Contents

1 MODULE I — TENSOR ALGEBRA AND CALCULUS ON MANIFOLDS

This module will first review a representation of vectors, long familiar to any postgraduate physics student, in terms of their **components in a particular basis**. An alternative description in terms of components in a so-called **cobasis**, or **dual** basis will be introduced and its meaning explored. We will, however, put a lot of emphasis on a vector as an object that can be discussed without explicit reference to a basis (or cobasis).

Our powerful geometric approach will then allow a conceptually simple generalisation of vectors to tensors. While everyone agrees that vectors are indispensable to the mathematical description of many physical quantities, the equally great importance of tensors is not always fully appreciated. For example, it is difficult to understand electromagnetism if one insists on regarding the electric and magnetic fields as just two vector fields connected by Maxwell equations, instead of the six non-zero components of the rank-2 Faraday tensor[†] $\mathbf{F}$. Mathematicians (and many mathematical physicists) will argue that an even deeper understanding is achieved by considering $\mathbf{F}$ as a **p-form** (with $p = 2$). Thus, we will be led to introduce the language of p-forms. The need to describe how vectors and tensors change in time and space will lead us to introduce the idea of a manifold, We will also learn how p-forms can be differentiated and integrated.

1.1 Vector Spaces and Linear Mappings

1.1.1 Vector spaces in a nutshell

Definition 1.1. A **vector space** $\mathcal{V}$ over a field $\mathbb{F}$ is a (possibly infinite) set of objects on which an operation called addition and another called s-multiplication (multiplication by a scalar) are defined, and which is is **closed** under these operations. That is, any two elements $\mathbf{u}$ and $\mathbf{v}$ of $\mathcal{V}$ satisfy:

$$(a+b)(\mathbf{u}+\mathbf{v}) \;=\; (a\,\mathbf{u} \,+\, a\,\mathbf{v} \,+\, b\,\mathbf{u} \,+\, b\,\mathbf{v}) \;\in\; \mathcal{V}$$

$\forall\, a,\ b \in \mathbb{F}$; in what follows, $\mathbb{F} = \mathbb{R}$. This addition operation is commutative and associative; it can be composed with an inverse, the result being the zero element. As for s-multiplication, it obeys: $(ab)\,\mathbf{u} = a\,(b\,\mathbf{u})$. We will call (informally) elements of a vector space **vectors**, keepng in mind that deciding whether an object is a vector can only be done by reference to the vector space of which it is an element. In this view, the properties of a vector space matter more than any of its elements.

Example 1.1. $\mathbb{R}^n$, the set of all ordered n-tuples of real numbers, with addition defined as adding entries with the same place in the n-tuple, and s-multiplication by λ defined as multiplying each entry by λ, is perhaps the best-known and most important vector space.

Let $\mathcal{V}$ and $\mathcal{W}$ be two vector spaces that share the same field of scalars. We shall be interested in the set of all **linear mappings**, $\mathrm{Hom}\,(\mathcal{V},\mathcal{W}) \coloneqq \{\mathbf{T} \,:\, \mathcal{V} \,\to\, \mathcal{W}\}$, such that, $\forall\, \mathbf{T}_i \in \mathrm{Hom}\,(\mathcal{V},\mathcal{W})$:

$$(a\,\mathbf{T}_i + \mathbf{T}_j)(\mathcal{V}) \;=\; a\,\mathbf{T}_i(\mathcal{V}) \,+\, \mathbf{T}_j(\mathcal{V})$$

where the operations on the left are defined in $\mathrm{Hom}\,(\mathcal{V},\mathcal{W})$, and those on the right on $\mathcal{W}$. Henceforth, we will use the notation $\mathcal{L}(\mathcal{V},\,\mathcal{W})$ instead of Hom for our *space* of linear mappings. Of course, one can define linear mappings on $\mathcal{L}\,(\mathcal{V},\mathcal{W})$, ie., we can **compose** linear mappings, and the composition of two linear maps is itself a linear map.

1.2 Where Do Vectors Live?

From what we emphasised above, the answer has to be: in a vector space! Therefore, we should learn how to identify (or construct) such vector spaces. In what follows, we shall see that such a familiar space as spacetime cannot be endowed with a *vector*-space structure.

[†]These notes generally follow the conventions set by the ISO (International Standards Organisation) for mathematical typography, with one important exception: as in BF, vectors and tensors are in bold upright ($\mathbf{u}$) instead of bold italic font ($\boldsymbol{u}$). Sans-serif fonts denote matrices, eg. M.

1.2.1 Manifolds and coordinates

Definition 1.2. Let M be a set of elements, or "points", with a so-called topology that introduces the notion of **open balls** (or **neighbourhoods**) around points, in turn allowing a definition of continuity. If

- M can be entirely covered by a union U of possibly overlapping open (without boundary) subsets U_1, U_2, etc., (as given by the topology), each mapped in a one-to-one way to an open subset of $\mathbb{R}^n$ by a **coordinate map**: $x : U_i \longrightarrow \mathbb{R}^n$, $y : U_j \longrightarrow \mathbb{R}^n$, etc.;
- each coordinate map is differentiable, in the sense that a neighbourhood *around* any point $\mathcal{P}$ in M, which we know must exist by virtue of the topology, is mapped to a neighbourhood of the image via x of $\mathcal{P}$ in $\mathbb{R}^n$;
- when there is an overlapping region $U_i \cap U_j$ in M, only coordinate maps for which the transition map (aka coordinate transformation) $y \circ x^{-1} : \mathbb{R}^n \longrightarrow U_i \cap U_j \longrightarrow \mathbb{R}^n$ between them is (once) differentiable are allowed;

then we say that M is a **differentiable manifold**, usually abbreviated to just **manifold**. The minimum number n of parameters—each a map $x^i : U \longrightarrow \mathbb{R}$ $(i = 1, \ldots, n)$—that uniquely specify every point in an open subset is the **dimension** of the manifold. Real differentiable (C^1) manifolds can be shown to be **smooth**, ie., C^∞.

Each open subset (itself a manifold) together with its coordinate map is called a **coordinate chart**, $(U_i,\ x)$, or **local coordinate system**, on M. Any collection of coordinate charts that covers the whole of M is called an **atlas**.

Example 1.2.

- $\mathbb{R}^n$ can be promoted to a manifold; it is mapped into iself and can be covered with just one coordinate chart, Cartesian coordinates, or **standard (natural) coordinates**. Other charts are of course possible, eg. polar coordinates on $\mathbb{R}^2$, but in general more than one (see below) is needed to cover the manifold.

- A conical surface, even a semi-infinite one, can never be a manifold because of its tip.

- A vector space $\mathcal{V}$ can be made into a manifold that can be covered with one chart $(\mathcal{V}, \Phi)$, where Φ maps elements of $\mathcal{V}$ to their components in $\mathbb{R}^n$ in that basis. Conversely, however, a manifold is *not* in general a vector space! Considering the manifold of points on Earth's surface, there is no meaning to adding the position of Toronto to that of London.

- Even though $\mathbb{R}^n$ can be endowed with a manifold structure, a unit ball in $\mathbb{R}^n$, defined in Cartesian coordinates by $\sum x_i^2 \leq 1$, is not a manifold because it has an edge on which it is not differentiable. The *open* unit ball, $\sum x_i^2 < 1$, is a manifold. So is the **unit sphere**, S^n, defined by $\displaystyle\sum^{n} x_i^2 = 1$ and embedded in $\mathbb{R}^{n+1}$.

  Let us look at S^1, the unit circle in the plane $\mathbb{R}^2$, and S^2, the 2-dim sphere in $\mathbb{R}^3$. These are the archetypal examples of (closed) curves in $\mathbb{R}^2$ and (closed) surfaces in $\mathbb{R}^3$.

  S^1 being a 1-dim manifold, we wish to build an atlas for it. One way of doing this is with two *open* patches, $y = \pm\sqrt{1 - x^2}$ with the two points at $x = \pm 1$ excluded (why?), and the $+/-$ sign corresponding to the submanifold in the upper/lower half-plane. Then each point of any of the two submanifolds is in one-to-one correspondence with some $x \in \mathbb{R}$, with $x < 1$. To cover all of S^1, we can then repeat the procedure with two submanifolds in correspondence with $x > 0$ and $x < 0$, and an atlas with four charts has been constructed.

  S^1 also has another local coordinate, θ, related to x by the multi-valued coordinate transformation: $\theta = \tan^{-1}(y/x) = \tan^{-1}(\sqrt{1/x^2 - 1})$. To avoid a given point being in correspondence with more than one value of θ, the interval mapped to in $\mathbb{R}$ must be $[0,\ 2\pi)$.

  Similarly, an atlas can be constructed for S^2 out of patches corresponding to, first, $z > 0$ and $z < 0$ (leaving out the circle $z = 0$), then $y > 0$ and $y < 0$ (leaving out the points corresponding to $(y = 0, z = 0)$,

and finally $x > 0$ and $x < 0$. Each point in each patch can be mapped unambiguously to $\mathbb{R}^2$, and there exist functions $f(u, v)$ which map (x, y), (x, z) or (y, z) into coordinates u and $\pm\sqrt{1-u^2-v^2}$, both differentiable functions provided that $u^2 + v^2 < 1$, which is always the case in any given patch.

On S^2 we could also use the local spherical coordinates $\theta = \cos^{-1} z$ and $\phi = \tan^{-1}(y/x)$. The region of $\mathbb{R}^2$ mapped to is $(0,\ \pi)$ and $[0,\ 2\pi]$, with the poles removed. This time, the problem comes from a failure of 1-to-1 correspondence at $\theta = 0$ and $\theta = \pi$: for each of these the value of ϕ is undetermined. More patches are needed to cover S^2.

Notice that we have looked at S^1 and S^2 as being embedded, or immersed, in a higher-dimensional manifold, $\mathbb{R}^2$ and $\mathbb{R}^3$. Whitney's embedding theorems guarantee that any smooth M^n is a smooth submanifold of $\mathbb{R}^{m>2n}$; there are stronger results in important restricted cases. Embedding curves and surfaces in, eg., $\mathbb{R}^3$ is great for visualisation purposes, but we are really interested in the *intrinsic* properties of a manifold which should be independent of whatever higher-dimensional manifold in which it may or may not be immersed.

Beyond the technicalities, it is enough almost all the time to view a manifold as a set which can be parametrised in a smooth way.

1.2.2 Curves, directional derivatives and vectors

The naïve notion of a vector as a straight arrow from one point to another in $\mathbb{R}^n$ cannot be extended to arbitrary manifolds M, on which straightness will in general have no well-defined meaning (think of straight arrows *on* a sphere). As noted before, manifolds are not vector spaces; so where do vectors that we want to attach to a point in M actually live? And is it possible to think of a vector as a local object that involves only *that* point, in a way that is independent of any coordinate chart?

Definition 1.3. A **curve** $\Gamma \in M$ on a manifold M is a *mapping*, at least C^1 (no kinks!), that sends each value of a real parameter λ to a unique point $\mathcal{P}$ in M. More succintly, $\Gamma : \mathbb{R} \longrightarrow M$. In effect, λ is a coordinate on Γ, but note the direction of the mapping that, following tradition, is opposite the one used in definition 1.2. The curve at $\mathcal{P}$ is given by $\Gamma(\lambda_0) = \mathcal{P}$. We can also write $\Gamma(\lambda) = \Gamma \circ \lambda$.

Definition 1.4. Now introduce the *vector* space, $C^\infty(M) \coloneqq \{\, f : M \longrightarrow \mathbb{R} \,\}$, of all smooth, real-valued functions f that map a point in M, such that:

$$(f+g)(\mathcal{P}) = f(\mathcal{P}) + g(\mathcal{P}) \qquad \forall\, f,\, g \in C^\infty(M) \qquad\qquad (a\,f)(\mathcal{P}) = a\,f(\mathcal{P}) \qquad a \in \mathbb{R} \tag{1.1}$$

where the addition and s-multiplication on the left-hand side of the equations are operations on $C^\infty(M)$, while those on the right-hand side are on $\mathbb{R}$. We also have the composition: $f \circ \Gamma : \mathbb{R} \xrightarrow{\Gamma} M \xrightarrow{f} \mathbb{R}$. Thus, $f \circ \Gamma$ could be called $f(\lambda)$, with $\lambda \in \mathbb{R}$.

