta}}\,E_{\vec{\alpha}+\vec{\beta}}\,,
\end{aligned} \tag{3.251}$$

where $\vec{\alpha}\ne\vec{\beta}$.[16]

It is useful, therefore, to have a compact way of summarising all the necessary information about the lengths and angles for the simple roots. This can be done in what is called

[16]To be precise, with what we have studied so far we will know which of the constants $\mathcal{N}_{\vec{\alpha},\vec{\beta}}$ is zero and which is non-zero, since we will know the root system, so we will know which commutators $[E_{\vec{\alpha}},E_{\vec{\beta}}]$ produce non-zero results and which produce zero. If one needs to know the values of the non-vanishing constants $\mathcal{N}_{\vec{\alpha},\vec{\beta}}$, they can be worked out from the chain of equations we derived in section 3.5.4 by repeatedly acting with the raising and lowering operators.

a *Dynkin Diagram.* Recall that the angles θ between simple roots satisfy $\frac{1}{2}\pi \le \theta < \pi$, and that θ can only take the discrete values

$$\theta = \qquad 90^o\,, \qquad 120^o\,, \qquad 135^o\,, \qquad 150^o\,. \tag{3.252}$$

As we saw above, in the case of $SU(3)$ the angle between its two simple roots is 120^o.

In a Dynkin diagram, each simple root is represented by a circle. The angle between two simple roots is indicated by the number of lines joining the two roots, according to the rule

$$\begin{aligned} \theta = 90^o &\qquad \text{No lines} \\ \theta = 120^o &\qquad \text{One line} \\ \theta = 135^o &\qquad \text{Two lines} \\ \theta = 150^o &\qquad \text{Three lines}\,. \end{aligned} \tag{3.253}$$

There is also the question of indicating the lengths of the simple roots. Although it is not immediately obvious, it turns out that in fact the simple roots in any given simple Lie algebra can only have two possible lengths. We can call these short roots and long roots. Thus we merely need a notation for each circle, representing a simple root, to indicate whether that particular root is a short one or a long one. This is done by filling in the circle (i.e. a black circle) if it is a short root, while leaving it hollow (i.e. a white circle) if it is a long root.[17]

In our $SU(3)$ example, the two simple roots have equal length. In all cases where all the roots have equal length, the convention is to call them long roots, and thus represent them by open circles. A Lie algebra where all the simple roots have the same length is called a *Simply-laced* Lie algebra.

The Dynkin diagram for $SU(2)$ will consist of just a single circle, since there is just one simple root:

$$\circ \tag{3.254}$$

For $SU(3)$, we have two simple roots, of equal length, and with an angle of 120^o between them. The $SU(3)$ Dynkin diagram is therefore

$$\circ\!\!-\!\!\!-\!\!\circ \tag{3.255}$$

[17] Of course, when drawing Dynkin diagrams on a blackboard there is a reversal of the roles, in the sense that a filled-in circle (a short root) will actually be white, while an open circle (a long root) will be black.

3.5.8 Fundamental Weights

Consider a simple Lie algebra $\mathcal{G}$ that has rank m. There are m simple roots, which we shall call $\vec{\alpha}_i$, with the index i that labels the simple roots ranging from $1 \le i \le m$.[18] Consider an arbitrary finite-dimensional irreducible representation D, and suppose that its highest-weight state is

$$|\vec{\mu}, D\rangle\,. \tag{3.256}$$

In other words, we have $H_i|\vec{\mu}, D\rangle = \mu_i\,|\vec{\mu}, D\rangle$, and $\vec{\mu}$ is bigger than the weights of any of the other states in the representation.

From its definition, it therefore follows that $\vec{\mu} + \vec{\gamma}$ is not a weight in the representation D, for any positive root $\vec{\gamma}$. In fact, it suffices to say that $\vec{\mu} + \vec{\alpha}_i$ is not a weight for any of the m simple roots $\vec{\alpha}_i$. Thus we have the statement

$$E_{\vec{\alpha}_i}\,|\vec{\mu}, D\rangle = 0\,. \tag{3.257}$$

Recall now the master formula (3.199), i.e.

$$\frac{2\vec{\alpha}\cdot\vec{\mu}}{\vec{\alpha}^2} = -(p-q)\,, \tag{3.258}$$

which was derived for any state $|\vec{\mu}, D\rangle$ in an irreducible representation D, and any root vector $\vec{\alpha}$, where the non-negative integers p and q were defined by the fact that $\vec{\mu} + p\,\vec{\alpha}$ and $\vec{\mu} - q\,\vec{\alpha}$ are weights in the representation, but $\vec{\mu} + (p+1)\,\vec{\alpha}$ and $\vec{\mu} - (q+1)\,\vec{\alpha}$ are not. Taking $\vec{\alpha}$ in (3.258) to be any of the simple roots $\vec{\alpha}_i$, it follows from (3.257) that $p = 0$ for each i, and so we have

$$\frac{2\vec{\alpha}_i\cdot\vec{\mu}}{\vec{\alpha}_i^2} = q_i \qquad \text{for each } i\,, \tag{3.259}$$

where the q_i are non-negative integers.

Since we have established that the $\vec{\alpha}_i$ are m linearly independent m-component vectors, it follows that the q_i specify the highest-weight vector $\vec{\mu}$ completely. Each set of non-negative integers q_i determines a highest-weight vector, and so each set of q_i specifies an irreducible representation D of the Lie algebra $\mathcal{G}$. The complete set of states in D are then built up by repeatedly acting on the highest-weight state $|\vec{\mu}, D\rangle$ with the lowering operators $E_{-\vec{\alpha}_i}$,

[18] Take care not to confuse the index i on $\vec{\alpha}_i$, which labels the different simple roots, and the index i that we typically use to label the components of a given vector, such as when we write $\vec{\mu} = (\mu_1, \mu_2, \ldots, \mu_m)$, and we call these components μ_i. The way to distinguish between the two usages is that if the index i appears on a quantity with an arrow, as in $\vec{\alpha}_i$, then it is labelling a set of vectors (such as simple roots), whereas if the index i appears on a quantity without an arrow, such as μ_i, it is labelling the components of a specific vector.

where $\vec{\alpha}_i$ are the simple roots. Needless to say, the master formula (3.258) will play a central role in working out what the full set of states in D are.

It is convenient at this point ot introduce the so-called *Fundamental Weight Vectors* $\vec{\mu}_i$, which are defined to be the highest-weight vectors corresponding to taking all of the $q_i = 0$ except for one, which is taken to be unity. There are m possible choices; we define

$$\begin{aligned}
\vec{\mu}_1 \quad &\leftrightarrow \quad q_i = (1,0,0,\ldots,0) \\
\vec{\mu}_2 \quad &\leftrightarrow \quad q_i = (0,1,0,\ldots,0) \\
&\vdots \\
\vec{\mu}_m \quad &\leftrightarrow \quad q_i = (0,0,\ldots,0,1)\,.
\end{aligned} \tag{3.260}$$

From their definition, it follows using (3.258) that

$$\frac{2\vec{\alpha}_i\cdot\vec{\mu}_j}{\vec{\alpha}_i^2} = \delta_{ij}\,. \tag{3.261}$$

The m irreducible representations whose highest-weight vectors are the fundamental weight vectors $\vec{\mu}_i$ are called the *Fundamental Representations* of the Lie algebra $\mathcal{G}$.

The highest-weight vector of any irreducible representation D, specified by a set of non-negative integers $q_i = (q_1, q_2, \ldots, q_m)$, is given by

$$\vec{\mu} = \sum_i q_i\,\vec{\mu}_i\,. \tag{3.262}$$

This way of characterising a representation by means of the integers q_i is known as describing it by means of the *Highest-weight* labelling.

3.5.9 Examples in $SU(3)$

As we saw already (see (3.246)), the simple root vectors for $SU(3)$ are given by

$$\vec{\alpha}_1 = (\tfrac{1}{2}, \tfrac{1}{2}\sqrt{3})\,, \qquad \vec{\alpha}_2 = (\tfrac{1}{2}, -\tfrac{1}{2}\sqrt{3})\,. \tag{3.263}$$

It follows from (3.261) that the two fundamental weight vectors for $SU(3)$ are given by

$$\vec{\mu}_1 = (\tfrac{1}{2}, \tfrac{1}{2\sqrt{3}})\,, \qquad \vec{\mu}_2 = (\tfrac{1}{2}, -\tfrac{1}{2\sqrt{3}})\,. \tag{3.264}$$

We have in fact already encountered the fundamental weight vector $\vec{\mu}_1$, when we looked at the 3-dimensional representation **3** of $SU(3)$. The three states are listed in (3.217), and it is evident by inspection that the state with the highest weight is $|\frac{1}{2}, \frac{1}{2\sqrt{3}}\rangle$. We now recognise the **3** representation as the one characterised by taking $q_i = (1,0)$.

Let us look now at the other fundamental representation of $SU(3)$, specified by $q_i = (0,1)$. Thus has $\vec{\mu}_2 = (\frac{1}{2}, -\frac{1}{2\sqrt{3}})$ as its highest-weight vector. The idea now is to build up the rest of the states in this representation. We know straight away that $\vec{\mu}_2 - \vec{\alpha}_2$ is a weight, but $\vec{\mu}_2 - \vec{\alpha}_1$ and $\vec{\mu}_2 - 2\vec{\alpha}_2$ are not. (This follows from the master formula (3.258), together with the previous observation that, by definition, $p_i = 0$ for any of the simple roots $\vec{\alpha}_i$ acting on the highest-weight states.) So we know there is a state

$$E_{-\vec{\alpha}_2}\, |\vec{\mu}_2\rangle = c\, |\vec{\mu}_2 - \vec{\alpha}_2\rangle\,, \tag{3.265}$$

for some non-zero constant c. Note that $\vec{\mu}_2 - \vec{\alpha}_2 = (0, 1/\sqrt{3})$. Now we descend a level, by acting on $|\vec{\mu}_2 - \vec{\alpha}_2\rangle$ with lowering operators. We know that $E_{-\vec{\alpha}_2}$ will annihilate it, since we established $\vec{\mu}_2 - 2\vec{\alpha}_2$ is not a weight.