Definition 1.5. The **velocity** for a curve Γ at $\mathcal{P}$ is the linear map $\mathbf{v}_{(\Gamma,\mathcal{P})} : C^\infty(M) \longrightarrow \mathbb{R}$, defined as:

$$\mathbf{v}_{(\Gamma,\mathcal{P})}(f) \coloneqq \mathrm{d}_\lambda(f \circ \Gamma)\big|_{\lambda_0} \qquad \left(= \mathrm{d}_\lambda f\big|_{\lambda_0}\right) \tag{1.2}$$

where we use the notation d_λ as short-hand for $\mathrm{d}/\mathrm{d}\lambda$.

It should be clear that such a curve is only one of an infinite number containing $\mathcal{P}$ that have their own velocity at $\mathcal{P}$. Indeed, we could parametrise another curve Θ and write, following eq. (1.5): $\mathbf{w}(f) = \mathrm{d}_\lambda f$ at λ_1, where $\Theta(\lambda_1) = \mathcal{P}$. We say that the velocities are **tangent** to the manifold at $\mathcal{P}$.

1.2.3 The tangent space of a manifold

Definition 1.6. The **tangent space** $\mathcal{T}_\mathcal{P}$ to a manifold M^n — *even though* M^n *itself is not a vector space!* — at a point $\mathcal{P} \in M^n$, is a set that can be equipped with a vector-space structure, consisting of all the velocity vectors $\mathbf{v}_\mathcal{P}$ tangent to M^n at $\mathcal{P}$, In fact, *all* vectors defined on M^n at $\mathcal{P}$ live in $\mathcal{T}_\mathcal{P}$, not in M^n. $\mathcal{T}_\mathcal{P}$ is always of *finite* dimension. If M^n is viewed as embedded in $\mathbb{R}^N$ ($N \geq n$), $\mathcal{T}_\mathcal{P} = \mathbb{R}^n$.

The set of all points in a manifold, together with their tangent spaces, is called the **tangent bundle** of the manifold.

This definition rests on a very bold assertion, namely, that the velocities as defined above (definition 1.5) are in fact vectors. It is instructive to prove it, that is, to show that $\mathcal{T}_\mathcal{P}$ can indeed be made into a vector space. First, let us specify what is meant by addition and s-multiplication on $\mathcal{T}_\mathcal{P}$.

Definition 1.7. The addition operation on $\mathcal{T}_\mathcal{P}$ is a map, $\mathcal{T}_\mathcal{P} + \mathcal{T}_\mathcal{P} \longrightarrow \mathcal{L}(C^\infty(M), \mathbb{R})$, such that, $\forall\, f \in C^\infty(M)$ and any two curves $(\Gamma, \Theta) \in M$ intersecting at $\mathcal{P} \in M$:

$$(\mathbf{v}_{(\Gamma,\mathcal{P})} + \mathbf{v}_{(\Theta,\mathcal{P})})(f) \coloneqq \mathbf{v}_{(\Gamma,\mathcal{P})}(f) + \mathbf{v}_{(\Theta,\mathcal{P})}(f)$$

Again, the addition operation on the left is between mappings, whereas that on the right is on $\mathbb{R}$. As for s-multiplication, it is a map, $\mathbb{R} \times \mathcal{T}_\mathcal{P} \longrightarrow \mathcal{L}(C^\infty(M), \mathbb{R})$, such that, $\forall\, a \in \mathbb{R}$:

$$(a \cdot \mathbf{v}_{(\Gamma,\mathcal{P})})(f) \coloneqq a\, \mathbf{v}_{(\Gamma,\mathcal{P})}(f)$$

The question now is: do these operations close? In other words, can we find some curve $\Theta \in M$ such that: $a \cdot \mathbf{v}_{(\Gamma,\mathcal{P})} = \mathbf{v}_{(\Theta,\mathcal{P})}$, and perhaps another curve $\Sigma \in M$ such that: $\mathbf{v}_{(\Gamma,\mathcal{P})} + \mathbf{v}_{(\Theta,\mathcal{P})} = \mathbf{v}_{(\Sigma,\mathcal{P})}$?

To construct such a curve for s-multiplication, we first redefine the parameter of the curve Γ as the linear function, $\mu : \mathbb{R} \longrightarrow \mathbb{R}$, of λ: $\mu = a\lambda + \lambda_0$, with λ now the parameter of a curve Θ such that $\Theta(\lambda) = \Gamma(\mu)$ Therefore, $\Theta(0) = \Gamma(\lambda_0) = \mathcal{P}$. As in definition 1.3 we can write: $\Gamma(\mu) = (\Gamma \circ \mu(\lambda))$. Insert this information into the expression for the velocity for Θ at $\mathcal{P}$:

$$\begin{aligned}
\mathbf{v}_{(\Theta,\,\mathcal{P})}(f) &= \mathrm{d}_\lambda(f \circ \Theta)\big|_{\lambda=0} = \mathrm{d}_\lambda(f \circ \Gamma \circ \mu)\big|_{\lambda=0} \\
&= \mathrm{d}_\mu(f \circ \Gamma \circ \mu)\big|_{\mu(\lambda=0)=\lambda_0}\, \mathrm{d}_\lambda\mu\big|_{\lambda=0} \\
&= a\, \mathbf{v}_{(\Gamma,\,\mathcal{P})}(f)
\end{aligned}$$

Therefore, we have found a curve Θ such that the operation $a \cdot \mathbf{v}_{(\Gamma,\mathcal{P})}$ gives the velocity for that curve at $\mathcal{P}$.

Up to now, in our discussion of tangent spaces, we have been able to avoid any reference to coordinate charts. Unfortunately, when it comes to proving that addition of two velocities in $\mathcal{T}_\mathcal{P}$ gives a velocity in $\mathcal{T}_\mathcal{P}$, we cannot add the curve mappings directly since this has no meaning. Instead, assume that both curves Γ et Θ are in some open subset $U \subset M$ parametrised by coordinate functions denoted collectively by $x : U \longrightarrow \mathbb{R}^n$. Along with their intrinsic parameter λ, the curves themselves are automatically also parametrised by x, which describe what the curves "look like" in U.

We write the parametrisation of the curves in U as binary maps from $\mathbb{R}$ to $\mathbb{R}^n$: $(x \circ \Gamma)(\lambda)$ and $(x \circ \Theta)(\lambda)$ (or, if one wishes, $x(\lambda)$). Let Γ and Θ go through point $\mathcal{P}$ at values λ_1 and λ_2 ot their respective parameter. Then construct a curve Σ parametrised in U by:

$$(x \circ \Sigma)(\lambda) = (x \circ \Gamma)(\lambda_1 + \lambda) + (x \circ \Theta)(\lambda_2 + \lambda) - (x \circ \Gamma)(\lambda_1)$$

Although there might appear to be an obvious cancellation in this expression, it is not allowed because the coordinate functions are not linear and thus do not distribute over the additions in $\mathbb{R}^n$ in the arguments on the right. At $\lambda = 0$, however, the cancellation does occur, leaving $\Sigma_x(0) = \Theta(\lambda_2) = \mathcal{P}$, so that our curve Σ_x runs through point $\mathcal{P}$ at $\lambda = 0$.

We also need the derivative of the ν^{th} x coordinate of the curve Σ, evaluated at $\mathcal{P}$:

$$\begin{aligned} \mathrm{d}_\lambda(x^\nu \circ \Sigma)\big|_0 &= \mathrm{d}_\lambda\big[(x^\nu \circ \Gamma)(\lambda_1+\lambda) + (x^\nu \circ \Theta)(\lambda_2+\lambda) - (x^\nu \circ \Gamma)(\lambda_1)\big]\big|_0 \\ &= \mathrm{d}_{\lambda_1+\lambda}(x^\nu \circ \Gamma)\big|_{\lambda_1}\, \mathrm{d}_\lambda(\lambda_1+\lambda)\big|_0 + \mathrm{d}_{\lambda_2+\lambda}(x^\nu \circ \Theta)\big|_{\lambda_2}\, \mathrm{d}_\lambda(\lambda_2+\lambda)\big|_0 \\ &= \mathrm{d}_{\lambda_1+\lambda}(x^\nu \circ \Gamma)\big|_{\lambda_1} + \mathrm{d}_{\lambda_2+\lambda}(x^\nu \circ \Theta)\big|_{\lambda_2} \end{aligned} \tag{1.3}$$

We are now ready to compute the velocity for curve Σ at $\mathcal{P}$, starting from definition 1.5 in which we insert the identity map:

$$\begin{aligned} \mathbf{v}_{(\Sigma,\mathcal{P})}(f) &\coloneqq \mathrm{d}_\lambda(f \circ \Sigma)\big|_0 = \mathrm{d}_\lambda\big[(f \circ x^{-1}) \circ (x \circ \Sigma)\big]\big|_0 \\ &= \sum_\nu^n \Big[\, \mathrm{d}_\lambda(x^\nu \circ \Sigma)\big|_0 \, \big[\partial_\nu (f \circ x^{-1})\big]\big|_{(x^\nu \circ \Sigma)(0) = x^\nu(\mathcal{P})} \Big] \end{aligned} \tag{1.4}$$

where the index ν in the chain rule runs over the number of parameters (or local coordinates) that specify each point in U. The first factor on the right has been evaluated in eq. (1.3) and, running the last chain of equalities backward, there comes:

$$\begin{aligned} \mathbf{v}_{(\Sigma,\mathcal{P})}(f) &= \sum_\nu \Big[\big[\partial_\nu (f \circ x^{-1})\big]\big|_{x^\nu(\mathcal{P})}\, \mathrm{d}_\lambda(x^\nu \circ \Gamma)\big|_{\lambda_1}\Big] + \sum_\nu \Big[\big[\partial_\nu (f \circ x^{-1})\big]\big|_{x^\nu(\mathcal{P})}\, \mathrm{d}_\lambda(x^\nu \circ \Theta)\big|_{\lambda_2}\Big] \\ &= \mathrm{d}_\lambda\big[(f \circ x^{-1}) \circ (x \circ \Gamma)\big]\big|_{\lambda_1} + \mathrm{d}_\lambda\big[(f \circ x^{-1}) \circ (x \circ \Theta)\big]\big|_{\lambda_2} \\ &= \mathbf{v}_{(\Gamma,\mathcal{P})}(f) + \mathbf{v}_{(\Theta,\mathcal{P})}(f) \end{aligned}$$

Thus, adding the velocities for two curves meeting at some point yields the velocity for some other curve intersecting the others at that same point, and the tangent space of a curve at a point can indeed support a vector space structure! Do note that the result does not depend on whatever coordinate chart we might have used in the intermediate steps of the proof.

The coordinate-dependent expression that we derived in eq. (1.4) contains unexpected information. Taking a curve $\Gamma \subset U$ and a chart (U, x), with $\Gamma(\lambda = 0) = \mathcal{P}$. Our expression is:

$$\mathbf{v}_{(\Gamma,\mathcal{P})}(f) = \sum_\nu \Big[\, \mathrm{d}_\lambda(x^\nu \circ \Gamma)\big|_0 \, \big[\partial_\nu (f \circ x^{-1})\big]\big|_{(x^\nu \circ \Gamma)(0) = x^\nu(\mathcal{P})} \Big]$$

We have already remarked that we could write $\mathrm{d}_\lambda(x^\nu \circ \Gamma)\big|_0 = \mathrm{d}_\lambda x^\nu\big|_0$. And we may suspect that the second factor has something to do with the partial derivative of f. Indeed, f acts on U, not $\mathbb{R}$, without any reference to a local coordinate system on U. In order to have a calculable derivative, we take a detour via $\mathbb{R}^n$: $U \underset{x}{\longrightarrow} \mathbb{R}^n \underset{f\circ x^{-1}}{\longrightarrow} \mathbb{R}$. Since $f \circ x^{-1}$ maps $\mathbb{R}^n$ to $\mathbb{R}$, its usual derivatives can be calculated and behave for practical purposes like the standard $\partial_{x^\nu} f(x^\mu)$., that is: $\partial_{x^\nu} f\big|_{\mathcal{P}} \coloneqq \partial_\nu (f \circ x^{-1})\big|_{\mathcal{P}}$. Although we don't always write the dependence of f on x explicitly, here it is essential.