The only remaining option is to act with $E_{-\vec{\alpha}_1}$. Applying the master formula (3.258), we have

$$\frac{2\vec{\alpha}_1 \cdot (\vec{\mu}_2 - \vec{\alpha}_2)}{\vec{\alpha}_1^2} = 1 = -(p - q)\,. \tag{3.266}$$

But $p = 0$, since $\vec{\mu}_2 - \vec{\alpha}_2 + \vec{\alpha}_1$ is not a weight. We can see this, i.e. that $E_{\vec{\alpha}_1}\, |\vec{\mu}_2 - \vec{\alpha}_2\rangle = 0$, by considering $E_{\vec{\alpha}_1}\, E_{-\vec{\alpha}_2}\, |\vec{\mu}_2\rangle$. Now, we know that $[E_{\vec{\alpha}_1}, E_{-\vec{\alpha}_2}] = 0$, since we proved in general that the difference of simple roots is never a root. Therefore, proving $E_{\vec{\alpha}_1}\, E_{-\vec{\alpha}_2}\, |\vec{\mu}_2\rangle = 0$ is equivalent to proving $E_{-\vec{\alpha}_2}\, E_{\vec{\alpha}_1}\, |\vec{\mu}_2\rangle = 0$, and this is obvious, since $|\vec{\mu}_2\rangle$ is the highest-weight state and so $E_{\vec{\alpha}_1}\, |\vec{\mu}_2\rangle = 0$. Having proved $p = 0$, equation (3.266) shows that $q = 1$, and so $\vec{\mu}_2 - \vec{\alpha}_2 - \vec{\alpha}_1$ is a weight.

Similar arguments show that $E_{-\vec{\alpha}_1}$ and $E_{-\vec{\alpha}_2}$ both annihilate $|\vec{\mu}_2 - \vec{\alpha}_2 - \vec{\alpha}_1\rangle$, and so the construction of the representation with highest-weight vector $\vec{\mu}_2$ is complete. It has three states, with weights

$$\vec{\mu}_2 = (\tfrac{1}{2}, -\tfrac{1}{2\sqrt{3}})\,, \qquad \vec{\mu}_2 - \vec{\alpha}_2 = (0, \tfrac{1}{\sqrt{3}})\,, \qquad \vec{\mu}_2 - \vec{\alpha}_2 - \vec{\alpha}_1 = (-\tfrac{1}{2}, -\tfrac{1}{2\sqrt{3}}) \tag{3.267}$$

These can be plotted on a weight diagram, depicted in Figure 6. We can see that it is just an upside-down version of the original **3** representation. For reasons that will become apparent, it is called the $\overline{\mathbf{3}}$ representation.

3.5.10 Weyl Reflections

Strictly speaking, our derivation of the three states of the $\overline{\mathbf{3}}$ representation of $SU(3)$ is not yet complete. Although we could argue from the master formula (3.258) that no other weights could arise, it does still leave open the question of whether there might exist more

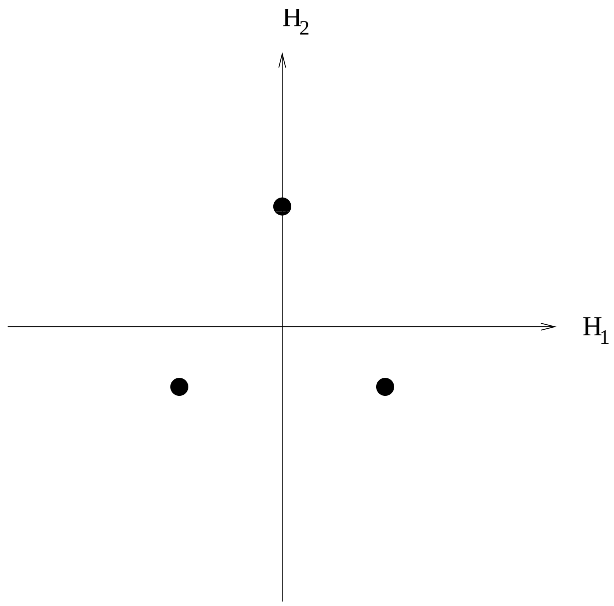

Figure 6: The weight diagram for the $\overline{\mathbf{3}}$ representation of $SU(3)$

than one state with any of the three weights that we found. One useful tool that helps to rule out such a possibility is called the *Weyl Reflection Formula*. This says the following:

If $\vec{\mu}$ is a weight of a state in an irreducible representation D, and $\vec{\alpha}$ is any root, then

$$\vec{\mu}' = \vec{\mu} - \Big(\frac{2\vec{\alpha}\cdot\vec{\mu}}{\vec{\alpha}^2}\Big)\,\vec{\alpha} \tag{3.268}$$

is also a weight of a state in the representation D. The proof consists of showing that if $\vec{\mu}'$ is substituted into the master formula, it passes the test of being a permissible weight. The key point is that, from (3.268), we have

$$\frac{2\vec{\alpha}\cdot\vec{\mu}'}{\vec{\alpha}^2} = -\frac{2\vec{\alpha}\cdot\vec{\mu}}{\vec{\alpha}^2}\,, \tag{3.269}$$

and hence if $\vec{\mu}$ passes the test for being a weight, then so does $\vec{\mu}'$.

A consequence of the Weyl reflection formula is that the degeneracy of states with weight $\vec{\mu}$ is identical to the degeneracy of states with weights $\vec{\mu}'$.

Notice that the Weyl reflection formula (3.268) constructs a weight $\vec{\mu}'$ by reflection of $\vec{\mu}$ in the hyperplane[19] orthogonal to the root $\vec{\alpha}$. To see this, consider the vector

$$\vec{v} = \vec{\mu} - \Big(\frac{\vec{\alpha}\cdot\vec{\mu}}{\vec{\alpha}^2}\Big)\,\vec{\alpha}\,. \tag{3.270}$$

Obviously we have $\vec{v}\cdot\vec{\alpha} = 0$. We also obviously have that $\vec{\mu}+\vec{\mu}'$ lies along $\vec{v}$; in fact from (3.268) and (3.270) we have $\vec{\mu}+\vec{\mu}' = 2\vec{v}$. This proves the assertion.

[19]In our $SU(3)$ example, which is rank 2, the roots and weights live in a 2-dimensional space, and so the "hyperplane" orthogonal to a root $\vec{\alpha}$ is a line.

The set of all reflections, for all the roots $\vec{\alpha}$, is called the *Weyl Group*. It is a discrete symmetry of the weight daigram.

In our $SU(3)$ example, take another look at the root diagram of Figure 5. Recall that we identified the simple roots $\vec{\alpha}_1$ and $\vec{\alpha}_2$ in (3.246), and the remaining positive root $\vec{\alpha}_1+\vec{\alpha}_2$ in (3.247). In Figure 5, $\vec{\alpha}_1$ is therefore the dot in the top right, and $\vec{\alpha}_2$ is the dot in the bottom right. $\vec{\alpha}_1+\vec{\alpha}_2$ is the dot in the middle right. The hyperplanes (i.e. lines) orthogonal to these vectors therefore comprise the H_2 axis (perpendicular to $\vec{\alpha}_1+\vec{\alpha}_2$), and lines at $+30^o$ and -30^o to the H_1 axis. In other words these lines make 60^o angles to each other. Imagine these lines as mirrors, and it is clear that a dot placed at a generic point on the weight diagram will acquire 5 image points under these reflections, making six dots in total, at the vertices of a regular hexagon. This is illustrated in Figure 7 below.

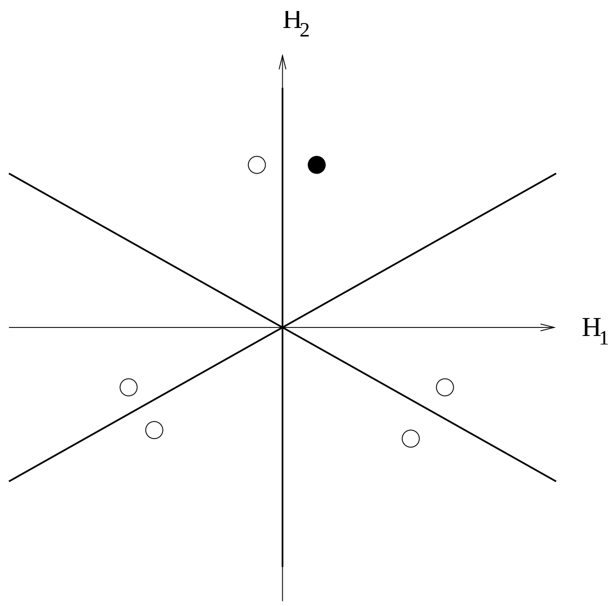

Figure 7: The Weyl reflection lines for $SU(3)$. The NW-SE line is perpendicular to $\vec{\alpha}_1$, the SW-NE line is perpendicular to $\vec{\alpha}_2$, and the vertical line is perpendicular to $\vec{\alpha}_1+\vec{\alpha}_2$. A generically-placed dot (the black circle) acquires 5 image points (the open circles) under the set of Weyl reflections

A dot that is placed at a special point, sitting actually on the surface of one of the mirrors, will only acquire 2 image points.

Now let us go back to our $\overline{\mathbf{3}}$ representation of $SU(3)$. Start with the highest-weight state $\vec{\mu}_2$, which is the bottom right vertex of the triangle in Figure 6. This is clearly at one of the special points, which sits on one of the mirrors. It therefore acquires just two image points, which are exactly the two other weights in the representation, which we already calculated.

This is illustrated in Figure 8 below.

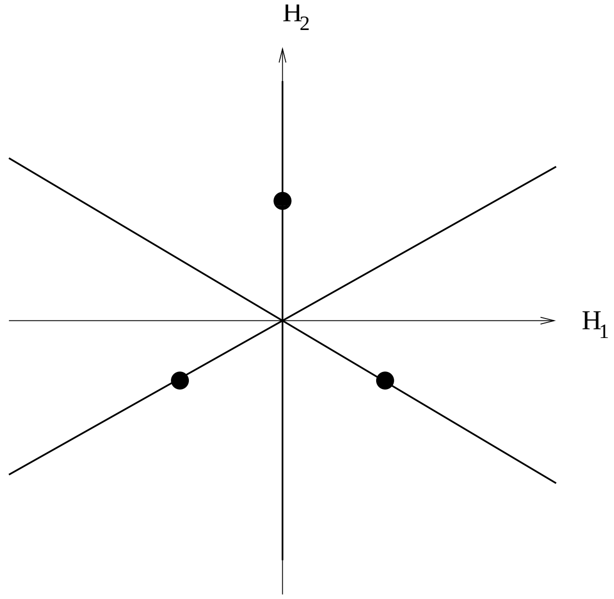

Figure 8: The highest weight (bottom right dot) of the $\overline{3}$ representation of $SU(3)$ acquires 2 image dots under Weyle reflections, filling out the entire irreducible representation.

At this stage, we know from our general discussion of the Weyl group that the degeneracies of all the Weyl-reflected weights are the same. Since for the $\overline{\mathbf{3}}$ representation we are able to get all three weights from the highest-weight state by reflection, we need only demonstrate that the highest-weight state is unique (i.e. not degenerate), and we will have proved that all three states are non-degenerate, and hence that there really are exactly three states and no more.