Having made this technical point, we discard the arbitrary function f and write the velocity vector as:

$$\mathbf{v}_{(\Gamma,\mathcal{P})} = \sum_\nu \mathrm{d}_\lambda x^\nu\big|_0 \, (\boldsymbol{\partial}_\nu)_{\mathcal{P}} \tag{1.5}$$

1.2.4 Vectors: components and bases

To interpret eq. (1.5), we need first to introduce a different perspective on vectors, one that does make reference to local coordinate systems.

Definition 1.8. If any $\mathbf{v} \in \mathcal{V}$ can be written as a linear combination[†]:

$$\mathbf{v} \;=\; \sum_{\alpha}^{n<\infty} v^\alpha\, \mathbf{e}_\alpha \;\equiv\; v^\alpha\, \mathbf{e}_\alpha \qquad \text{Einstein's convention : summation over repeated indices implied!} \tag{1.6}$$

of a set $\{\mathbf{e}_\alpha \in \mathcal{V}\}$, then that set is said to **span**, or to be a **set of generators** of, $\mathcal{V}$.

If, furthermore, this set is **linearly independent**, in the sense that demanding that $\mathbf{v} = 0$ forces all coefficients v^α in eq. (1.6) to vanish, then it is a **basis** of $\mathcal{V}$. The number n of vectors in the largest linearly independent set defines the dimension of $\mathcal{V}$, and we often write $\mathcal{V}^n$. Conversely, the number of elements of every basis of $\mathcal{V}^n$ is the dimension of $\mathcal{V}^n$.

The (real, and unique!) coefficients v^α in eq. (1.6) are called the **components** of the vector $\mathbf{v}$ in this basis. This one-to-one *correspondence* between $\mathcal{V}^n$ and $\mathbb{V}^n$ can be represented by a $n \times 1$ matrix:

$$\mathbf{v} \;\longmapsto\; \begin{pmatrix} v^1 \\ v^2 \\ \vdots \\ v^n \end{pmatrix}$$

Warning! $\mathbf{v}$ *and its components are different beasts and should never be confused*. Byron and Fuller (BF) do not make this distinction clear enough. Also, always remember that the index on $\mathbf{e}_\alpha$ identifies the *vector*, not a component of the vector.

Example 1.3. The **standard**, or natural, basis $\mathbb{R}^n$ is the set $\{\mathbf{e}_\alpha\}$ ($\alpha = 1, 2, \ldots, n$), where each n-tuple labelled by a value of α has 1 in the α^{th} position and 0 in all other positions.

Now, going back to eq. (1.5), and acting on a function f at point $\mathcal{P}$: $\mathbf{v}_{(\Gamma,\mathcal{P})}(f) \;= \mathrm{d}_\lambda x^\nu\big|_0\, \boldsymbol{\partial}_\nu(f)\big|_{\mathcal{P}}$, we now notice that it looks like the **directional derivative** of f in the direction of $\mathbf{v}$, which in basic calculus is written $\partial_{\mathbf{v}} f \coloneqq \mathbf{v}\cdot\boldsymbol{\nabla} f$. *This motivates us to* identify *any tangent vector* $\mathbf{t} \in \mathcal{T}_{\mathcal{P}}$ *with the directional derivative at* $\mathcal{P}$ *in the direction of* $\mathbf{v}$. Thus:

Definition 1.9. Given an arbitrary differentiable function f on a manifold M^n, parametrised by a local coordinate system by $f(x^1 \ldots, x^n)$, then the action of a vector $\mathbf{v}$ on f at a point $\mathcal{P}$ is defined as:

$$\mathbf{v}(f)\Big|_{\mathcal{P}} \;\coloneqq\; \partial_{\mathbf{v}} f \;=\; v^\nu\, \boldsymbol{\partial}_\nu f\,\Big|_{\mathcal{P}} \tag{1.7}$$

where the components $v^\nu = \mathbf{v}(x^\nu) = \mathrm{d}_\lambda x^\nu\big|_0$.

It is natural to think of $\boldsymbol{\partial}_\nu$ in eq. (1.5) as basis vectors for the tangent space of M at $\mathcal{P}$. But are they linearly independent? Take $f = x^\nu$, the coordinate functions for U; then $a^\mu \boldsymbol{\partial}_{x^\mu} x^\nu\big|_{\mathcal{P}} = a^\mu \partial_\mu (x^\nu \circ x^{-1})\big|_{\mathcal{P}} = a^\mu \delta_\mu{}^\nu = a^\nu$, where $a^\nu \in \mathbb{R}$. If $a^\mu \boldsymbol{\partial}_{x^\mu} x^\nu\big|_{\mathcal{P}} = 0$, $a^\nu = 0$, which shows that the $\boldsymbol{\partial}_\nu$ do form a basis of the tangent space. Thus:

Definition 1.10. The tangent space $\mathcal{T}_{\mathcal{P}}$ to a manifold M^n, at a point $\mathcal{P} \in M^n$, admits a basis $\{\boldsymbol{\partial}_\nu\big|_{\mathcal{P}}\}$ ($\nu = 1, \ldots, n$) called the **coordinate** (aka chart-induced) basis for the n local coordinates x^ν that parametrise M^n.

To find these coordinate basis vectors, we freeze all the variables that parametrise the manifold, except one. Varying that remaining parameter generates a so-called **coordinate curve** whose tangent at a point is the partial derivative with respect to one of the parameters on which the coordinate functions x^μ of a point depend, Then the components of $\boldsymbol{\partial}_\mu$ are simply the partial derivatives with respect to the parameters (coordinates) *on* the manifold of the position vector $\mathbf{x} \in \mathbb{R}^n$ of a point. An example should make this procedure clearer:

[†]Infinite linear combinations (series) require extra topological structure on $\mathcal{V}$ so as to allow the notion of convergence.

Example 1.4. On S^2 (embedded in $\mathbb{R}^3$), a point is mapped into the spherical coordinates (θ, ϕ), with $\theta \neq (0,\, \pi)$; it can also be described by the $\mathbb{R}^3$ coordinates $(\sin\theta\, \cos\phi, \sin\theta\, \sin\phi, \cos\theta)$. Freezing one of these generates a great circle on the sphere. Then $(\sin\theta\, \cos\phi,\ \theta\, \sin\phi, \cos\theta)$, with θ fixed, describe a circle of radius $\sin\theta$ at "colatitude" θ, and $\boldsymbol{\partial_\phi}$ is a coordinate vector which can be visualised in $\mathbb{R}^3$ by the vector with components:

$$\partial_\phi(\sin\theta\, \cos\phi, \sin\theta\, \sin\phi, \cos\theta) \;=\; (-\sin\theta\, \sin\phi, \sin\theta\, \cos\phi, 0)$$

At each point on S^2 parametrised by (θ, ϕ), this is a vector tangent to the circle at colatitude θ.

Similarly, there is a spherical-coordinate vector, $\boldsymbol{\partial_\theta}$, tangent to a meridian going through that same point, with components $(\cos\theta\, \cos\phi, \cos\theta\, \sin\phi, -\sin\theta)$. $\boldsymbol{\partial_\theta}$ and $\boldsymbol{\partial_\phi}$ together form a basis for vectors in the plane tangent to S^2 at that point. *These vectors do not live in* S^2! Instead, *any* vector on S^2 attached to that point lives in the tangent $\mathbb{R}^2$ plane. Also, each point on S^2 has its own tangent plane.

1.2.5 The space dual to a vector space

Given a basis $\{\mathbf{e}_\alpha\}$ for $\mathcal{V}^m$ and a basis $\{\mathbf{e}_{\beta'}\}$ for $\mathcal{W}^n$, we can easily construct a basis for the space of linear mappings, $\mathcal{L}(\mathcal{V}^m,\, \mathcal{W}^n)$, introduced in section 1.1.1. This will be the set $\{\mathbf{E}^\gamma_{\beta'}\}$ such that:

$$\mathbf{E}^\gamma_{\beta'}(\mathbf{e}_\alpha) \;=\; \delta_\alpha{}^\gamma\, \mathbf{e}_{\beta'} \qquad \gamma = 1,\, \ldots,\, m;\ \beta' = 1,\, \ldots,\, n \tag{1.8}$$

where $\delta_\alpha{}^\gamma$ is the Kronecker delta, and the left-hand side should be understood as the **action** (eg., matrix multiplication) of the mappings $\mathbf{E}^\gamma_{\beta'}$ on the vectors $\mathbf{e}_\alpha$

Do keep in mind that indices on a **bold**-character object will always label the object itself, *not* its components, which will never be bold. Thus, a particular linear mapping in the basis we have constructed takes the form:

$$\mathbf{T} \;=\; T^{\beta'}{}_\alpha\, \mathbf{E}^\alpha_{\beta'} \tag{1.9}$$

where the *matrix* T with *elements* $T^{\beta'}{}_\alpha$ (with the row index on the left) represents the *mapping* $\mathbf{T}$ in the *basis* $\{\mathbf{E}^\alpha_\beta\}$. Thus, the action of $\mathbf{T}$ on some vector $\mathbf{v}$ is:

$$\mathbf{T}(\mathbf{v}) \;=\; T^{\beta'}{}_\alpha\, v^\nu\, \mathbf{E}^\alpha_{\beta'}(\mathbf{e}_\nu) \;=\; T^{\beta'}{}_\alpha\, v^\nu\, \delta_\nu{}^\alpha\, \mathbf{e}_{\beta'} \;=\; T^{\beta'}{}_\alpha\, v^\alpha\, \mathbf{e}_{\beta'}$$

where the $T^{\beta'}{}_\alpha\, v^\alpha$ are seen to be the components of the vector $\mathbf{T}(\mathbf{v})$ in basis $\mathbf{e}_{\beta'}$ of $\mathcal{W}$. Whenever $\mathcal{W} = \mathcal{V}$, that basis would be the same as the initial basis.

In this last equation, as well as in eq. (1.9) and in $\mathbf{v} = v^\nu \mathbf{e}_\nu$, the left-hand side is explicitly basis-independent; this notation we shall call **index-free**, or **geometric**. The right-hand side, in so-called **index notation**, makes explicit reference to a basis even though, taken *as a whole*, it is still basis-independent. Both notations have advantages and disadvantages which we shall discuss later. Fluency in both is highly recommended.

A very interesting subset of the set of linear mappings is $\mathrm{Hom}\,(\mathcal{V}, \mathbb{R}) = \mathcal{L}(\mathcal{V},\, \mathbb{R})$, that is, the set of all linear functions on a vector space which takes their value in the same scalar field over which $\mathcal{V}$ is defined, here the real numbers. It forms a space $\mathcal{V}^*$ **dual** to $\mathcal{V}$. Since $\mathcal{L}(\mathcal{V}^m,\, \mathcal{W}^n)$ has dimension $m \times n$, $\mathcal{V}^*$ and $\mathcal{V}$ have the same[†] dimension. The elements of $\mathcal{V}^*$ are (informally! — they really only make sense in relation to the underlying dual space) called **covectors**, or **linear functionals** (in linear algebra), or **1-forms**. An example would be the definite integral of a polynomial.

Go back to the expansion of a general mapping over a basis, eq. (1.9). If such a mapping is an element of $\mathcal{V}^*$, $\beta' = 1$ and, dropping that index, we write: $\mathbf{T} \;= T_\alpha\, \mathbf{E}^\alpha$, where $\mathbf{T}$ is represented by the elements T_α ($\alpha = 1,\, \ldots,\, n$) of a *row vector*, and $\mathbf{T}(\mathbf{v}) = T_\alpha v^\alpha \in \mathbb{R}$. *Notation change alert*: from now on, we switch to greek letters to denote 1-forms.

Eq. (1.8) also undergoes considerable simplification. Putting $\beta' = 1$, and noting that $\mathbf{e}_1$ on the right-hand side, being a basis for $\mathbb{R}$, is just a number that can be set to 1, we obtain the following important definition:

[†]This assumes that $\mathcal{V}$'s dimension is finite!