It is easy to prove, for any representation of any simple Lie algbera, that the highest-weight state is unique, i.e. that there is a unique state with the highest weight $\vec{\mu}$. Recall that, by definition, the states in a representation are built up by acting in all posible ways with raising and lowering operators on a state with the highest weight $\vec{\mu}$. Without yet assuming that there is a unique highest-weight state, consider one such, say $|\vec{\mu}\rangle$, and now build up the representation. It will be obtained as the set of all non-vanishing states of the form

$$E_{\vec{\gamma}_1}\, E_{\vec{\gamma}_2} \cdots E_{\vec{\gamma}_n}\, |\vec{\mu}\rangle\,, \qquad \text{for all } n\,, \tag{3.271}$$

where each of $(\vec{\gamma}_1, \vec{\gamma}_2, \ldots, \vec{\gamma}_n)$ is any of the root vectors of the algebra. We can immediately exclude all positive roots, since by repeated commutation we can move them to the right, where they hit $|\vec{\mu}\rangle$ and annihilate it. They get progressively less positive as they form commutators with negative-root generators along the way, but the net effect is that any non-vanishing state obtained in (3.271) is actually expressible as a (possibly smaller number) of purely negative-root generators acting on $|\vec{\mu}\rangle$. (This is an important point, and so if you don't immediately see it, try playing around with an example in order to see why it is true.)

Having established that we need only consider states in (3.271) where all the operators have negative root-vectors, it is now manifest that we will never build a second state with the weight $\vec{\mu}$ of the highest weight. Thus the highest-weight state is unique.

Going back to our example of the $\overline{\mathbf{3}}$ of $SU(3)$, this completes the proof that it really is a representation with exactly three states, i.e. with no degeneracy.

3.5.11 Complex Representations

Suppose T_a are the generators of a Lie algebra $\mathcal{G}$, in some represenation D. We have

$$[T_a, T_b] = \mathrm{i}\, f_{ab}{}^c\, T_c\,, \tag{3.272}$$

where the structure constants $f_{ab}{}^c$ are real. Complex conjugating, we therefore have

$$[T_a^*, T_b^*] = -\mathrm{i}\, f_{ab}{}^c\, T_c^*\,, \tag{3.273}$$

and so we see that the $(-T_a^*)$ generate the same algebra as T_a. The representation of $\mathcal{G}$ using the the generators $(-T_a^*)$ is called the complex conjugate of the representation D of the generators T_a, and it is denoted by $\overline{D}$.

If $\overline{D}$ is equivalent to D, then we say that D is a *real* representation. If $\overline{D}$ is inequivalent to D, then we say that D is a *complex* representation.

Suppose that $\vec{\mu}$ is a weight in the representation D, i.e. there is a corresponding state with

$$H_i\, |\vec{\mu}, D\rangle = \mu_i\, |\vec{\mu}, D\rangle\,. \tag{3.274}$$

It then follows that $\overline{D}$ has a corresponding state with weight $-\vec{\mu}$. The reason for this is that the Cartan generators for $\overline{D}$ are $-H_i^*$, and furthermore, we know that H_i has the same eigenvalues as H_i^*. This is because the Cartan generators are Hermitean, and so they eigenvalues are real. The upshot is that the *highest weight* of $\overline{D}$ is the negative of the *lowest weight* of D. Since the highest weight determines the entire representation, it follows that

D is real *if its lowest weight is the negative of its highest weight*

If this is not the case, the representation is complex.

Let us consider some examples. In the defining representation of $SU(3)$, i.e. the **3** representation depicted in Figure 4. Its highest weight is the fundamental weight vector $\vec{\mu}_1$; this corresponds to the top right dot in the weight diagram. Its lowest weight is the reflection across the H_2 axis, i.e. the dot on the top left, which is $-\vec{\mu}_2$, where $\vec{\mu}_2$ is the other fundamental weight vector of $SU(3)$. Manifestly, the lowest-weight vector is not the negative of the highest-weight vector, and so we conclude that the **3** of $SU(3)$ is a complex representation. Indeed, as we already saw, there is another three-dimensional representation which is just the upside-down version of the **3**, namely the $\overline{\mathbf{3}}$ that we constructed, and that is depicted in Figure 6. This has $\vec{\mu}_2$ as its highest weight.

In terms of the highest-weight labelling, where the representation is specified by the integers q_i and the highest weight is $\vec{\mu} = \sum_i q_i \, \vec{\mu}_i$, the **3** and the $\overline{\mathbf{3}}$ representations are the $(1,0)$ and the $(0,1)$ respectively. It is no coincidence that one is obtained from the other by exchanging the first and second components of their highest-weight labels.

Consider an $SU(3)$ representation defined by $q_i = (m,n)$. By definition, this has as its highest weight the vector

$$\vec{\mu} = m\,\vec{\mu}_1 + n\,\vec{\mu}_2\,, \tag{3.275}$$

where $\vec{\mu}_1$ and $\vec{\mu}_2$ are the fundamental weight vectors defined earlier, which satisfy $2\vec{\alpha} \cdot \vec{\mu}_j/(\vec{\alpha}_i^2) = \delta_{ij}$. It follows therefore that the lowest-weight state in the (m,n) representation has weight

$$-n\,\vec{\mu}_1 - m\,\vec{\mu}_2\,. \tag{3.276}$$

The highest weight of the complex conjugate representation $(m,n)^*$ is therefore given by

$$n\,\vec{\mu}_1 + m\,\vec{\mu}_2\,, \tag{3.277}$$

from which we see that $(m,n)^* = (n,m)$. It is now very easy to recognise which representations of $SU(3)$ are real, and which are complex: A representation (m,n) is *real* if $m=n$, and *complex* if $m \neq n$.

Let us discuss two more examples of $SU(3)$ representations before moving on to other matters. First, consider the representation $(1,1)$, which, in view of the above discussion, is real. In fact we already know this representation. By definition, its highest-weight vector is

$$\vec{\mu} = \vec{\mu}_1 + \vec{\mu}_2 = (1,0)\,. \tag{3.278}$$

(Recall that the fundamental weight vectors were given in (3.264).) Now, recall that when we studied the adjoint representation of $SU(3)$, we found that the three positive root vectors were the two simple roots $\vec{\alpha}_1$ and $\vec{\alpha}_2$ given in (3.246), and the vector $\vec{\alpha}_1 + \vec{\alpha}_2$. This last is obviously the highest-weight vector in the adjoint representation. From (3.246) we have

$$\vec{\alpha}_1 + \vec{\alpha}_2 = (1,0)\,. \tag{3.279}$$

Thus we see that the highest-weight vector of the $q_i = (1,1)$ representation is precisely the highest-weight vector of the adjoint representation. It follows that the $(1,1)$ representation *is* the adjoint representation.

If we didn't already know everything about the adjoint representation of $SU(3)$, we could easily construct it from the knowledge of the simple roots, and the highest-weight vector in (3.278). By definition, since $q_1 = 1$ and $q_2 = 1$, we know that $\vec{\mu} - \vec{\alpha}_1$ and $\vec{\mu} - \vec{\alpha}_2$ are weights but $\vec{\mu} - 2\alpha_1$ and $\vec{\mu} - 2\vec{\alpha}_2$ are not. Applying the master formula (3.258) to the weight $\vec{\mu} - \vec{\alpha}_1$, we find

$$\frac{2\vec{\alpha}_2 \cdot (\vec{\mu} - \vec{\alpha}_1)}{2\vec{\alpha}_2^2} = 2 = -(p - q)\,. \tag{3.280}$$

Since $\vec{\mu} - \vec{\alpha}_1 + \vec{\alpha}_2$ is not a weight (we know this because $\vec{\alpha}_1 - \vec{\alpha}_2$ is not a root), we have $p = 0$, and hence $q = 2$. This means $\vec{\mu} - \vec{\alpha}_1 - \vec{\alpha}_2$ and $\vec{\mu} - \vec{\alpha}_1 - 2\vec{\alpha}_2$ are weights, but $\vec{\mu} - \vec{\alpha}_1 - 3\vec{\alpha}_2$ is not. Interchanging the rôles of $\vec{\alpha}_1$ and $\vec{\alpha}_2$ when applying the master formula, we also learn that $\vec{\mu} - 2\vec{\alpha}_1 - \vec{\alpha}_2$ is a weight but $\vec{\mu} - 3\vec{\alpha}_1 - \vec{\alpha}_2$ is not. Finally, we find that $2\vec{\alpha}_2 \cdot (\vec{\mu} - 2\vec{\alpha}_1 - \vec{\alpha}_2)/(\vec{\alpha}_2^2) = 1$, and since $\vec{\mu} - 2\vec{\alpha}_1$ is not a weight, we have $q = 1$ and so $(\vec{\mu} - 2\vec{\alpha}_1 - 2\vec{\alpha}_2$ is a weight. Applying the master formula to all our newly-found weights, we discover that there can be no more, and the process has terminated. The weights we have found by this process comprise six non-zero weights, which live on the six vertices of the hexagon in Figure 5, and the weight $\vec{\mu} - \vec{\alpha}_1 - \vec{\alpha}_2 = 0$. This lives at the origin. It was not included in Figure 5 because there, we were specifically plotting the *roots*, i.e. the non-zero-weight vectors of the adjoint represenation. When we plot the *weight diagram* of the adjoint representation, which should of course include the zero-weight states too.

For the record, the non-zero weights found above lie in Figure 5 as follows. At the far right we have $\vec{\mu}$. Top-right is $\vec{\mu} - \vec{\alpha}_2$, and bottom-right is $\vec{\mu} - \vec{\alpha}_1$. Far-left is $\vec{\mu} - 2\vec{\alpha}_1 - 2\vec{\alpha}_2$; top-left is $\vec{\mu} - \vec{\alpha}_1 - 2\vec{\alpha}_2$; and bottom-left is $\vec{\mu} - 2\vec{\alpha}_1 - \vec{\alpha}_2$.

There is a small subtlety about the zero-weight vector $\vec{\mu} - \vec{\alpha}_1 - \vec{\alpha}_2$. There are actually two linearly-independent zero-weight states, which we can write as

$$E_{-\vec{\alpha}_1}\, E_{-\vec{\alpha}_2}\, |\vec{\mu}\rangle\,, \qquad \text{and} \qquad E_{-\vec{\alpha}_2}\, E_{-\vec{\alpha}_1}\, |\vec{\mu}\rangle\,. \tag{3.281}$$

The reason why these are independent is that the commutator $[E_{-\vec{\alpha}_1}, E_{-\vec{\alpha}_2}]$ is non-zero (it gives a constant times $E_{-\vec{\alpha}_1 - \vec{\alpha}_2}$), and so the two orderings of the lowering operators in (3.281) can, and indeed do, give different states. This can be proved by a rather simple argument.

It is, of course, precisely to be expected that there should be two linearly-independent zero-weight states in the adjoint representation; they are nothing but the Cartan states

$|H_1\rangle$ and $|H_2\rangle$.

For a second, and final example of an $SU(3)$ representation, consider the $q_i = (2, 0)$ representation, which is complex. Its highest-weight vector is $\vec{\mu} = 2\vec{\mu}_1 = (1, 1/\sqrt{3})$. Since $q_1 = 2$ we know $\vec{\mu} - \vec{\alpha}_1$ and $\vec{\mu} - 2\vec{\alpha}_1$ are weights but $\vec{\mu} - 3\vec{\alpha}_1$ is not. We can proceed again using the master formula (3.258), to build up all the weights. Recall that one can also make use of the Weyl reflection properties derived earlier. Either way, one soon arrives at the conclusion that there are six weights in the representation, namely

$$\begin{aligned} &\vec{\mu} = (1, \tfrac{1}{\sqrt{3}})\,, && \vec{\mu} - \vec{\alpha}_1 = (\tfrac{1}{2}, -\tfrac{1}{2\sqrt{3}})\,, \\ &\vec{\mu} - 2\vec{\alpha}_1 = (0, -\tfrac{2}{\sqrt{3}})\,, && \vec{\mu} - \vec{\alpha}_1 - \vec{\alpha}_2 = (0, \tfrac{1}{\sqrt{3}})\,, \\ &\vec{\mu} - 2\vec{\alpha}_1 - 2\vec{\alpha}_2 = (-1, \tfrac{1}{\sqrt{3}})\,, && \vec{\mu} - 2\vec{\alpha}_1 - \vec{\alpha}_2 = (-\tfrac{1}{2}, -\tfrac{1}{2\sqrt{3}})\,. \end{aligned} \tag{3.282}$$

The weight diagram for this six-dimensional representation is given in Figure 9 below.

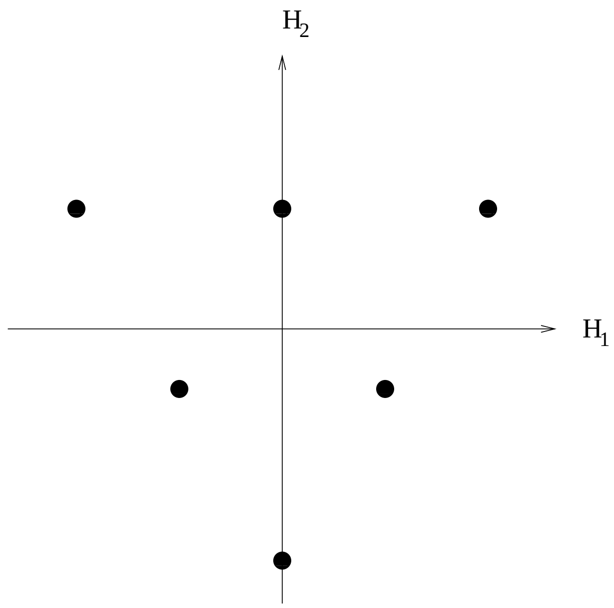

Figure 9: The **6** representation of $SU(3)$. The highest-weight state $\vec{\mu} = 2\vec{\mu}_1$ is at the top right of the triangle.

3.5.12 Two Theorems about $\vec{\alpha}$ Strings

As we have been seeing, one can extract an enormous amount of information from the "master formula" (3.258). In fact, it lies at the heart of the entire procedure for constructing irreducible representations of Lie algebras. Some of the things one learns from using (3.258) are specific to the details of the particular Lie algebra one is studying, as encoded in the information about the lengths of the simple roots and the angles between them. Other things are rather general, such as the results about the allowed angles between simple roots.

Here are two more general results, which can be derived from the master formula (3.258). It is useful to know these, when constructing the root system, because they can save a lot of time. They are both concerned with what we may call $\vec{\alpha}$ *Strings*, namely sets of roots in the root system that are obtained by adding or subtracting the root vector $\vec{\alpha}$ repeatedly to a given root vector $\vec{\gamma}$:

(1) *A string of roots* $\vec{\gamma} + k\,\vec{\alpha}$ *has no gaps.*

Imagine that we start with the integer k being sufficiently small (which could mean large and negative) that $\vec{\gamma} + k, \vec{\alpha}$ is not a root, and we increase k until we get an allowed root. We now keep increasing k until again we reach a vector that is not an allowed root. The theorem states that if we keep increasing k further, it is not possible to find any further allowed roots. In other words, an $\vec{\alpha}$ string of roots cannot have segments of allowed roots with a gap of disallowed vectors in between.

We can prove this by the time-honoured procedure of supposing the contrary, and arriving at a contradiction. Suppose, therefore, that we did have a gap in a string of allowed roots. We can characterise this by supposing that we have roots

$$\cdots, \qquad \vec{\beta} - 2\vec{\alpha}\,, \qquad \vec{\beta} - \vec{\alpha}\,, \qquad \vec{\beta}\,, \tag{3.283}$$

and then a gap where there are no roots, and then a further set of allowed roots

$$\vec{\beta}'\,, \qquad \vec{\beta}' + \vec{\alpha}\,, \qquad \vec{\beta}' + 2\vec{\alpha}\,, \qquad \cdots\,, \tag{3.284}$$

where

$$\vec{\beta}' = \vec{\beta} + n\,\vec{\alpha}\,, \qquad n \ge 2\,. \tag{3.285}$$

In particular, we are supposing that $\vec{\beta} + \vec{\alpha}$ is not a root, and $\vec{\beta}' - \vec{\alpha}$ is not a root.

Applying the master formula (3.258), these last two facts imply that

$$\begin{aligned} \frac{2\vec{\alpha}\cdot\vec{\beta}}{\vec{\alpha}^2} &= -(p-q) = q\,, \\ \frac{2\vec{\alpha}\cdot\vec{\beta}'}{\vec{\alpha}^2} &= -(p'-q') = -p'\,. \end{aligned} \tag{3.286}$$

Using (3.285) in the second of these equations, we therefore obtain

$$q + p' + 2n = 0\,. \tag{3.287}$$

This is a contradiction, since q and p' are non-negative, and $n \geq 2$. Therefore, the $\vec{\alpha}$-string cannot have gaps.

2) *No string of roots can have more than 4 roots in the chain*

Again, the proof is by contradiction. Suppose we had a string of 5 or more roots. Without loss of generality, we could then pick a root, let's call it $\vec{\beta}$, somewhere in the middle, such that we have roots

$$\cdots\,, \qquad \vec{\beta}-2\vec{\alpha}\,, \qquad \vec{\beta}-\vec{\alpha}\,, \qquad \vec{\beta}\,, \qquad \vec{\beta}+\vec{\alpha}\,, \qquad \vec{\beta}+2\vec{\alpha}\,, \qquad ,\cdots\,, \tag{3.288}$$

where $\vec{\alpha}$ is a simple root.

Now, we know that that

$$2\vec{\alpha} = (\vec{\beta}+2\vec{\alpha}) - \vec{\beta} \qquad \text{and} \qquad 2(\vec{\beta}+\vec{\alpha}) = (\vec{\beta}+2\vec{\alpha}) + \vec{\beta} \tag{3.289}$$

are not roots, since if $\vec{\gamma}$ is a root $2\vec{\gamma}$ can never be a root. (This follows from the fact that $[E_{\vec{\gamma}}, E_{\vec{\gamma}}] = 0$.) Applying the master formula (3.258), with $\vec{\alpha}$ replaced by $\vec{\beta}$, andd $\vec{\mu}$ replaced by $(\vec{\beta}+2\vec{\alpha})$, we know from (3.289) that $p = q = 0$ and so

$$\frac{\vec{\beta}\cdot(\vec{\beta}+2\vec{\alpha})}{\vec{\beta}^2} = 0\,. \tag{3.290}$$

By the same token, we know that

$$-2\vec{\alpha} = (\vec{\beta}-2\vec{\alpha}) - \vec{\beta} \qquad \text{and} \qquad 2(\vec{\beta}-\vec{\alpha}) = (\vec{\beta}-2\vec{\alpha}) + \vec{\beta} \tag{3.291}$$

are not roots, and so applying the master formula here we obtain

$$\frac{\vec{\beta}\cdot(\vec{\beta}-2\vec{\alpha})}{\vec{\beta}^2} = 0\,. \tag{3.292}$$

Adding (3.290) and (3.292) we arrive at the conclusion

$$\frac{\vec{\beta}^2}{\vec{\beta}^2} = 0\,, \tag{3.293}$$

which is a contradiction. Hence we cannot have more than 4 roots in a string of roots $\vec{\gamma} + k\,\vec{\alpha}$.

3.6 Root Systems for the Classical Algebras

3.6.1 The $SU(N)$ Algebras: A_n

An arbitrary $N\times N$ unitary matrix can be written as $U = e^{\mathrm{i}\,H}$, where H is hermitean. The unit-determinant condition $\det U = 1$ is equivalent to the tracelessness condition $\operatorname{tr} H = 0$.

Therefore the generators of $SU(N)$ are the set of all hermitean traceless $N \times N$ matrices, T_a. Let us choose a basis so that

$$\mathrm{tr}(T_a\, T_b) = \tfrac{1}{2}\delta_{ab}\,. \tag{3.294}$$

The maximal set of mutually commuting matrices amongst the T_a can most conveniently be taken to be the diagonal matrices, so these will form the Cartan subalgebra. Thus we can take

$$\begin{aligned}
H_1 &= \tfrac{1}{\sqrt{2}}\,\mathrm{diag}\,(1,-1,0,0,0,\ldots,0,0)\,,\\
H_2 &= \tfrac{1}{2\sqrt{3}}\,\mathrm{diag}\,(1,1,-2,0,0,\ldots,0,0)\,,\\
H_3 &= \tfrac{1}{2\sqrt{6}}\,\mathrm{diag}\,(1,1,1,-3,0,\ldots,0,0)\,,\\
&\vdots\\
H_j &= \tfrac{1}{\sqrt{2j(j+1)}}\mathrm{diag}\,(1,1,1,\ldots,1,-j,0,0,\ldots,0,0)\,,\\
&\vdots\\
H_{N-1} &= \tfrac{1}{\sqrt{2N(N-1)}}\,\mathrm{diag}\,(1,1,1,1,1,\ldots,1,1,-(N-1))\,,
\end{aligned} \tag{3.295}$$

where in the penultimate line H_j has j entries of 1 before the $-j$. These are normalised so that

$$\mathrm{tr}(H_i\, H_j) = \tfrac{1}{2}\delta_{ij}\,. \tag{3.296}$$

Since there are $(N-1)$ of them, we conclude that $SU(N)$ has rank $(N-1)$. In Cartan's classification, $SU(N)$ is denoted by A_{N-1}. The A indicates the special unitary sequence of algebras, and the subscript indicates the rank.