Definition 1.11. If $\{\mathbf{e}_\alpha\}$ is a basis of a vector space $\mathcal{V}^n$, then its *unique* **dual basis** (**cobasis**) in $\mathcal{V}^*$, $\{\boldsymbol{\omega}^\alpha\}$, satisfies:

$$\boldsymbol{\omega}^\alpha(\mathbf{e}_\beta) \;=\; \delta^\alpha{}_\beta \qquad \alpha,\, \beta = 1,\, \ldots,\, n \tag{1.10}$$

where $\boldsymbol{\omega}^\alpha \equiv \mathbf{E}^\alpha$, and the left-hand side is just classic matrix multiplication. Other basis for $\mathcal{V}^*$ are of course possible, but this is the most useful one.

From this we derive the action (sometimes used to define $\boldsymbol{\omega}^\alpha$ instead of eq. (1.10)) of an element $\boldsymbol{\omega}^\alpha$ of the cobasis of $\mathcal{V}^*$ on a vector $\mathbf{v} \in \mathcal{V}$:

$$\boldsymbol{\omega}^\alpha(\mathbf{v}) \;=\; \boldsymbol{\omega}^\alpha(v^\beta \mathbf{e}_\beta) \;=\; v^\beta\, \boldsymbol{\omega}^\alpha(\mathbf{e}_\beta) \;=\; v^\beta\, \delta^\alpha{}_\beta \;=\; v^\alpha$$

from which we conclude that the cobasis element $\boldsymbol{\omega}^\alpha$ projects out, or picks out, the corresponding component of $\mathbf{v}$. This will probably come as some surprise to many, who are used to think of v^α as the projection of $\mathbf{v}$ on $\mathbf{e}_\alpha$.

What happens if we act on some $\mathbf{e}_\alpha$ with a 1-form (covector) $\boldsymbol{\sigma} = \sigma_\beta \boldsymbol{\omega}^\beta$? Well,

$$\boldsymbol{\sigma}(\mathbf{e}_\alpha) \;=\; \sigma_\beta\, \boldsymbol{\omega}^\beta(\mathbf{e}_\alpha) \;=\; \sigma_\beta\, \delta^\beta_\alpha \;=\; \sigma_\alpha$$

This shows that the α^{th} component of a 1-form in a cobasis $\{\boldsymbol{\omega}^\alpha\}$ is calculated by acting with the 1-form on the corresponding element of the basis to which the cobasis is dual.

Recall the one-to-one correspondence between a vector $\mathbf{v}$ and the n-tuple of its components in a basis $\{\mathbf{e}_\alpha\}$, $(v^1,\, \ldots\, ,\, v^n) \in \mathbb{R}^n$. An analog correspondence exists between a 1-form, $\boldsymbol{\sigma}$, and its components σ_α:

$$\mathbf{v} \;\longmapsto\; (v^1\; v^2\; \ldots\; v^n)^{\mathrm{T}} \qquad\qquad \boldsymbol{\sigma} \;\longmapsto\; \begin{pmatrix}\sigma_1 & \sigma_2 & \ldots & \sigma_n\end{pmatrix}$$

Therefore, we can also think of $\boldsymbol{\sigma}$ as a linear mapping from $\mathbb{R}^n$ to $\mathbb{R}$, or as a procedure to obtain the number $\boldsymbol{\sigma}(\mathbf{v}) = \sigma_\alpha v^\alpha$ out of the vector $\mathbf{v}$ via standard multiplication of a row vector with components σ_α by a column vector with components v_β:

$$(v^1\; v^2\; \ldots\; v^n)^{\mathrm{T}} \underset{\boldsymbol{\sigma}}{\longmapsto} \begin{pmatrix}\sigma_1 & \sigma_2 & \ldots & \sigma_n\end{pmatrix} (v^1\; v^2\; \ldots\; v^n)^{\mathrm{T}} \;=\; \sigma_\alpha\, v^\alpha \tag{1.11}$$

1.2.6 Differential of a function and basis dual to a coordinate basis

Definition 1.12. Let $x^\mu \in \mathbb{R}^n$ ($\mu = 1, \ldots, n$) be the coordinate functions of *arbitrary* coordinates at a point $\mathcal{P} \in M^n$, and $f \in C^\infty(M)$ a real-valued diffeentiable function on M^n. Let also $\mathbf{t} \in \mathcal{T}_\mathcal{P}$ be a vector tangent to M^n at $\mathcal{P}$. Then the **differential** of f at $\mathcal{P}$, $\mathbf{d}f$, is defined as the 1-form in $\mathcal{T}^*_\mathcal{P}$ which, when $\mathbf{t}$ is inserted in its input slot, yields the basis-independent action of $\mathbf{t}$ on f at $\mathcal{P}$ from eq. (1.7):

$$[\mathbf{d}f](\mathbf{t}) \;\coloneqq\; \mathbf{t}(f) \;=\; \partial_\mathbf{t} f \tag{1.12}$$

Now, if we take $\mathbf{t}_\mathcal{P} = (\boldsymbol{\partial}_\nu)_\mathcal{P}$, the vector tangent to the coordinate curve for x^ν, and also $f = x^\mu$, eq. (1.12) at $\mathcal{P}$ gives immediately:

$$(\mathbf{d}x^\mu)_\mathcal{P}(\boldsymbol{\partial}_{\boldsymbol{x}^\nu})_\mathcal{P} \;=\; \boldsymbol{\partial}_{x^\nu}(x^\mu) \;=\; \partial_\nu(x^\mu \circ x^{-1})(x_\mathcal{P}) \;=\; \delta^\mu{}_\nu$$

Choosing $\{\boldsymbol{\partial}_\mu\}$ as basis for $\mathcal{T}_\mathcal{P}$, we conclude from eq. (1.10) that $\{\mathbf{d}x^\mu\}$ *is the basis, dual to* $\{\boldsymbol{\partial}_\mu\}$*, of the* **cotangent space**, $\mathcal{T}^*_\mathcal{P}$, *dual to* $\mathcal{T}_\mathcal{P}$. In other words, when $\mathbf{e}_\mu = \boldsymbol{\partial}_\mu$, then $\boldsymbol{\omega}^\mu = \mathbf{d}x^\mu$. A general 1-form, when written in a coordinate cobasis as $\boldsymbol{\sigma} = \sigma_\alpha \mathbf{d}x^\alpha$, is often called a **differential form**. Warning: in general, a differential 1-form is *not* the differential of a function!

To find the components of $\mathbf{d}f$ in the $\{\mathbf{d}x^\mu\}$ coordinate cobasis, recall that the action of a cobasis on a vector outputs the corresponding component of the vector: $\mathbf{d}x^\mu(\mathbf{t}) = t^\mu$. Then, from eq. (1.7) and eq. (1.12):

$$[\mathbf{d}f](\mathbf{t}) \;=\; \mathbf{t}(f) \;=\; t^\mu \partial_\mu f \;=\; \partial_\mu f\, \mathbf{d}x^\mu(\mathbf{t}) \tag{1.13}$$

so that the components are $\partial_\mu f$. If we hink of f as a 0-form, the differential of f is the **gradient** 1-form $\mathbf{d}f$:

$$\mathbf{d}f \;=\; \partial_\mu f\, \mathbf{d}x^\mu \tag{1.14}$$

We recognise the well-known expression for the differential of a function in calculus, where it is taken to be a scalar, a number. But $\mathbf{d}f$, interpreted as the infinitesimal change of f *does not know in which direction this change should be evaluated*. Only when a vector is inserted in its slot, as in eq. (1.13), can it output a number, the change of f in the direction of the vector.

As for the usual calculus interpretation of $\mathbf{d}x^\mu$ as the difference between the components of two coordinate vectors at infinitesimally close points, this may be acceptable in $\mathbb{R}^n$, but not if x^μ are coordinates on an arbitrary manifold, since $\mathbf{d}x^\mu$, like all vectors and p-forms at a point, does not live on the manifold, but in the tangent (co)space. Only in $\mathbb{R}^n$ can one ignore with impunity this crucial distinction between a **base manifold** and its tangent space at a point.

1.2.7 Vectors as linear mappings

Since $\mathcal{L}(\mathcal{V},\,\mathbb{R})$, or $\mathcal{V}^*$, is technically a vector space, it has its own dual space, $\mathcal{L}(\mathcal{V}^*,\,\mathbb{R})$, or $\mathcal{V}^{**}$. We realise that nothing prevents us (*with finite-dimensional spaces*) from considering the elements $\mathbf{v} \in \mathcal{V}$ as *themselves linear mappings* on $\mathcal{V}^*$, and identifying $\mathcal{V}^{**}$ with $\mathcal{V}$! Then $\mathbf{e}_\alpha(\boldsymbol{\omega}^\beta) = \delta_\alpha{}^\beta$, and we would find: $\mathbf{v}(\boldsymbol{\sigma}) = v^\alpha\sigma_\alpha$, exactly as in eq. (1.11) above.

These considerations suggest that we can also view a 1-form (covector), as a kind of machine[†] whose input a vector and whose output is a number; similarly, a vector can function as a device with a 1-form as input and a number as output. To summarise, and noting that indices on vectors and covectors are labels specifying which of them, not which of their components, we construct a table that tells us how to calculate the output of a 1-form acting on a vector, and of a vector acting on a 1-form:

1-form	**Input vector**	**Output**	**Vector**	**Input 1-form**	**Output**
Cobasis element $\boldsymbol{\omega}^\alpha$	Basis element $\mathbf{e}_\beta$	$\boldsymbol{\omega}^\alpha(\mathbf{e}_\beta) = \delta^\alpha{}_\beta$	$\mathbf{e}_\alpha$	$\boldsymbol{\omega}^\beta$	$\mathbf{e}_\beta(\boldsymbol{\omega}^\alpha) = \delta^\alpha{}_\beta$
Cobasis element $\boldsymbol{\omega}^\alpha$	$\mathbf{v}$	$\boldsymbol{\omega}^\alpha(\mathbf{v}) = v^\alpha$	$\mathbf{e}_\alpha$	$\boldsymbol{\sigma}$	$\mathbf{e}_\alpha(\boldsymbol{\sigma}) = \sigma_\alpha$
$\boldsymbol{\sigma}$	Basis element $\mathbf{e}_\alpha$	$\boldsymbol{\sigma}(\mathbf{e}_\alpha) = \sigma_\alpha$	$\mathbf{v}$	$\boldsymbol{\omega}^\alpha$	$\mathbf{v}(\boldsymbol{\omega}^\alpha) = v^\alpha$
$\boldsymbol{\sigma}$	$\mathbf{v}$	$\boldsymbol{\sigma}(\mathbf{v}) = \sigma_\alpha v^\alpha$	$\mathbf{v}$	$\boldsymbol{\sigma}$	$\mathbf{v}(\boldsymbol{\sigma}) = v^\alpha\sigma_\alpha$

Note that $\boldsymbol{\sigma}(\mathbf{v}) = \mathbf{v}(\boldsymbol{\sigma}) = \sigma_\alpha v^\alpha$ is basis-independent, but only if $\boldsymbol{\sigma}$ is referred to the cobasis of the basis in which $\mathbf{v}$ is written. At this stage, there is no natural (basis-independent) isomorphism between a vector space and its dual space, because there is as yet no unique connection between a given $\mathbf{v} \in \mathcal{V}$ and an element of $\in \mathcal{V}^*$. So, tempting as it is to identify it with the scalar product of two vectors, let us resist that urge. The σ_α are components of a 1-form, not of a vector! As we shall soon discover, vector and 1-form components behave differently under the same linear transformation. For the moment $\sigma_\alpha v^\alpha$ is the real number resulting from the 1-form $\boldsymbol{\sigma}$ acting on the vector $\mathbf{v}$. This mapping is represented by a linear combination of the components of $\mathbf{v}$ with the coefficients of the combination the components of $\boldsymbol{\sigma}$ or, equivalently, by matrix multiplication.

For a given vector $\mathbf{v}$, there exists a unique set of parallel $(n$-1)-dimensional hyperplanes that can provide a geometric picture of 1-forms. This is very easy to do when $n = 2$ (vector in a plane). Then any number $a = \sigma_1 v^1 + \sigma_2 v^2$ determines a straight line perpendicular to $\mathbf{v}$ with equation $\sigma_2 = a/v^2 - \sigma_1 v^1/v^2$. The lines generated by different values of a all have identical slope $-v^1/v^2$.