The $N \times N$ matrices generate the N-dimensional represenation of $SU(N)$, i.e. the *defining representation.* The matrices act by matrix multiplication on the N states of the vector space,

$$|\vec{\nu}_1\rangle = \begin{pmatrix}1\\0\\0\\\vdots\\0\\0\end{pmatrix}, \qquad |\vec{\nu}_2\rangle = \begin{pmatrix}0\\1\\0\\\vdots\\0\\0\end{pmatrix}, \qquad \cdots, \qquad |\vec{\nu}_N\rangle = \begin{pmatrix}0\\0\\0\\\vdots\\0\\1\end{pmatrix}. \tag{3.297}$$

Their weights, $\vec{\nu}_k$, i.e. their eigenvalues under H_j, are easily seen to be

$$\vec{\nu}_1 = (\tfrac{1}{2}, \tfrac{1}{2\sqrt{3}}, \tfrac{1}{2\sqrt{6}}, \ldots, \tfrac{1}{\sqrt{2j(j+1}}, \ldots, \tfrac{1}{\sqrt{2N(N-1)}})\,,$$

$$
\begin{aligned}
\vec{\nu}_2 &= (-\tfrac{1}{2}, \tfrac{1}{2\sqrt{3}}, \tfrac{1}{2\sqrt{6}}, \ldots, \tfrac{1}{\sqrt{2j(j+1}}, \ldots, \tfrac{1}{\sqrt{2N(N-1)}})\,, \\
\vec{\nu}_3 &= (0, -\tfrac{1}{\sqrt{3}}, \tfrac{1}{2\sqrt{6}}, \ldots, \tfrac{1}{\sqrt{2j(j+1}}, \ldots, \tfrac{1}{\sqrt{2N(N-1)}})\,, \\
&\;\vdots \\
\vec{\nu}_N &= (0, 0, 0, \ldots, 0, \ldots, -\tfrac{(N-1)}{\sqrt{2N(N-1)}})\,. \qquad (3.298)
\end{aligned}
$$

When we introduced the notion of positivity or negativity of weight vectors, we used the rule that the sign of the first non-zero component working from the left would determine the sign of the root. As we emphasised then, this is a completely arbitrary choice. In fact here, it is preferrable to work in from the right; i.e. , we shall say that a weight is positive (negative) if its first non-zero component working from the right is positive (negative). Under this scheme, we have

$$\vec{\nu}_1 > \vec{\nu}_2 > \vec{\nu}_3 > \cdots > \vec{\nu}_{N-1} > \vec{\nu}_N\,. \qquad (3.299)$$

The raising and lowering operators will be built from complex combinations of the off-diagonal hermitean matrices. Specifically, we may define the $N \times N$ matrix E_{ij}, which has zeros everywhere except at row i, column j, where the component is equal to $1/\sqrt{2}$. It is manifest that these act as raising and lowering operators on the set of fundamental states $|\vec{\nu}_k\rangle$, according to the rule

$$E_{ij}\,|\nu_k\rangle = \tfrac{1}{\sqrt{2}}\,\delta_{jk}\,|\vec{\nu}_i\rangle\,. \qquad (3.300)$$

Now the differences between the weights are necessarily roots, since in general $E_{\vec{\alpha}}\,|\vec{\mu}\rangle = c\,|\vec{\mu}+\vec{\alpha}\rangle$. Thus we know that roots are given by

$$\vec{\nu}_i - \vec{\nu}_j\,, \qquad \text{for any} \;\; i \ne j\,. \qquad (3.301)$$

In fact since we get $N(N-1)$ roots by this construction, we see that these constitute *all* the roots of $SU(N)$.[20] From (3.299) we see that the positive roots are given by

$$\vec{\nu}_i - \vec{\nu}_j\,, \qquad i < j\,. \qquad (3.302)$$

The simple roots are then clearly given by

$$\vec{\alpha}_i = \vec{\nu}_i - \vec{\nu}_{i+1}\,, \qquad 1 \le i \le N-1\,. \qquad (3.303)$$

[20] $SU(N)$ has dimension N^2-1, and it has rank $(N-1)$, so there are $(N^2-1)-(N-1) = N(N-1)$ roots.

Explicitly, they are given by

$$\begin{aligned}
\vec{\alpha}_1 &= (1, 0, 0, \ldots, 0, 0)\,, \\
\vec{\alpha}_2 &= (-\tfrac{1}{2}, \tfrac{\sqrt{3}}{2}, 0, \ldots, 0, 0)\,, \\
\vec{\alpha}_3 &= (0, -\tfrac{1}{\sqrt{3}}, \sqrt{\tfrac{2}{3}}, \ldots, 0, 0)\,, \\
&\vdots \\
\vec{\alpha}_j &= (0, 0, 0, \ldots, -\sqrt{\tfrac{j-1}{2j}}, \sqrt{\tfrac{j+1}{2j}}, 0, \ldots, 0, 0)\,, \\
&\vdots \\
\vec{\alpha}_{N-1} &= (0, 0, 0, \ldots, 0, -\sqrt{\tfrac{N-2}{2(N-1)}}, \sqrt{\tfrac{N}{2(N-1)}}\ )\,.
\end{aligned} \tag{3.304}$$

It is straightforward to check from the above that the simple roots of $SU(N)$ satisfy

$$\begin{aligned}
\vec{\alpha}_i \cdot \vec{\alpha}_i &= 1\,, \qquad \text{for each } i\,, \\
\vec{\alpha}_i \cdot \vec{\alpha}_{i+1} &= -\tfrac{1}{2}\,, \qquad \text{for each } i\,, \\
\vec{\alpha}_i \cdot \vec{\alpha}_j &= 0\,, \qquad i \ne j \text{ and } i \ne j \pm 1\,.
\end{aligned} \tag{3.305}$$

Note that we can summarise all these dot products in the single equation

$$\vec{\alpha}_i \cdot \vec{\alpha}_j = \delta_{i,j} - \tfrac{1}{2}\delta_{i,j+1} - \tfrac{1}{2}\delta_{i,j-1}\,. \tag{3.306}$$

From these, it follows that the angle between any pair of *adjacent* simple roots is 120^o, while the angle between any pair of non-adjacent simple roots is 90^o. All the simple roots have the same length. It therefore follows that the Dynkin diagram for $SU(n)$ is

$$\circ\!-\!\!-\!\circ\!-\!\!-\!\circ\!-\!\!-\!\circ\cdots\circ\!-\!\!-\!\circ\!-\!\!-\!\circ\,, \tag{3.307}$$

where there are $N-1$ circles. We can label them $\vec{\alpha}_1, \vec{\alpha}_2, \vec{\alpha}_3, \ldots, \vec{\alpha}_{N-2}, \vec{\alpha}_{N-1}$.

There is a more convenient way to parameterise the roots of $SU(N)$. Let us first make a shift by 1, and consider $SU(n+1)$, which has rank n and is called A_n in the Dynkin classification scheme. We then introduce a set of $(n+1)$ mutually-orthogonal unit vectors $\vec{e}_i$ in $\mathbb{R}^{n+1}$, for $1 \le i \le n+1$, satisfying

$$\vec{e}_i \cdot \vec{e}_j = \delta_{ij}\,. \tag{3.308}$$

We can choose a basis where

$$\vec{e}_i = (0, 0, 0, \ldots, 0, 1, 0, \ldots, 0, 0)\,, \tag{3.309}$$

where the only non-zero entry is the 1 in the i'th component. The root vectors of A_n then lie in the n-dimensional hyperplane orthogonal to

$$\vec{v} = \vec{e}_1 + \vec{e}_2 + \vec{e}_3 + \cdots + \vec{e}_n + \vec{e}_{n+1}\,. \tag{3.310}$$

They are given by

$$\vec{e}_i - \vec{e}_j\,, \tag{3.311}$$

and are positive if $i > j$, and negative if $i < j$.[21] The simple roots are clearly given by

$$\vec{\alpha}_i = \vec{e}_i - \vec{e}_{i+1}\,, \qquad 1 \le i \le n\,. \tag{3.312}$$

For example, we can write the non-simple positive root $\vec{e}_1 - \vec{e}_3$ as

$$\vec{e}_1 - \vec{e}_3 = (\vec{e}_1 - \vec{e}_2) + (\vec{e}_2 - \vec{e}_3) = \vec{\alpha}_1 + \vec{\alpha}_2\,. \tag{3.313}$$

From (3.312) we clearly have

$$\vec{\alpha}_i \cdot \vec{\alpha}_j = 2\delta_{i,j} - \delta_{i,j+1} - \delta_{i,j-1}\,. \tag{3.314}$$

Up to an overall normalisation factor (which is totally irrelevant as far as determing the structure of the algebra is concerned), this is equivalent to what we had in equation (3.306).

3.6.2 The $SO(N)$ Algebras: B_n and D_n

The $SO(N)$ algebra is generated by $N \times N$ matrices that are imaginary and antisymmetric. This can be seen by exponentiating to get $SO(N)$ group elements. Thus if A is antisymmetric, then $e^{\mathrm{i}\,A}$ is orthogonal:

$$(e^{\mathrm{i}\,A})^T\,(e^{\mathrm{i}\,A}) = (e^{\mathrm{i}\,A^T})\,(e^{\mathrm{i}\,A}) = (e^{-\mathrm{i}\,A})\,(e^{\mathrm{i}\,A}) = \mathbb{1}\,. \tag{3.315}$$

Our general rule is that we take our generators to be Hermitean, and so if they are antisymmetric, they must be imaginary.

Here, we must divide the discussion into two cases, depending on whether N is even or odd. First, let us consider the even case, $N = 2n$. The $SO(2n)$ algebras are called D_n in the Dynkin classification.

[21] We have now reverted to determining the sign of a vector by the sign of its first non-zero component starting from the left.

For $D_n = SO(2n)$, we consider the set of all imaginary antisymmetric $2n \times 2n$ matrices. We can take the Cartan generators, of which there are n, to be

$$H_1 = \begin{pmatrix} \sigma_2 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}, \qquad H_2 = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix},$$

$$H_3 = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}, \qquad \cdots \qquad H_n = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ 0 & 0 & 0 & \cdots & \sigma_2 \end{pmatrix},$$

where each entry represents a 2×2 matrix, and σ_2 is the second Pauli matrix,

$$\sigma_2 = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}. \tag{3.316}$$

Note that $\mathrm{tr}(H_i H_j) = 2\delta_{ij}$. Note also that $D_n = SO(2n)$ has rank n.

We can now consider the states of the $2n$-dimensional defining representation (corresponding to the $2n \times 2n$ matrices acting on the $2n$-dimensional vector space). We can write these states as

$$|1\rangle = \begin{pmatrix} 1 \\ \mathrm{i} \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \qquad |2\rangle = \begin{pmatrix} 1 \\ -\mathrm{i} \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \qquad |3\rangle = \begin{pmatrix} 0 \\ 0 \\ 1 \\ \mathrm{i} \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \qquad |4\rangle = \begin{pmatrix} 0 \\ 0 \\ 1 \\ -\mathrm{i} \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \tag{3.317}$$

and so on, up to

$$|2n-1\rangle = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \mathrm{i} \end{pmatrix}, \qquad |2n\rangle = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ -\mathrm{i} \end{pmatrix}. \tag{3.318}$$

These vectors have been chosen because they are eigenvectors under the Cartan generators $H)i$. In fact they clearly have weight vectors as follows:

$$\begin{aligned}
|1\rangle: &\qquad (1,0,0,\ldots,0,0) = \vec{e}_1\,, \\
|2\rangle: &\qquad (-1,0,0,\ldots,0,0) = -\vec{e}_1\,, \\
|3\rangle: &\qquad (0,1,0,\ldots,0,0) = \vec{e}_2\,, \\
|4\rangle: &\qquad (0,-1,0,\ldots,0,0) = -\vec{e}_2\,, \\
&\vdots \\
|2n-1\rangle: &\qquad (0,0,0,\ldots,0,1) = \vec{e}_n\,, \\
|2n\rangle: &\qquad (0,0,0,\ldots,0,-1) = -\vec{e}_n\,.
\end{aligned} \tag{3.319}$$

As in our $SU(N)$ discussion, we can now read off the root vectors of $SO(2n)$, since we know that the difference between any pair of weights in the defining representation must be a root. Thus the full set of roots is given by

$$\pm\vec{e}_i \pm \vec{e}_j\,, \qquad i \le j\,, \tag{3.320}$$

where the $\pm$ signs can be chosen independently. The positive roots are then

$$\vec{e}_i \pm \vec{e}_j\,, \qquad i < j\,, \tag{3.321}$$

and we can easily then see that the simple roots are given by

$$\begin{aligned}
\vec{\alpha}_i &= \vec{e}_i - \vec{e}_{i+1}\,, \qquad 1 \le i \le n-1\,, \\
\vec{\alpha}_n &= \vec{e}_{n-1} + \vec{e}_n\,.
\end{aligned} \tag{3.322}$$

In other words, one can easily check that using these, one can build all the positive roots in (3.321) by taking sums of the $\vec{\alpha}_i$ with non-negative integer coefficients.

It follows from (3.322) that all n simple roots have the same length:

$$\vec{\alpha}_1^2 = \vec{\alpha}_2^2 = \vec{\alpha}_3^2 = \cdots = \vec{\alpha}_n^2 = 2\,, \tag{3.323}$$

and that their dot products are given by

$$\begin{aligned}
&\vec{\alpha}_1\cdot\vec{\alpha}_2 = \vec{\alpha}_2\cdot\vec{\alpha}_3 = \vec{\alpha}_3\cdot\vec{\alpha}_4 = \cdots = \vec{\alpha}_{n-2}\cdot\vec{\alpha}_{n-1} = -1\,, \\
&\vec{\alpha}_{n-2}\cdot\vec{\alpha}_n = -1\,, \qquad \vec{\alpha}_{n-1}\cdot\vec{\alpha}_n = 0\,.
\end{aligned} \tag{3.324}$$

All other dot products not listed here are zero. The Dynkin diagram for $D_n = SO(n)$ is shown in Figure 10 below.[22]

$$
\begin{array}{ccccccccccc}
\vec{\alpha}_1 & & \alpha_2 & & & & \vec{\alpha}_{n-3} & & \vec{\alpha}_{n-2} & & \vec{\alpha}_{n-1} \\
\circ & — & \circ & — & \cdots\cdots & — & \circ & — & \circ & — & \circ \\
& & & & & & & & | & & \\
& & & & & & & & \circ & & \\
& & & & & & & & \vec{\alpha}_n & &
\end{array}
$$

Figure 10. The Dynkin diagram for $SO(2n)$, which is called D_n in the Dynkin classification. It has rank n, and it is simply-laced.

Now, let us consider $SO(2n+1)$, which is known as B_n in the Dynkin classification. Like $SO(2n)$, this has rank n. We can take the Cartan generators to be

$$
H_1 = \begin{pmatrix} \sigma_2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}, \qquad H_2 = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \sigma_2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix},
$$

$$
H_3 = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \sigma_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}, \qquad H_n = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_2 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}, \qquad (3.325)
$$

where again σ_2 is the Pauli matrix given in (3.316). Note that here, every entry in the matrix represents a 2×2 submatrix, except for the entries down the far right column, and along the bottom row, which are just numbers (i.e. 1×1 matrices). The reason for this is, of course, that the matrices here are $(2n+1)\times(2n+1)$ in dimension, and so there

[22]Note that it is customarily drawn with the right-hand end twisted anti-clockwise through 45^o, so that there are two "ears" formed by $\vec{\alpha}_{n-1}$ and $\vec{\alpha}_n$. This is entirely equivalent, since only the pattern of connecting lines and the type of circle (open or closed) has any significance. The reason for displaying it as in Figure 10 is simply because I don't know how to construct the necessary 45^o lines using Latex.

is a left-over strip around the right and the bottom, after we have filled in the rest with 2×2 blocks. This is also why we don't get an extra Cartan generator when we move from $SO(2n)$ to $SO(2n+1)$.

The states of the $(2n+1)$-dimensional defining representation will comprise the $(2n+1)$-component column vectors

$$|1\rangle = \begin{pmatrix} 1\\ \mathrm{i}\\ 0\\ 0\\ \vdots\\ 0\\ 0\\ 0 \end{pmatrix}, \qquad |2\rangle = \begin{pmatrix} 1\\ -\mathrm{i}\\ 0\\ 0\\ \vdots\\ 0\\ 0\\ 0 \end{pmatrix}, \qquad |3\rangle = \begin{pmatrix} 0\\ 0\\ 1\\ \mathrm{i}\\ \vdots\\ 0\\ 0\\ 0 \end{pmatrix}, \qquad |4\rangle = \begin{pmatrix} 0\\ 0\\ 1\\ -\mathrm{i}\\ \vdots\\ 0\\ 0\\ 0 \end{pmatrix}, \tag{3.326}$$

and so on, up to

$$|2n-1\rangle = \begin{pmatrix} 0\\ 0\\ 0\\ 0\\ \vdots\\ 1\\ \mathrm{i}\\ 0 \end{pmatrix}, \qquad |2n\rangle = \begin{pmatrix} 0\\ 0\\ 0\\ 0\\ \vdots\\ 1\\ -\mathrm{i}\\ 0 \end{pmatrix}, \qquad |2n+1\rangle = \begin{pmatrix} 0\\ 0\\ 0\\ 0\\ \vdots\\ 0\\ 0\\ 1 \end{pmatrix}. \tag{3.327}$$

Note that the first $2n$ of these are just like the $2n$-component state vectors of $SO(2n)$, except that they now have an extra 0 entry at the bottom. The weights of these states under the Cartan generators H_i can be read off by inspection:

$$\begin{aligned}
|1\rangle : &\qquad (1,0,0,\ldots,0,0) = \vec{e}_1\,,\\
|2\rangle : &\qquad (-1,0,0,\ldots,0,0) = -\vec{e}_1\,,\\
|3\rangle : &\qquad (0,1,0,\ldots,0,0) = \vec{e}_2\,,\\
|4\rangle : &\qquad (0,-1,0,\ldots,0,0) = -\vec{e}_2\,,\\
&\vdots\\
|2n-1\rangle : &\qquad (0,0,0,\ldots,0,1) = \vec{e}_n\,,\\
|2n\rangle : &\qquad (0,0,0,\ldots,0,-1) = -\vec{e}_n\,,
\end{aligned}$$

$$|2n+1\rangle: \qquad = (0,0,0,\ldots,0,0)\,, \tag{3.328}$$

As before, the raising and lowering operators in the $SO(2n+1)$ algebra will map amongst the states of the defining representation, and so we can read off the root vectors as the differences between their weight vectors. Thus we have that the roots are given by

$$\begin{aligned} & \pm\vec{e}_i \pm \vec{e}_j\,, \qquad i<j\,, \\ \text{and} \qquad & \pm\vec{e}_i\,. \end{aligned} \tag{3.329}$$

The positive roots are

$$\begin{aligned} & \vec{e}_i \pm \vec{e}_j\,, \qquad i<j\,, \\ \text{and} \qquad & \vec{e}_i\,, \end{aligned} \tag{3.330}$$

and so the simple roots are given by

$$\begin{aligned} \vec{\alpha}_i &= \vec{e}_i - \vec{e}_{i+1}\,, \qquad 1 \le i \le n-1\,, \\ \vec{\alpha}_n &= \vec{e}_n\,. \end{aligned} \tag{3.331}$$

From (3.331) we see that

$$\vec{\alpha}_1^2 = \vec{\alpha}_2^2 = \cdots = \vec{\alpha}_{n-2}^2 = \vec{\alpha}_{n-1}^2 = 2\,, \qquad \vec{\alpha}_n^2 = 1\,. \tag{3.332}$$

Thus $\vec{\alpha}_i$ for $1 \le i \le n-1$ are long roots, and $\vec{\alpha}_n$ is a short root. Unlike $A_n = SU(n+1)$, and $D_n = SO(2n)$, therefore, $B_n = SO(2n+1)$ is not simply-laced. The remaining non-vanishing dot products are

$$\vec{\alpha}_1 \cdot \vec{\alpha}_2 = \vec{\alpha}_2 \cdot \vec{\alpha}_3 = \cdots = \vec{\alpha}_{n-2} \cdot \vec{\alpha}_{n-1} = \vec{\alpha}_{n-1} \cdot \vec{\alpha}_n = -1\,. \tag{3.333}$$

From these results, it follows that each simple root makes an angle of 120^o with the adjacent one, except for $\vec{\alpha}_n$, which makes an angle of 135^o with $\vec{\alpha}_{n-1}$. The Dynkin diagram for $B_n = SO(2n+1)$ is shown in Figure 11 below.

$$\begin{array}{ccccccccccc} \vec{\alpha}_1 & & \alpha_2 & & & & \vec{\alpha}_{n-2} & & \vec{\alpha}_{n-1} & & \vec{\alpha}_n \\ \circ & — & \circ & — & \cdots\cdots & — & \circ & — & \circ & = & \bullet \end{array}$$

Figure 11. The Dynkin diagram for $SO(2n+1)$, which is called B_n in the Dynkin classification. It has rank n, and it is not simply-laced.