[†]So far as I know, this metaphor was first proposed by Misner, Thorne and Wheeler (MTW) in their monumental textbook, *Gravitation*.

1.2.8 Transformations on bases, cobases, and components

Let (U_1, x) and (U_2, y) be two overlapping charts (see definition 1.2) on a manifold M, with x and y their coordinate functions, respectively. Consider a point $\mathcal{P} \in U_1 \cap U_2$. Let us obtain the relation between $\boldsymbol{\partial}_{x^\mu}\big|_{x_\mathcal{P}}$ and $\boldsymbol{\partial}_{y^\nu}\big|_{y_\mathcal{P}}$, the coordinates bases for the two charts. These are maps, which we let act on some arbitrary differentiable function f. We remember that because f acts on the manifold, we must write $\boldsymbol{\partial}_{x^\mu} f\big|_{x_\mathcal{P}} = \partial_\mu (f \circ x^{-1})\big|_{x_\mathcal{P}}$. Insert $y^{-1} \circ y$ and use the multidimensional version of the chain rule $(f \circ g)'(\mathcal{P}) = g'(\mathcal{P})\, f'[g(\mathcal{P})]$ (written in the order opposite the usual one):

$$\begin{aligned} \boldsymbol{\partial}_{x^\mu} f\Big|_{x_\mathcal{P}} &= \partial_\mu \Big[(f \circ y^{-1}) \circ (y \circ x^{-1})\Big]_{x_\mathcal{P}} \\ &= \partial_\mu (y \circ x^{-1})^\nu\big|_{x_\mathcal{P}}\, \partial_\nu (f \circ y^{-1})\big|_{(y \circ x^{-1})(x_\mathcal{P})} \\ &= \partial_{x^\mu} y^\nu\big|_{x_\mathcal{P}}\, \boldsymbol{\partial}_{y^\nu} f\big|_{y_\mathcal{P}} \end{aligned} \tag{1.15}$$

A vector $\mathbf{v} \in \mathcal{T}_\mathcal{P}$ must remain invariant under change of chart. That is: $\mathbf{v} = v_x^\mu \boldsymbol{\partial}_{x^\mu}\big|_{x_\mathcal{P}} = v_y^\lambda \boldsymbol{\partial}_{y^\lambda}\big|_{y_\mathcal{P}}$. Inserting the transformation law for the coordinate bases, we immediately find he transformation law for the components of $\mathbf{v}$:

$$v_y^\nu \;=\; \partial_{x^\mu} y^\nu\big|_{x_\mathcal{P}}\, v_x^\mu \tag{1.16}$$

What is remarkable of both transformations is that they are *linear* and *homogeneous*, even though the transformations beween (U_1, x) and (U_2, y) can be non-linear. Thus, in coordinate bases, the coefficients, $\partial_{x^\mu} y^\nu$, in the transformation law are the elements of the **Jacobian matrix of the transformation** evaluated at $\mathcal{P}$. Conversely, if $v_x^\nu = v_y^\mu \partial_{y^\mu} x^\nu$, one shows easily (do it!), using the chain rule on partial derivatives, that $\mathbf{v}$ is indeed unchanged by the transformation.

In general bases, the transformations must be assumed homogeneous and linear, and take the form:

$$\mathbf{e}_\mu \;=\; A^{\alpha'}{}_\mu\, \mathbf{e}_{\alpha'} \;=\; \mathbf{e}_{\alpha'} A^{\alpha'}{}_\mu \tag{1.17}$$

where the prime refers to the y coordinates in eq. (1.15) and (1.16). Components transform as:

$$v^{\alpha'} \;=\; v^\mu\, A^{\alpha'}{}_\mu \;=\; A^{\alpha'}{}_\mu\, v^\mu \tag{1.18}$$

This is the more traditional definition of a vector.

The two ways of writing $v^{\alpha'}$ in eq. (1.18) are equivalent, but the second one is a matrix product. The second expression in eq. (1.17) (and also eq. (1.15)), however, is *not* matrix multiplication , because the subscript of the basis vector is a *label* for a whole vector, not for a component of this vector.

$\mathbf{A}$ being non-singular, and therefore invertible, the action of the inverse transformation $\mathbf{A}^{-1}$ is represented by:

$$\mathbf{v} \;=\; \mathbf{A}^{-1}\mathbf{v}' \qquad \Longleftrightarrow \qquad v^\mu \;=\; (\mathbf{A}^{-1})^\mu{}_{\nu'} v^{\nu'} \tag{1.19}$$

Do not confuse matrix and index notation! Whereas matrix notation is readily translated into index notation, the reverse generally requires some rearrangement. This is because index notation does not care about ordering—one of its virtues—but matrix notation most certainly does.

Let $\{\mathbf{e}_\mu\}$ and $\{\mathbf{e}_{\mu'}\}$ be two bases in $\mathcal{V}^n$, connected by $\mathbf{e}_\mu = \mathbf{e}_{\nu'} A^{\nu'}{}_\mu$, where the $A^{\nu'}{}_\mu$ are the coefficients of the matrix $\mathbf{A}$ representing a linear transformation $\mathbf{A}$. Let $\{\boldsymbol{\omega}^\alpha\}$ and $\{\boldsymbol{\omega}^{\alpha'}\}$ be their two respective cobases in $\mathcal{V}^*$. Then, writing $\boldsymbol{\omega}^\alpha = B^\alpha{}_{\beta'} \boldsymbol{\omega}^{\beta'}$ where the $B^\alpha{}_{\beta'}$ are the matrix coefficients of the corresponding transformation $\mathbf{B}$ between the cobases, it can be shown (EXERCISE) that $\mathbf{B}$ is the inverse of $\mathbf{A}$, ie. $B^\alpha_{\nu'} A^{\nu'}{}_\beta = \delta^\alpha{}_\beta$ in index notation and $\mathbf{B} = \mathbf{A}^{-1}$ in matrix notation. This means that the transformation that takes the *unprimed* cobasis to the *primed* cobasis must be $\mathbf{A}$: $\boldsymbol{\omega}^{\alpha'} = A^{\alpha'}{}_\beta \boldsymbol{\omega}^\beta$.

In the same way as for vector components, we can then obtain (EXERCISE) the transformation law of the components σ_α of a 1-form $\boldsymbol{\sigma}$. Since $\boldsymbol{\sigma}$ must be cobasis-independent, $\sigma_\alpha \boldsymbol{\omega}^\alpha = \sigma_{\beta'} \boldsymbol{\omega}^{\beta'}$ yields:

$$\sigma_{\alpha'} \;=\; \sigma_\mu\, (\mathbf{A}^{-1})^\mu{}_{\alpha'} \tag{1.20}$$

while the inverse matrix, $\mathbf{B} = \mathbf{A}^{-1}$, takes the components in the opposite direction.

The following table, with $\mathbf{A} \equiv \mathbf{L}$ for clarity, summarises all the possible transformations, with the derivative expressions for components applying in coordinate bases:

$\mathbf{e}_{\alpha'} = \mathbf{e}_\beta (\mathbf{L}^{-1})^\beta{}_{\alpha'} = \mathbf{e}_\beta \partial_{\alpha'} x^\beta$	$\mathbf{e}_\alpha = \mathbf{e}_{\beta'} L^{\beta'}{}_\alpha = \mathbf{e}_{\beta'} \partial_\alpha x^{\beta'}$
$v^{\alpha'} = L^{\alpha'}{}_\beta \, v^\beta = \partial_\beta x^{\alpha'} \, v^\beta$	$v^\alpha = (\mathbf{L}^{-1})^\alpha{}_{\beta'} v^{\beta'} = \partial_{\beta'} x^\alpha \, v^{\beta'}$
$\boldsymbol{\omega}^{\alpha'} = L^{\alpha'}{}_\beta \boldsymbol{\omega}^\beta = \partial_\beta x^{\alpha'} \, v^\beta$	$\boldsymbol{\omega}^\alpha = (\mathbf{L}^{-1})^\alpha{}_{\beta'} \boldsymbol{\omega}^{\beta'} = \partial_{\beta'} x^\alpha \, \boldsymbol{\omega}^{\beta'}$
$\sigma_{\alpha'} = \sigma_\beta \, (\mathbf{L}^{-1})^\beta{}_{\alpha'}$	$\sigma_\alpha = \sigma_{\beta'} L^{\beta'}{}_\alpha = \sigma_{\beta'} \, \partial_\alpha x^{\beta'}$
$\sigma_\alpha \, v^\alpha = \sigma_{\beta'} \, v^{\beta'}$	

Care should be exercised when comparing this table to the expressions given in §2.9 and in Box 8.4 of MTW which refer to Lorentz transformations. In their potentially confusing but standard notation, the matrix with elements $L^\alpha{}_{\beta'}$ is actually the *inverse* of the matrix with elements $L^{\beta'}{}_\alpha$; we prefer making this explicit by writing $(\mathbf{L}^{-1})^\alpha{}_{\beta'}$.

Another word of caution: transformations in coordinate bases may well produce components in non-normalised bases, even if one starts from a normalised basis. This does not occur in the case of rotations and Lorentz boosts, but it will when we transform from Cartesian to curvilinear coordinates.

Also, if we wish to stick to a coordinate basis, we cannot call $\partial_\mu f$ the components of the gradient *vector*, $\boldsymbol{\nabla} f$. These do not transform as the components of a vector, as can be seen by calculating $\partial_{\mu'} f$ in terms of $\partial_\nu f$ using the chain rule (EXERCISE) and comparing with the form of the above transformation matrix in terms of derivatives. Moreover, everyone knows that in curvilinear coordinates (polar, spherical, cylindrical), the components of the gradient vector are not just the partial derivatives with respect to the coordinates. But even in Cartesian coordinates, which is a coordinate basis, $\partial_\mu f$ are not the components of a vector, but of a 1-form!

1.3 At Last, Tensors!

Our previous discussions make it straightforward to extend the concept of linear mappings to that of **multilinear** mappings, ie. mappings which are linear in each of their arguments, with the other arguments held fixed.

With $\mathcal{V}^n$ and $\mathcal{V}^*$ its dual space, equipped respectively with coordinate basis $\{\boldsymbol{\partial}_{\nu_i}\}$ and cobasis $\{\mathbf{d}x^{\mu_i}\}$ ($1 \leq i \leq n$), we construct the space of multilinear mappings, $\{T : \mathcal{V}^* \times \ldots \times \mathcal{V}^* \times \mathcal{V}^n \times \ldots \times \mathcal{V}^n \longrightarrow \mathbb{R}\}$, in the following way:

Definition 1.13. Contravariant tensors $\mathbf{T} \in \mathcal{T}^r$ of rank r are real multilinear functions of r *1-forms*:

$$\mathbf{T}(\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_r) = \sigma_{\mu_1} \ldots \sigma_{\mu_r} \, \mathbf{T}(\mathbf{d}x^{\mu_1}, \ldots, \mathbf{d}x^{\mu_r}) = T^{\mu_1 \cdots \mu_r} \, \sigma_{\mu_1} \ldots \sigma_{\mu_r} \tag{1.21}$$

Covariant tensors $\mathbf{S} \in \mathcal{T}_s$ of rank s are real multilinear functions of s *vectors*:

$$\mathbf{S}(\mathbf{u}_1, \ldots, \mathbf{u}_s) = u^{\nu_1} \ldots u^{\nu_s} \, \mathbf{S}(\boldsymbol{\partial}_{\nu_1}, \ldots, \boldsymbol{\partial}_{\nu_s}) = S_{\nu_1 \ldots \nu_s} \, u^{\nu_1} \ldots u^{\nu_s} \tag{1.22}$$

Mixed tansors of **type** (r, s) are real functions of r covectors and s vectors:

$$\begin{aligned} \mathbf{Q}(\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_r, \mathbf{u}_1, \ldots, \mathbf{u}_s) &= \sigma_{\mu_1} \ldots \sigma_{\mu_r} \, u^{\nu_1} \ldots u^{\nu_s} \mathbf{Q}(\mathbf{d}x^{\mu_1}, \ldots, \mathbf{d}x^{\mu_r}, \boldsymbol{\partial}_{\nu_1}, \ldots, \boldsymbol{\partial}_{\nu_s}) \\ &= Q^{\mu_1 \cdots \mu_r}{}_{\nu_1 \ldots \nu_s} \, \sigma_{\mu_1} \ldots \sigma_{\mu_r} \, u^{\nu_1} \ldots u^{\nu_s} \end{aligned} \tag{1.23}$$

$T^{\mu_1 \cdots \mu_r} = \mathbf{T}(\mathbf{d}x^{\mu_1}, \ldots, \mathbf{d}x^{\mu_r})$ are the **contravariant** components of $\mathbf{T}$, $S_{\nu_1 \ldots \nu_s} = \mathbf{S}(\boldsymbol{\partial}_{\nu_1}, \ldots, \boldsymbol{\partial}_{\nu_s})$ the **covariant** components of $\mathbf{S}$, and $Q^{\mu_1 \cdots \mu_r}{}_{\nu_1 \ldots \nu_s} = (\mathbf{d}x^{\mu_1}, \ldots, \mathbf{d}x^{\mu_r}, \boldsymbol{\partial}_{\nu_1}, \ldots, \boldsymbol{\partial}_{\nu_s})$ the **mixed** components of $\mathbf{Q}$ *with respect to the chosen basis and cobasis*.