3.6.3 The $Sp(2n)$ Algebras: C_n

$Sp(2n)$ is generated by $2n \times 2n$ matrices X that satisfy

$$XG + GX^T = 0\,, \tag{3.334}$$

where $G = -G^T$ is some non-degenerate antisymmetric matrix. We can write G, and the generators X, as tensor products of 2×2 and $n \times n$ matrices. We take

$$G = \sigma_2 \otimes \mathbb{1}\,, \tag{3.335}$$

where $\mathbb{1}$ is the $n \times n$ unit matrix, and σ_2 is the second Pauli matrix, as given in (3.316). The tensor product can be understood as follows: one thinks of the $2n \times 2n$ matrix as being composed of four $n \times n$ blocks, with each block composed of the second matrix factor (the $n \times n$ matrix after the $\otimes$ sign) multiplied by the corresponding component of the 2×2 matrix. Thus

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \otimes A = \begin{pmatrix} aA & bA \\ cA & dA \end{pmatrix}. \tag{3.336}$$

In particular, we shall have

$$G = \begin{pmatrix} 0 & -\mathrm{i}\,\mathbb{1} \\ \mathrm{i}\,\mathbb{1} & 0 \end{pmatrix}. \tag{3.337}$$

However, one does not actually need to construct the $2n \times 2n$ matrices explicitly like this. One can perfectly well just manipulate the matrices in their tensor product forms. The rules for multiplication of matrices written in tensor-product form are simply

$$(A \otimes B)\,(C \otimes D) = (AC) \otimes (BD)\,. \tag{3.338}$$

The generators X are first of all, as always, Hermitean matrices, and in addition they must satisfy (3.334). With G given by (3.335), it follows that the set of all X can be obtained from the following sets of matrices:

$$\mathbb{1} \otimes A\,, \qquad \sigma_1 \otimes S_1\,, \qquad \sigma_2 \otimes S_2\,, \qquad \sigma_3 \otimes S_3\,. \tag{3.339}$$

Here A denotes arbitrary $n \times n$ imaginary antisymmetric matrices, S_1, S_2 and S_3 denote arbitrary $n \times n$ real symmetric matrices, and σ_i are the three Pauli matrices, as given in (3.80). Counting the total number of real generators X, we therefore get

$$\dim\, Sp(2n) = \tfrac{1}{2} n(n-1) + 3 \times \tfrac{1}{2} n(n+1) = n(2n+1)\,. \tag{3.340}$$

In the explicit $2n \times 2n$ format, as in (3.336), one has

$$\begin{pmatrix} A & 0 \\ 0 & A \end{pmatrix}, \qquad \begin{pmatrix} 0 & S_1 \\ S_1 & 0 \end{pmatrix}, \qquad \begin{pmatrix} 0 & -\mathrm{i}\,S_2 \\ \mathrm{i}\,S_2 & 0 \end{pmatrix}, \qquad \begin{pmatrix} S_3 & 0 \\ 0 & -S_3 \end{pmatrix}. \tag{3.341}$$

One can easily verify that all these matrices satisfy the defining relation (3.334).

The subset $1\!\!1\otimes A$ and $\sigma_3\otimes S_3$, with the additional condition that S_3 be traceless, generate an $SU(n)$ subalgebra of $Sp(2n)$, since we shall have

$$\begin{pmatrix} A+S_3 & 0 \\ 0 & A-S_3 \end{pmatrix} = \begin{pmatrix} T & 0 \\ 0 & -T^* \end{pmatrix}, \tag{3.342}$$

with $T = A + S_3$ being Hermitean and traceless. (Recall that A is imaginary and antisymmetric, whilst S_3 is real and symmetric.) It is convenient, therefore, to choose the $Sp(2n)$ Cartan subalgebra to include the Cartan subalgebra of $SU(n)$. We can therefore choose $(n-1)$ of the $Sp(2n)$ Cartan subalgebra matrices to be the given by taking matrices T in (3.342) that are just the diagonal traceless $SU(n)$ Cartan matrices given in (3.295). There is one more matrix in $Sp(2n)$ that commutes with these, namely

$$H_n = \frac{1}{\sqrt{2n}}\,\sigma_3\otimes 1\!\!1\,. \tag{3.343}$$

Thus we have in total n Cartan generators, so $Sp(2n)$ has rank n. It is known as C_n in the Cartan classification.

The full set of generators in the $SU(n)$ subalgebra comprise $Sp(2n)$ matrices that commute with H_n. Thus we can first enumerate the $Sp(2n)$ roots that lie in the $SU(n)$ subalgebra; they will simply be give by the differences of weights $\vec\nu_i$ of the defining representation of $SU(n)$, which were given in (3.298. These are $(n-1)$-component vectors (since $SU(n) = A_{n-1}$ has rank $(n-1)$), and so we can write the corresponding $Sp(2n)$ roots as

$$(\vec\nu_i - \vec\nu_j, 0)\,, \tag{3.344}$$

where we have appended a zero as the n'th component (since the $SU(n)$ matrices all have zero weight under H_n). The remaining $Sp(2n)$ generators that are not contained in $SU(n)$ consist of matrices which can be organised into raising and lowering operators of the form

$$(\sigma_1 \pm \mathrm{i}\,\sigma_2)\otimes S_{k\ell}\,, \tag{3.345}$$

where $S_{k\ell}$ is the symmetric $n\times n$ matrix with components $(S_{k\ell})_{ij}$ given by

$$(S_{k\ell})_{ij} = \delta_{ik}\,\delta_{j\ell} + \delta_{i\ell}\,\delta_{jk}\,. \tag{3.346}$$

They satisfy

$$[H_n, (\sigma_1 \pm \mathrm{i}\,\sigma_2)\otimes S_{k\ell}] = \pm\frac{2}{\sqrt{2n}}\,(\sigma_1 \pm \mathrm{i}\,\sigma_2)\otimes S_{k\ell}\,, \tag{3.347}$$

and

$$[H_i, (\sigma_1 \pm \mathrm{i}\,\sigma_2)\otimes S_{k\ell}] = \pm(\vec\nu_k + \vec\nu_\ell)_i\,(\sigma_1 \pm \mathrm{i}\,\sigma_2)\otimes S_{k\ell}\,, \qquad 1\le i\le n-1\,. \tag{3.348}$$

The full set of $Sp(2n)$ roots are therefore given by

$$\begin{aligned} &(\vec{\nu}_i - \vec{\nu}_j, 0)\,, \qquad i \neq j\,, \\ &\pm\Big(\vec{\nu}_i + \vec{\nu}_j, \sqrt{\frac{2}{n}}\Big)\,, \qquad \text{all}\quad i, j\,. \end{aligned} \tag{3.349}$$

The positive roots comprise the subset

$$\begin{aligned} &(\vec{\nu}_i - \vec{\nu}_j, 0)\,, \qquad i < j\,, \\ &\pm\Big(\vec{\nu}_i + \vec{\nu}_j, \sqrt{\frac{2}{n}}\Big)\,, \qquad \text{all}\quad i, j\,, \end{aligned} \tag{3.350}$$

and hence the simple roots are

$$\begin{aligned} \vec{\alpha}_i &= (\vec{\nu}_i - \vec{\nu}_{i+1}, 0)\,, \qquad 1 \le i \le n-1\,, \\ \vec{\alpha}_n &= \Big(2\vec{\nu}_n, \sqrt{\frac{2}{n}}\Big)\,. \end{aligned} \tag{3.351}$$

These therefore satisfy dot-product relations as follows. For $1 \le i \le n-1$, they are the same as for $SU(n)$, namely

$$\begin{aligned} \vec{\alpha}_i \cdot \vec{\alpha}_j &= 1\,, \qquad i = j\,, \\ \vec{\alpha}_i \cdot \vec{\alpha}_j &= -\tfrac{1}{2} \qquad i = j \pm 1\,, \\ \vec{\alpha}_i \cdot \vec{\alpha}_j &= 0\,, \qquad \text{otherwise}\,. \end{aligned} \tag{3.352}$$

The dot-product relations involving the n'th simple root are

$$\begin{aligned} &\vec{\alpha}_i \cdot \vec{\alpha}_n = 0\,, \qquad 1 \le i \le n-2\,, \\ &\vec{\alpha}_{n-1} \cdot \vec{\alpha}_n = -1\,, \\ &\vec{\alpha}_n \cdot \vec{\alpha}_n = 2\,. \end{aligned} \tag{3.353}$$

We see that the simple roots $\vec{\alpha}_i$ with $1 \le i \le n-1$ are all "short," having length 1, whilst $\vec{\alpha}_n$ is "long," with length $\sqrt{2}$. The Dynkin diagram for $Sp(2n) = C_n$ is given by

$$\begin{array}{ccccccccccc} \vec{\alpha}_1 & & \alpha_2 & & & & \vec{\alpha}_{n-2} & & \vec{\alpha}_{n-1} & & \vec{\alpha}_n \\ \bullet & — & \bullet & — & \cdots\cdots & — & \bullet & — & \bullet & = & \circ \end{array}$$

Figure 11. The Dynkin diagram for $Sp(2n)$, which is called C_n in the Dynkin classification. It has rank n, and it is not simply-laced.

Note that a simpler way to write the root vectors is by defining a set of n orthonomral unit vectors $\vec{e}_i$ in $\mathbb{R}^n$, satisfying $\vec{e}_i \cdot \vec{e}_j = \delta_{ij}$. The positive roots are given by

$$\vec{e}_i \pm \vec{e}_j \,, \qquad i < j \,, \qquad \text{and} \;\; 2\vec{e}_i \,. \tag{3.354}$$

The simple roots are given by

$$\vec{\alpha}_i = \vec{e}_i - \vec{e}_{i+1} \,, \qquad 1 \le i \le n-1 \,, \qquad \text{and} \;\; 2\vec{e}_n \,. \tag{3.355}$$

3.6.4 The Exceptional Lie Algebras

So far, we have examined in detail the so-called classical Lie algebras, which are the ones that are definined by the action matrices on a vector space. In the case of the orthogonal, unitary and symplectic groups, the matrices are required to preserve a metric on the vector space. We have seen how to analyse all the classical groups in terms of the Cartan decomposition, we have found their root systems, and hence we have constructed their Dynkin diagrams.

In this section, we discuss the remaining simple Lie algebras. It might come as a surprise that there exist any more, and indeed they were discovered much later. As it turns out there are precisely five more simple Lie algebras, in addition to the infinite sequences of the A_n, B_n, C_n and D_n algebras that we have already met. They are named G_2, F_4, E_6, E_7 and E_8, and they are known as the *exceptional Lie algebras.* The reason why they were discovered later is that they are not defined in terms of their action via matrix multiplication on vector spaces; i.e. they do not correspond to groups of metric-preserving matrices. Instead, we define them by directly constructing their root systems, which, as we have seen, are fully encoded in the Dynkin diagram.

Effectively, then, the idea is that we establish the necessary and sufficient condititions under which a Dynkin diagram is *valid.* All valid Dynkin diagrams define Lie algebras, and so by classifying all valid Dynkin diagrams, we classify all Lie algberas.

There is insufficient time in this lecture course to present the classification procedure in detail, so at this stage we shall just give the basic facts, accompanied with a brief summary of how the results are proved.

We begin with the following observations. The simple roots of any simple Lie algebra, of rank m, satisfy:

1. They are m linearly-independent m-vectors.

2. If $\vec{\alpha}$ and $\vec{\beta}$ are simple roots, then

$$\frac{2\vec{\alpha} \cdot \vec{\beta}}{\vec{\alpha} \cdot \vec{\alpha}} \tag{3.356}$$

is a non-positive integer.

3. The simple roots must be *indecomposable*, i.e. their Dynkin diagram nmust be *connected*. If the Dynkin diagram comprised two or more disconnected pieces, then the Lie algebra would not be simple.