Following the metaphor of tensors as machines, to output a number from a (r, s) tensor, one must supply r 1-forms and s vectors as input, one for each slot.

1.3.1 The tensor product

There is an important kind of multilinear mapping we can construct, this time out of *known* building blocks.

Definition 1.14. The **Kronecker (tensor) product space** of $\mathcal{V}_1^*$ and $\mathcal{V}_2^*$ is a set of bilinear mappings $\mathcal{L}(\mathcal{V}_1, \mathcal{V}_2, \mathbb{R})$, denoted by $\mathcal{V}_1^* \times \mathcal{V}_2^*$, with as **product elements** the covariant tensor $\boldsymbol{\sigma} \otimes \boldsymbol{\tau}$:

$$\boldsymbol{\sigma} \otimes \boldsymbol{\tau}(\mathbf{u},\, \mathbf{v}) \;=\; \boldsymbol{\sigma}(\mathbf{u})\, \boldsymbol{\tau}(\mathbf{v}) \tag{1.24}$$

for all $\mathbf{u} \in \mathcal{V}_1$, $\mathbf{v} \in \mathcal{V}_2$, $\boldsymbol{\sigma} \in \mathcal{V}_1^*$, and $\boldsymbol{\tau} \in \mathcal{V}_2^*$.

Similarly, the product space $\mathcal{L}(\mathcal{V}_1^*,\, \mathcal{V}_2^*,\, \mathbb{R}) = \mathcal{V}_1 \times \mathcal{V}_2$, has as elements the contravariant tensor of rank 2:

$$\mathbf{u} \otimes \mathbf{v}(\boldsymbol{\sigma},\, \boldsymbol{\tau}) \;=\; \mathbf{u}(\boldsymbol{\sigma})\, \mathbf{v}(\boldsymbol{\tau}) \tag{1.25}$$

Finally, there are tensor product spaces $\mathcal{V}_1 \otimes \mathcal{V}_2^*$ with product elements $\mathbf{u} \otimes \boldsymbol{\sigma}(\boldsymbol{\tau},\, \mathbf{v}) = \mathbf{u}(\boldsymbol{\tau})\boldsymbol{\sigma}(\mathbf{v})$, and $\mathcal{V}_1^* \times \mathcal{V}_2$, with product elements $\boldsymbol{\sigma} \otimes \mathbf{v}(\mathbf{u},\, \boldsymbol{\tau}) = \boldsymbol{\sigma}(\mathbf{u})\, \mathbf{v}(\boldsymbol{\tau})$. If a tensor is the tensor product of other tensors, we say that it is **decomposable**.

It is important to note that the tensor product *is not commutative*!.

Example 1.5. Let P be the vector space whose elements are polynomials of some degree n. Such a space can be constructed provided we define addition and s-multiplication of polynomials. Then we can construct a map $\mathbf{T} \,:\, P \times P \,\to\, \mathbb{R}$ defined by $\int_0^1 p(x)\, q(x)\, \mathrm{d}x$, where $p,\; q \in P$. This *bilinear* map—call it the inner product— which takes as inputs two vectors and outputs a number, is a $(0, 2)$ tensor.

Now take $\mathcal{V}_1 = \mathcal{V}_2 = \mathcal{V}$. If $\{\boldsymbol{\partial}_\mu\}$ and $\{\boldsymbol{\partial}_\nu\}$ are coordinate bases for $\mathcal{V}$, then $\{\boldsymbol{\partial}_\mu \otimes \boldsymbol{\partial}_\nu\}$ is a coordinate basis for $\mathcal{V} \otimes \mathcal{V}$. Similarly, if $\{\mathbf{d}x^\alpha\}$ and $\{\mathbf{d}x^\beta\}$ are coordinate bases for $\mathcal{V}^*$, then $\{\mathbf{d}x^\alpha \otimes \mathbf{d}x^\beta\}$ is a coordinate basis for $\mathcal{V}^* \otimes \mathcal{V}^*$.

We assert that *any* contravariant tensor of rank 2 lives in $\mathcal{V} \times \mathcal{V}$, and *any* covariant rank-2 tensor lives in $\mathcal{V}^* \times \mathcal{V}^*$, that is:

$$\mathbf{A} \;=\; A^{\mu\nu} \boldsymbol{\partial}_\mu \otimes \boldsymbol{\partial}_\nu, \qquad\qquad \mathbf{B} \;=\; B_{\alpha\beta}\, \mathbf{d}x^\alpha \otimes \mathbf{d}x^\beta \tag{1.26}$$

Therefore, the action of $\mathbf{A}$ on pairs of one-forms and of $\mathbf{B}$ on pairs of vectors is given by:

$$\begin{aligned} \mathbf{A}(\boldsymbol{\sigma},\, \boldsymbol{\tau}) \;&=\; A^{\mu\nu} \boldsymbol{\partial}_\mu \otimes \boldsymbol{\partial}_\nu\, (\boldsymbol{\sigma},\, \boldsymbol{\tau}) \;=\; A^{\mu\nu} \boldsymbol{\partial}_\mu(\boldsymbol{\sigma})\, \boldsymbol{\partial}_\nu(\boldsymbol{\tau}) \;=\; A^{\mu\nu} \sigma_\mu \tau_\nu \\ \mathbf{B}(\mathbf{u},\, \mathbf{v}) \;&=\; B_{\alpha\beta}\, \mathbf{d}x^\alpha \otimes \mathbf{d}x^\beta(\mathbf{u},\, \mathbf{v}) \;=\; B_{\alpha\beta}\, \mathbf{d}x^\alpha(\mathbf{u})\, \mathbf{d}x^\beta(\mathbf{v}) \;=\; B_{\alpha\beta}\, u^\alpha v^\beta \end{aligned} \tag{1.27}$$

As we have said before, both $\mathbf{A}$ and $\mathbf{B}$ can be viewed as operators, or devices, requiring two 1-forms or two vectors, respectively, as *ordered input*, to output a product of numbers. But we can also input a single vector (1-form) and obtain a 1-form (vector) as output, so long as we specify into which of the two input slots it should be inserted. For instance, we could write $\mathbf{B}(\mathbf{u},\;\;)$, or $\mathbf{B}(\;\;,\, \mathbf{u})$, but just $\mathbf{B}(\mathbf{u})$ would (in general, but not always, as we shall see in a moment) be ambiguous. For instance:

$$\begin{aligned} \mathbf{B}(\mathbf{u},\;\;) \;&=\; B_{\alpha\beta}\, \mathbf{d}x^\alpha(\mathbf{u})\, \mathbf{d}x^\beta \;=\; \sigma_\beta\, \mathbf{d}x^\beta \\ \mathbf{B}(\;\;,\, \mathbf{u}) \;&=\; B_{\alpha\beta}\, \mathbf{d}x^\alpha \mathbf{d}x^\beta(\mathbf{u}) \;=\; \tau_\alpha\, \mathbf{d}x^\alpha \end{aligned}$$

where $\sigma_\beta = B_{\alpha\beta}\, u^\alpha$, and $\tau_\alpha = B_{\alpha\beta}\, u^\beta$. Unless the components $B_{\alpha\beta}$ happen to be symmetric in their indices, the two resulting 1-forms are not the same!

With this in mind, $\mathbf{A}(\boldsymbol{\sigma},\, \boldsymbol{\tau})$ is the same as $\boldsymbol{\sigma}\big(\mathbf{A}(\;\;,\, \boldsymbol{\tau})\big)$, where $\mathbf{A}(\;\;,\, \boldsymbol{\tau})$ is (exercise) a vector. In other words, the same machine can be put to quite different tasks.

More generally, if $\{\boldsymbol{\partial}_{\mu_1}\},\ldots,\{\boldsymbol{\partial}_{\mu_r}\}$ are bases of $\mathcal{V}$, and $\{\mathbf{d}x^{\nu_1}\},\ldots,\{\mathbf{d}x^{\nu_s}\}$ (co)bases of $\mathcal{V}^*$, then $\boldsymbol{\partial}_{\mu_1}\otimes\cdots\otimes\boldsymbol{\partial}_{\mu_r}\otimes\mathbf{d}x^{\nu_1}\otimes\cdots\otimes\mathbf{d}x^{\nu_s}$ forms a basis for $\mathcal{T}^r_s$. Therefore, any tensor can be written in terms of its components as:

$$\mathbf{T} \;=\; T^{\mu_1\ldots\mu_r}{}_{\nu_1\ldots\nu_s}\,\boldsymbol{\partial}_{\mu_1}\otimes\cdots\otimes\boldsymbol{\partial}_{\mu_r}\otimes\mathbf{d}x^{\nu_1}\otimes\cdots\otimes\mathbf{d}x^{\nu_s} \tag{1.28}$$

But, as we saw for rank-2 tensors, we can also input one less vector and get a 1-form as output, or one less 1-form to get a vector. More generally, reducing the number of input vectors from s to $s-m$ causes $\mathbf{T}$ to output a tensor of covariant rank $(0,m)$; reducing the number of input 1-forms from r to $r-q$ outputs a tensor of contravariant rank $(q,0)$.

It is important to remember that, in general, interchanging vectors or 1-forms in the input results in different output. In other words, one should be mindful of the ordering of the vectors and of the 1-forms that one feeds into a tensor.

1.3.2 Transposition, symmetric and skew-symmetric tensors

Interchanging any two contravariant or any two covariant slots of a tensor produces a **transpose** of this tensor. Strictly speaking, interchanging a covariant and a contravariant slot of a *tensor* does not make sense, Interchanging its slots generally changes a tensor.

> **Definition 1.15.** If a tensor remains unchanged under transposition of two of its slots of the same type, we say that it is **symmetric** in these slots. Its corresponding components are unchanged under permutation of indices.
>
> If a tensor switches sign under transposition of two of its slots of the same type, we say that it is **antisymmetric** in these slots, and the correspoinding components also switch sign. inserting the *same* 1-form (in the contravariant slots) or vector (in the covariant slots outputs zero
>
> Symmetry and antisymmetry are basis-independent properties.

Take a covariant antisymmetric tensor of rank 2: $\mathbf{F}=F_{\mu\nu}\,\mathbf{d}x^\mu\otimes\mathbf{d}x^\nu$. Then $\mathbf{F}(\mathbf{u},\,\mathbf{u})=F_{\mu\nu}u^\mu\,u^\nu=0$. To see this in index notation, just rename μ as ν and ν as μ and transpose indices in the antisymmetric $F_{\mu\nu}$ and symmetric $u^\mu\,u^\nu$ to get $F_{\mu\nu}u^\mu\,u^\nu=-F_{\mu\nu}u^\mu\,u^\nu$.

Similarly, inserting the same cobasis 1-forms in two antisymmetric contravariant slots, or basis vectors in two antisymmetric covariant slots, just makes the corresponding indices identical, and the corresponding components vanish.

Among important tensors are those which are **completely symmetric** in all their covariant indices and all their contravariant indices (eg. $A_{\mu_1\ldots\mu_s}=A_{\mu_{\pi_1}\ldots\mu_{\pi_s}}$ where $\mu_{\pi_1}\ldots\mu_{\pi_s}$ is any permutation of $\mu_1\ldots\mu_s$), and those which are completely antisymmetric (**skew-symmetric**, **alternating**) in all their covariant indices and all their contravariant indices.

A completely symmetric tensor of rank r in n dimensions has $\binom{n+r-1}{r}=(n+r-1)!/(n-1)!r!$ *independent* components. A skew-symmetric tensor has $\binom{n}{r}=n!/(n-r)!r!$ *independent* non-zero components in all coordinate systems.

In three dimensions, many physically relevant tensors are symmetric, eg. examples 1.6, 1.7 and 1.8 (moment of inertia, electrical polarisation, multipole moment) in B&F, as well as the Maxwell stress tensor. Antisymmetric 3-d rank-2 tensors are not usual, although I will argue toward the end of the module that a three dimensional magnetic field is more naturally described by a rank-2 antisymmetric covariant tensor than by a vector.

In four dimensions, we also have symmetric tensors, such as the important energy-momentum tensor which carries all the information about the energy and momentum density at a point, plus the flux of these quantities at that point. But there is also a famous antisymmetric rank-2 tensor, the Faraday field tensor F, whose six independent non-zero components are the components of electric and magnetic field 3-*vectors*.

Even when a tensor is neither completely symmetric nor skew-symmetric, it can be useful to construct a symmetrised or skew-symmetrised version of it. The simplest example is that of a rank-2 contravariant (or covariant)

tensor. In index notation,

$$T^{\mu\nu} \;=\; \frac{1}{2}(T^{\mu\nu} + T^{\nu\mu}) \;+\; \frac{1}{2}(T^{\mu\nu} - T^{\nu\mu}) \;\equiv\; T^{(\mu\nu)} \;+\; T^{[\mu\nu]}$$

where round (square) brackets around indices mean that they are symmetric (antisymmetric). In this case, we say that $\mathbf{T}$ has been decomposed into a symmetrised tensor, with components $T^{(\mu\nu)}$, and a skew-symmetric tensor, with components $T^{[\mu\nu]}$.

More generally, $(r, 0)$ and $(0, s)$ tensors can be symmetrised and antisymmetrised.

Definition 1.16. To symmetrise the components of a covariant (or contravariant) tensor $\mathbf{T}$, we can apply the following prescription to its components:

$$T_{(\mu_1 \ldots \mu_s)} \;=\; \frac{1}{s!} \sum_{\pi} T_{\mu_{\pi_1} \ldots \mu_{\pi_s}} \tag{1.29}$$

where the sums run over all permutations of $1, \ldots, s$.

Definition 1.17. To antisymmetrise the components of a covariant (or contravariant) tensor $\mathbf{T}$, we can apply the following prescription to its components:

$$T_{[\mu_1 \ldots \mu_s]} \;=\; \frac{1}{s!}\, \delta^{\nu_1 \ldots \nu_s}_{\mu_1 \ldots \mu_s}\, T_{\nu_1 \ldots \nu_s} \tag{1.30}$$

where the **general permutation symbol**, $\delta^{j_1 \ldots j_s}_{i_1 \ldots i_s}$, is defined as:

$$\delta^{j_1 \ldots j_s}_{i_1 \ldots i_s} \;=\; \begin{cases} +1 & j_1 \ldots j_s \text{ an even permutation of } i_1 \ldots i_s \\ -1 & j_1 \ldots j_s \text{ an odd permutation of } i_1 \ldots i_s \\ 0 & j_1 \ldots j_s \text{ not a permutation of } i_1 \ldots i_s \\ 0 & j_k \;=\; j_l \text{ or } i_k \;=\; i_l \text{ for some } k,\, l \end{cases} \tag{1.31}$$

The permutation symbol is seen to be antisymmetric in its upper and lower indices. In terms of components: Clearly, as expected, $T_{[\mu_1 \ldots \mu_s]} = 0$ whenever any two of its indices are the same.

$s!$ is the number of terms in all these summations, ie. the number of permutations of the indices of the tensor. The normalisation factor $1/s!$ ensures consistency in the event that the $T_{\mu_1 \ldots \mu_s}$ should already be symmetric or skew-symmetric. If $\mathbf{T}$ is symmetric, we can also write:

$$\mathbf{T} \;=\; T_{\mu_1 \ldots \mu_s} \sum_{\pi} \mathbf{d}x^{\mu_{\pi_1}} \otimes \cdots \otimes \mathbf{d}x^{\mu_{\pi_s}} \qquad\qquad \mu_1 \leq \mu_2 \leq \ldots \leq \mu_s$$

and, if $\mathbf{T}$ is skew-symmetric:

$$\mathbf{T} \;=\; T_{\mu_1 \ldots \mu_s}\, \delta^{\mu_1 \ldots \mu_s}_{\nu_1 \ldots \nu_s}\, \mathbf{d}x^{\nu_1} \otimes \cdots \otimes \mathbf{d}x^{\nu_s} \qquad\qquad \mu_1 < \mu_2 < \ldots < \mu_s$$

$(0, s)$ skew-symmetric tensors live in a space denoted by $\bigwedge^s(\mathcal{V}^*)$, whose elements are also called **p-forms** (p is traditionally used instead of s); very often, $\mathcal{V}^* = \mathbb{R}^n$.

EXERCISE: Symmetrise and antisymmetrise $\mathbf{F}(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\theta})$. Write $\mathbf{F}_{\mathrm{s}}$ and $\mathbf{F}_{\mathrm{a}}$ in an explicit basis such that $\mathbf{F} = F^{\mu\nu\lambda} \mathbf{e}_\mu \otimes \mathbf{e}_\nu \otimes \mathbf{e}_\lambda$. How many components do $\mathbf{F}_{\mathrm{s}}$ and $\mathbf{F}_{\mathrm{a}}$ have when $\mathbf{F}$ is defined over a 3-dim space? a 4-dim space? Can you reconstruct $F^{\mu\nu\lambda}$ from $F^{(\mu\nu\lambda)}$ and $F^{[\mu\nu\lambda]}$?

1.3.3 Transformations on tensors

Using the transformations for a basis and a cobasis in the table of section 1.2.8, it is straightforward to generalise the transformation laws obeyed by tensor components. First, write $\mathbf{T}$ in the original basis and in the new (primed) basis:

$$\mathbf{T} \;=\; T^{\mu_1\dots\mu_r}{}_{\nu_1\dots\nu_s}\,\boldsymbol{\partial}_{\mu_1}\otimes\cdots\otimes\boldsymbol{\partial}_{\mu_r}\otimes\mathbf{d}x^{\nu_1}\otimes\cdots\otimes\mathbf{d}x^{\nu_s} \;=\; T^{\alpha'_1\dots\alpha'_r}{}_{\beta'_1\dots\beta'_s}\,\boldsymbol{\partial}_{\alpha'_1}\otimes\cdots\otimes\boldsymbol{\partial}_{\alpha'_r}\otimes\mathbf{d}x^{\beta'_1}\otimes\cdots\otimes\mathbf{d}x^{\beta'_s}$$

We obtain:

$$T^{\alpha'_1\dots\alpha'_r}{}_{\beta'_1\dots\beta'_s} \;=\; T^{\mu_1\dots\mu_r}{}_{\nu_1\dots\nu_s}\,L^{\alpha'_1}{}_{\mu_1}\,\dots\,L^{\alpha'_r}{}_{\mu_r}\,(L^{-1})^{\nu_1}{}_{\beta'_1}\,\dots\,(L^{-1})^{\nu_s}{}_{\beta'_s} \tag{1.32}$$

In traditional treatments, this transformation law actually *defines* a tensor. Scalars (tensors of type $(0,\,0)$) must remain invariant; and we know how the components of vectors and 1-forms transform. What about, say, those of a tensor of type $(2,\,0)$?

$$T^{\alpha'\beta'} \;=\; T^{\mu\nu}\,L^{\alpha'}{}_{\mu}\,L^{\beta'}{}_{\nu} \;=\; L^{\alpha'}{}_{\mu}\,T^{\mu\nu}\,\tilde{L}_{\nu}{}^{\beta'} \qquad\Longleftrightarrow\qquad \mathsf{T}' \;=\; \mathsf{L}\,\mathsf{T}\,\tilde{\mathsf{L}}$$

where $\tilde{\mathsf{L}}$ is the transpose of L. Sometimes, as with 3-dim rotations, $\tilde{\mathsf{L}} = \mathsf{L}^{-1}$; sometimes, as with Lorentz boosts, $\tilde{\mathsf{L}} = \mathsf{L}$. Tensors of rank 2 $\big((2,\,0),\,(0,\,2),\,(1,\,1)\big)$ can be represented by $n\times n$-dim matrices T, where n is the dimension of the spaces $\mathcal{V}$ and $\mathcal{V}^*$ on which they are defined. Unless they are decomposable, this is not true of tensors of higher rank. But the index notation itself does not care about the rank of a tensor.

An immediate consequence of eq. (1.32) is that a tensor that is zero in one basis will remain zero in any other transformed basis. Thus, *any equation made of tensors that is valid in one basis must hold in any other basis*. This might seem trivial in the geometric notation $\mathbf{T}$ of tensors, but the nice thing is that relations between their components are also basis-independent!

In the old view of tensors defined in terms of transformations, it is possible to restrict the tensor character to objects which may have it under certain transformations, but not others. For instance, 4-dim tensors might owe their tensor character to how they transform under Lorentz transformations, while 3-dim tensors might be tensors only under rotations. Then writing equations in terms of 4-tensors guarantees that they are consistent with Einstein's Relativity, ie. an equation valid in one Lorentz frame is valid in any other Lorentz-transformed frame.

The transformation rules can always be used to establish whether an object is a tensor. For instance, In a space of dimension n, the Kronecker delta, with components $\delta^{\nu}{}_{\mu}$, has the $n\times n$ identity matrix as matrix representation. It is a rank-2 tensor. Indeed, from the transformation law, eq. (1.32):

$$\delta^{\mu'}{}_{\nu'} \;=\; L^{\mu'}{}_{\lambda}(L^{-1})^{\rho}{}_{\nu'}\,\delta^{\lambda}{}_{\rho} \;=\; L^{\mu'}{}_{\lambda}(L^{-1})^{\lambda}{}_{\nu'} \;=\; \mathsf{I}^{\mu'}{}_{\nu'}$$

which are the components of the identity matrix. Here we learn that there is something more to $\delta^{\mu}{}_{\nu}$ than just being a tensor: its components remain the same under changes of basis!

1.3.4 The Levi-Civita symbol

Definition 1.18. In a Cartesian orthonormal basis of a n-dim space, the **Levi-Civita symbol**, $\epsilon_{\mu_1\dots\mu_n}$, is defined in terms of the general permutation symbol, $\delta^{j_1\,\dots\,j_n}_{i_1\,\dots\,i_n}$ (eq. (1.31)), as:

$$\epsilon_{\mu_1\dots\mu_n} \;=\; \delta^{1\,\dots\,n}_{\mu_1\dots\mu_n}$$

It is skew-symmetric in its n indices, with $\epsilon_{1\dots n} = +1$, where the indices are in *ascending order*. In pseudo-Riemannian manifolds, it is traditional to use $\epsilon_{0\dots n-1}$, the 0 index corresponding to time.

The determinant of a matrix L is nothing but the antisymmetrised product of the matrix elements:

$$\det\mathsf{L} \;=\; \epsilon_{\nu_1\dots\nu_n}\,L^{\nu_1}{}_{1}\cdots L^{\nu_n}{}_{n} \qquad\Longleftrightarrow\qquad \epsilon_{\mu_1\dots\mu_n}\det\mathsf{L} \;=\; \epsilon_{\nu_1\dots\nu_n}\,L^{\nu_1}{}_{\mu_1}\cdots L^{\nu_n}{}_{\mu_n} \tag{1.33}$$

The right-hand equation means that the Levi-Civita symbol is a tensor only when its components are transformed with a matrix whose Jacobian determinant is 1 (eg. rotations in 3-dim space, Lorentz), in which case these components, like those of the Kronecker delta, are invariant under changes of basis. Indeed, if the Levi-Civita symbol is to be a tensor, the transformation laws on its components demand that:

$$1 \;=\; \epsilon_{1\ldots n} \;=\; \epsilon_{\nu'_1\ldots\nu'_n} L^{\nu'_1}{}_1 \cdots L^{\nu'_n}{}_n \;=\; \det \mathsf{L}$$

We shall discover a little later how the Levi-Civita general tensor (in all bases) can be constructed.