Any connected Dynkin diagram describes a simple Lie algebra. A system of vectors that satisfies conditions 1, 2 and 3 above is called a Π system. Every Π system corresponds to a simple Lie algebra. Our task, therefore, is to classify all possible Π systems.

We can begin by just focusing on the *angles* between the simple roots $\vec{\alpha}_i$. Thus we define the unit vectors

$$\vec{u}_i = \frac{\vec{\alpha}_i}{|\vec{\alpha}_i|}\,. \qquad (3.357)$$

We saw earlier that the simple roots can only have angles 90^o, 120^o, 135^o or 150^o between them. Thus when $i \neq j$ we have

$$\vec{u}_i \cdot \vec{u}_j = -\sqrt{\frac{r}{4}}\,, \qquad 0 \le r \le 3\,, \qquad (3.358)$$

where r is an integer, whilst

$$\vec{u}_i \cdot \vec{u}_j = 1\,, \qquad \text{when} \quad i = j\,. \qquad (3.359)$$

Since the dot product of a non-vanishing vector with itself is strictly positive, we have

$$\Big(\sum_{i=1}^N \vec{u}_i\Big) \cdot \Big(\sum_{j=1}^N \vec{u}_j\Big) > 0\,, \qquad (3.360)$$

and hence

$$\sum_{i=1}^N \vec{u}_i \cdot \vec{u}_i + 2\sum_{i<j} \vec{u}_i \cdot \vec{u}_j > 0\,, \qquad (3.361)$$

and hence we have

Theorem 1:

$$N + 2\sum_{i<j} \vec{u}_i \cdot \vec{u}_j > 0\,. \qquad (3.362)$$

Suppose that $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_p$ are a set of p orthonormal vectors in the root space, where $p \le \text{rank}\,\mathcal{G}$, satisfying

$$\vec{v}_i \cdot \vec{v}_j = \delta_{ij}\,. \qquad (3.363)$$

Then for any unit vector $\vec{u}$ in the root space, $\vec{u}\cdot\vec{v}_i$ is the direction cosine $\cos(\vec{u}, \vec{v}_i)$ of $\vec{u}$ with respect to $\vec{v}_i$, and we have

Theorem 2:

$$\sum_{i=1}^{p} (\vec{u} \cdot \vec{v}_i)^2 = \sum_{i=1}^{p} \cos^2(\vec{u}, \vec{v}_i) \le 1 \,. \tag{3.364}$$

Equality implies $\vec{u}$ lies in the subspace spanned by the $\vec{v}_i$, i.e. that $\vec{u}$ is linearly dependent on $\vec{v}_i$. Inequality implies that $\vec{u}$ and $\vec{v}_i$ are all linearly independent.

Using Theorems 1 and 2, we can classify all Dynkin diagrams. To do this, we establish a number of intermediate results.

(**1**) A Dynkin diagram cannot have loops. For example, we cannot have three circles where each is joined by a line to each of the other circles, forming a triangular loop. Here is the proof:

If two roots $\vec{u}_i$ and $\vec{u}_j$ are connected, then by (3.358) they satisfy

$$2\vec{u}_i \cdot \vec{u}_j \le -1 \,. \tag{3.365}$$

If N roots are connected in a loop, we must therefore have at least N lines, so

$$2 \sum_{i<j} \vec{u}_i \cdot \vec{u}_j \le -N \,. \tag{3.366}$$

However, by Theorem 1 (equation (3.362)), we have

$$2 \sum_{i<j} \vec{u}_i \cdot \vec{u}_j > -N \,. \tag{3.367}$$

Equations (3.366) and (3.367) contradict each other, and hence the supposition that loops can exist must be false.

(**2**) A Dynkin diagram cannot have more than two double lines. For example, the following cannot occur:

$$\circ - \circ - \circ - \circ = \circ = \circ = \circ \,. \tag{3.368}$$

Numbering the roots $1, 2, 3, \ldots, 7$, starting from the left, we have

$$\begin{aligned} 2\vec{u}_1 \cdot \vec{u}_2 = 2\vec{u}_2 \cdot \vec{u}_3 = 2\vec{u}_3 \cdot \vec{u}_4 = -1 \,, \\ 2\vec{u}_4 \cdot \vec{u}_5 = 2\vec{u}_5 \cdot \vec{u}_6 = 2\vec{u}_6 \cdot \vec{u}_7 = -\sqrt{2} \,. \end{aligned} \tag{3.369}$$

Plugging into the left-hand side of equation (3.362, we get

$$N + 2 \sum_{i<j} \vec{u}_i \cdot \vec{u}_j = 7 - (1 + 1 + 1 + \sqrt{2} + \sqrt{2} + \sqrt{2}) \,, \tag{3.370}$$

which is negative. This contradicts Theorem 1, since (3.362) says that in a valid Dynkin diagram this quantity should be positive. Hence the supposition that the diagram above could exist is false. One can similarly prove that no Dynkin diagram with more than one double root can exist.

(**3**) A Dynkin diagram cannot have more than one triple line. For example, the following cannot occur:

$$\circ\!-\!\circ\!-\!\circ \equiv \circ \equiv \circ\,. \tag{3.371}$$

For this diagram, we shall have

$$N + 2\sum_{i<j} \vec{u}_i \cdot \vec{u}_j = 5 - (1 + 1 + \sqrt{3} + \sqrt{3})\,, \tag{3.372}$$

which is negative. This contradicts equation (3.362) of Theorem 1, and hence the diagram is not a valid Dynkin diagram. Similar arguments show that no diagram with more than one triple line is valid.

(**4**) If the lines joining any two $\vec{u}_i$'s in a Dynkin diagram are cut, the result is a sum of two disconnected Dynkin diagrams.

Cutting the lines amounts to removing some root vectors from the root space. The remaining ones generate a subalgebra.

(**5**) The maximum number of lines that can connect to any vertex in a Dynkin diagram is 3. The proof is as follows:

Let vertices $\vec{v}_1, \vec{v}_2, \vec{v}_3, \ldots$ be connected to the vertex $\vec{u}$. Since there can be no loops, we must have $\vec{v}_i \cdot \vec{v}_j = 0$ for all $i \neq j$, and so $\vec{v}_i \cdot \vec{v}_j = \delta_{ij}$. Let the number of lines joining the vertex $\vec{v}_i$ to the vertex $\vec{u}$ be n_i. We therefore have

$$\vec{u} \cdot \vec{v}_i = -\sqrt{\frac{n_i}{4}}\,, \tag{3.373}$$

with $n_i = 1$, 2 or 3. Hence we have

$$\sum_i (\vec{u} \cdot \vec{v}_i)^2 = \sum_i \frac{n_i}{4}\,. \tag{3.374}$$

Now $\vec{u}$ must be linearly independent of the $\vec{v}_i$, since this is one of the defining properties of a Π system. By Theorem 2, we must therefore have

$$\sum_i (\vec{u} \cdot \vec{v}_i)^2 < 1\,. \tag{3.375}$$

Comparing with (3.374) we therefore have

$$\sum_i n_i < 4\,, \tag{3.376}$$

and so the total number of lines joining any vertex must be less than 4.

An immediate consequence of this property is that there can only be one Dynkin diagram with a triple line, namely

$$\circ \equiv \circ\,. \tag{3.377}$$

Recall that we are *not* yet worrying about the lengths of the simple roots; our current arguments are all concerned just with the angles between the simple roots. We are not at this stage making any statement about the relative lengths of the roots. Thus the diagram (3.377) is not being claimed to be a true Dynkin diagram; it is what one would see if one would see if one were blind to whether circles were open or filled in. As we shall see later, the actual Dynkin diagram involving a triple line is like (3.377), except that one circle is opem, and the other is filled.

(**6**) Any set of vertices $\vec{u}_i$ in a Dynkin diagram that are joined by a simple chain (i.e. vertices joined by *single* lines) can be shrunk to a single vertex, and the resulting diagram will again be a valid Dynkin diagram.

Thus, for example, one could shrink

$$\circ = \circ - \circ - \circ - \circ - \cdots - \circ = \circ \tag{3.378}$$

to

$$\circ = \circ = \circ\,, \tag{3.379}$$

and *if* the upper diagram were valid, then the lower would be too. Of course the lower one in this example is *not* valid, since the middle vertex has four lines joining it, which we proved to be impossible. The power of this "shrinking theorem" is that it enables to see immediately that the upper diagram (3.378) is not a valid Dynkin diagram either.

Proof:

We have presented above various properties that valid Dynkin diagrams must have. Due to lack of time, we will not present all the properties. Suffice it to say that after some effort, one can eventually establish a complete set of properties of valid Dynkin diagrams.

By applying these considerations, one can then give an enumeration of all valid Dynkin diagrams, and hence of all simple Lie algebras. The upshot is that in addition to the four series that we have already met, namely

$$A_n = SU(n+1)\,, \qquad B_n = SO(2n+1)\,, \qquad D_n = SO(2n)\,, \qquad C_n = Sp(2n)\,, \quad (3.380)$$

there are exactly five additional isolated cases, denoted by

$$G_2\,, \qquad F_4\,, \qquad E_6\,, \qquad E_7\,, \qquad E_8 \qquad\qquad (3.381)$$

in the Dynkin classification. As always, the subscript denotes the rank of the algebra. Their Dynkin digrams are

$$\begin{array}{ll} G_2 & \circ \equiv \bullet \\ F_4 & \circ\!-\!\circ = \bullet\!-\!\bullet \end{array}$$

and then E_6 is given by

$$\begin{array}{ccccccc} \mathrm{o} & - & \mathrm{o} & - & \mathrm{o} & - & \mathrm{o} \\ & & & & | & & \\ & & & & \mathrm{o} & & \\ & & & & | & & \\ & & & & \mathrm{o} & & \end{array}$$

E_7 is given by

$$\begin{array}{ccccccccccc} \mathrm{o} & - & \mathrm{o} & - & \mathrm{o} & - & \mathrm{o} & - & \mathrm{o} & - & \mathrm{o} \\ & & & & & & | & & & & \\ & & & & & & \mathrm{o} & & & & \end{array}$$

and E_8 is given by

$$\begin{array}{ccccccccccccc} \mathrm{o} & - & \mathrm{o} & - & \mathrm{o} & - & \mathrm{o} & - & \mathrm{o} & - & \mathrm{o} & - & \mathrm{o} \\ & & & & | & & & & & & & & \\ & & & & \mathrm{o} & & & & & & & & \end{array}$$

The dimensions of these *exceptional Lie algebras* are

$$G_2 \qquad 14$$

F_4	52
E_6	78
E_7	133
E_8	248

The algebra G_2 arises in a number of contexts in physics and mathematics. It is, for example, associated with a symmetry of the algebra of the octonions. In many ways E_8 is the most interesting of all. It also arises in various contexts in mathematics and physics. For example, it plays a very important role in string theory.