Fortunately, we can avoid using the transformation law if we build tensors from other objects known to be tensors. The following sub-sections present some important examples.

1.4 Two More Ways to Construct Tensors

1.4.1 Contracted tensors

Definition 1.19. The **contraction** of a mixed-type tensor is a linear mapping $\mathcal{T}^r_s \to \mathcal{T}^{r-1}_{s-1}$, ($r \geq 1$, $s \geq 1$). More precisely, going back to eq. (1.21), insert input (co)bases into *only two* tensor slots: $\boldsymbol{\sigma}^i = \mathbf{d}x^{\mu_i}$ ($1 \leq i \leq r$) and $\mathbf{u}_j = \boldsymbol{\partial}_{\nu_j}$ ($1 \leq j \leq s$), with $\nu_j = \mu_i \equiv \gamma$:

$$\mathbf{T}(\ldots, \mathbf{d}x^\gamma, \ldots, \boldsymbol{\partial}_\gamma, \ldots) \;=\; T^{\cdots\,\gamma\,\cdots}{}_{\cdots\,\gamma\,\cdots}\, \boldsymbol{\partial}_{\mu_1} \otimes \ldots \otimes \boldsymbol{\partial}_{\mu_{i-1}} \otimes \boldsymbol{\partial}_{\mu_{i+1}} \otimes \ldots \mathbf{d}x^{\nu_1} \otimes \ldots \otimes \mathbf{d}x^{\nu_{j-1}} \otimes \mathbf{d}x^{\nu_{j+1}} \otimes \ldots \tag{1.34}$$

Note the need to keep track of the position of the tensor slots. In terms of the components of a tensor, it just involves making one contravariant index μ the same as one of the covariant indices, ν, by multiplying the component by $\delta^\nu{}_\mu$, thus forcing a summation over these indices.

For instance, consider $\mathbf{T} \in \mathcal{T}^1_1$. The contraction of $\mathbf{T} = T^\alpha{}_\beta \boldsymbol{\partial}_\alpha \otimes \mathbf{d}x^\beta$ is a scalar, called its **trace**:

$$\mathrm{Tr}\,\mathbf{T} \;=\; \mathbf{T}(\mathbf{d}x^\mu,\, \boldsymbol{\partial}_\mu) \;=\; T^\alpha{}_\beta\, \boldsymbol{\partial}_\alpha(\mathbf{d}x^\mu)\, \mathbf{d}x^\beta(\boldsymbol{\partial}_\mu) \;=\; T^\mu{}_\mu \;=\; T^\mu{}_\nu\, \delta^\nu{}_\mu$$

When contracting tensors of type higher than 2, it is important to specify which indices are being contracted. Thus, the tensor $T^{\mu\nu}{}_\lambda \boldsymbol{\partial}_\mu \otimes \boldsymbol{\partial}_\nu \otimes \mathbf{d}x^\lambda$ has two possible contractions: the vectors $T^{\mu\nu}{}_\mu \boldsymbol{\partial}_\nu$ and $T^{\mu\nu}{}_\nu \boldsymbol{\partial}_\mu$.

1.4.2 Inner product

Up to now, in the space $\mathcal{T}^r_s$ of tensors, there has been no unique link between tensors of type $(r, 0)$, $(0, r)$, or $(r-q, q)$, all of the same total rank. If we wish to establish such a link, a new object must be introduced: a bilinear covariant tensor denoted by $\mathfrak{g} = g_{\mu\nu} \mathbf{d}x^\mu \otimes \mathbf{d}x^\nu$ in a coordinate basis.

Indeed, let us insert only one vector in, say, the second slot of $\mathfrak{g}$:

$$\mathfrak{g}(\quad, \mathbf{u}) \;=\; g_{\mu\nu}\, \mathbf{d}x^\mu\, \mathbf{d}x^\nu(\mathbf{u}) \;=\; g_{\mu\nu} u^\nu\, \mathbf{d}x^\mu \;\equiv\; \tilde{\mathbf{u}}$$

As expected, the result is a 1-form, which we have called $\tilde{\mathbf{u}}$. The correspondence will be unique if we demand that $\mathfrak{g}$ be *symmetric*, because then $\mathfrak{g}(\mathbf{u}, \quad) = \mathfrak{g}(\quad, \mathbf{u})$, or $g_{\nu\mu} = g_{\mu\nu} = \mathfrak{g}(\boldsymbol{\partial}_\mu, \boldsymbol{\partial}_\nu)$ when working in a coordinate basis. So, although sometimes called a form, $\mathfrak{g}$ is *not* a 2-form because it is not antisymmetric. As befits any $(0, 2)$ tensor, inserting vectors in its two input slots yields the number: $\mathfrak{g}(\mathbf{u}, \mathbf{v}) = g_{\mu\nu}\, u^\mu\, v^\nu = \mathbf{u}\,\mathbf{g}\,\mathbf{v}$, with the last expression in matrix form.

In effect, $\mathfrak{g}$ may be thought of as an *invertible* (ie. $\det \mathbf{g} \neq 0$) mapping from $\mathcal{V}$ to its dual space! Once we have defined it, it will establish a unique correspondence between a vector $\mathbf{u}$ and a 1-form $\tilde{\mathbf{u}}$.

Definition 1.20. The **inner product** of two vectors $\mathbf{u}$ and $\mathbf{v}$, $<\mathbf{u},\, \mathbf{v}>$, is defined as:

$$<\mathbf{u},\, \mathbf{v}> = \; \mathfrak{g}(\mathbf{u},\, \mathbf{v}) \tag{1.35}$$

Often, $<\mathbf{u},\, \mathbf{v}>$ is written $\mathbf{u} \cdot \mathbf{v}$.

Definition 1.21. For the components of $\mathfrak{g}$ we take: $g_{\mu\nu} =< \boldsymbol{\partial}_\mu, \boldsymbol{\partial}_\nu >$, which is just the naïve scalar product of the two basis vectors $\boldsymbol{\partial}_\mu$ and $\boldsymbol{\partial}_\nu$.

$\mathfrak{g}(\mathbf{u}, \mathbf{u}) = g_{\mu\nu}u^\mu u^\nu$ is called the **norm** of $\mathbf{u}$. If it is positive (negative) $\forall\, \mathbf{u}$, we say that $\mathfrak{g}$ is **positive (negative) definite**. Else, $\mathfrak{g}$ is **indefinite** and a non-zero **null** vector $\mathbf{u}$ exists for which $\mathfrak{g}(\mathbf{u}, \mathbf{u}) = 0$.

It is usual to speak of $g_{\mu\nu}u^\nu$ as the covariant components u_μ of $\mathbf{u}$ as well as of $\tilde{\mathbf{u}}$, In that sense $\mathbf{u}$ and $\tilde{\mathbf{u}}$ can both have covariant and conravariant components. When $\mathbf{u} = u^\mu\boldsymbol{\partial}_\mu$, we think of it as a vector; when we write it as $u_\nu \mathbf{d}x^\nu$, we think of it as a covector. Thus, we can say that the inner product sets up an isomorphism between a vector space and its dual. Because of this relation, we also have: $\boldsymbol{\partial}_\mu = g_{\mu\nu}\mathbf{d}x^\nu$.

As mentioned above, $\mathfrak{g}$ must be invertible, and we can always write: $\mathbf{u} = \mathfrak{g}^{-1}\mathfrak{g}(\mathbf{u}) = \mathfrak{g}^{-1}(\tilde{\mathbf{u}}) = u_\mu\mathfrak{g}^{-1}(\mathbf{d}x^\mu)$. Now $\mathfrak{g}^{-1}$ must take 1-forms to vectors, which means it must be a $(2,0)$ tensor $\mathfrak{g}^{-1} = (\mathfrak{g}^{-1})^{\mu\nu}\boldsymbol{\partial}_\mu \otimes \boldsymbol{\partial}_\nu$. Then:

$$\mathbf{u} \;=\; u_\mu\,(\mathfrak{g}^{-1})^{\alpha\beta}\boldsymbol{\partial}_\alpha\,\boldsymbol{\partial}_\beta(\mathbf{d}x^\mu) \;=\; u_\mu\,(\mathfrak{g}^{-1})^{\alpha\beta}\boldsymbol{\partial}_\alpha\,\delta^\mu{}_\beta \;=\; u_\mu\,(\mathfrak{g}^{-1})^{\alpha\mu}\,\boldsymbol{\partial}_\alpha$$

As will be justified soon, we identify $(\mathfrak{g}^{-1})^{\mu\nu}$ with the contravariant components of $\mathfrak{g}$, $g^{\mu\nu}$, and, comparing with $\mathbf{u} = u^\alpha\,\boldsymbol{\partial}_\alpha$, we conclude that $u^\mu = g^{\mu\nu}u_\nu$, u^μ being thought now as the contravariant components of the 1-form.

These mappings between $\mathcal{V}$ and $\mathcal{V}^*$ can be applied to any tensor $\mathbf{T} \in \mathcal{T}$; in other words, $\mathfrak{g}$ may be used to convert any contravariant index of a given tensor into a covariant one, while $\mathfrak{g}^{-1}$ may be used to convert any covariant index of a given tensor into a contravariant one. The total rank of the tensor is not changed.

Definition 1.22. Converting a contravariant index of a tensor $\mathbf{T}$ to a covariant one is called "**lowering the index**". It is always accomplished by contracting any one index of $g_{\mu\nu}$ with the index to be lowered, eg. $T^\mu{}_\nu = g_{\nu\lambda}T^{\mu\lambda} = \mathbf{T}\,\mathbf{g}$.

Converting a covariant index of a tensor $\mathbf{T}$ to a contravariant one is called "**raising the index**". It is always accomplished by contracting any one index of $g^{\mu\nu}$ with the index to be raised, eg. $T^\mu{}_\nu = g^{\mu\lambda}T_{\lambda\nu} = \mathbf{g}^{-1}\,\mathbf{T}$.

After raising an index, the corresponding input slot now accepts a 1-form as input instead of a vector. After lowering an index, the corresponding input slot now accepts a vector as input instead of a 1-form. But the tensor itself (as opposed to its components) does not change! For instance, let $\mathbf{T}$ be a $(2,1)$ tensor with components $T^{\mu\nu}{}_\lambda$. The $(1,2)$ tensor with components $T_\mu{}^\nu{}_\lambda = g_{\mu\alpha}T^{\alpha\nu}{}_\lambda$ is not a fundamentally different object, in the same way that we say that a vector (or a covector) can have both contravariant and covariant compoennts.

A word of caution here: before we introduced the inner product, we always wrote our matrices as $L^\mu{}_\nu$, with the upper index a row index. Why do we not write the matrix of $\mathfrak{g}$'s components the same way? Because $L^\mu{}_\nu$ is a *transformation* between two bases in $\mathcal{V}^n$, whereas $g_{\mu\nu}$ transforms from a basis in $\mathcal{V}^n$ to the *dual* basis. For instance, in $\mathbb{R}^3$, the basis dual to $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ cannot be reached by any combination of rotations and translations. Also, $L^\mu{}_\nu$ is *not* the component of a tensor, whereas $g_{\mu\nu}$ is.

As well as fully contravariant and covariant components $\mathfrak{g}$ also has mixed components which can easily be obtained by raising a covariant, or lowering a contravariant, index:

$$g^\mu{}_\nu \;=\; \mathfrak{g}(\mathbf{d}x^\mu, \boldsymbol{\partial}_\nu) \;\equiv <\mathbf{d}x^\mu, \boldsymbol{\partial}_\nu> =\; \mathbf{d}x^\mu(\boldsymbol{\partial}_\nu) \;=\; \delta^\mu{}_\nu \tag{1.36}$$

Since, as we have seen, $\delta^\mu{}_\nu$ is basis-independent, so is $g^\mu{}_\nu$, unlike $g_{\mu\nu}$and $g^{\mu\nu}$. But $g^\mu{}_\nu$ can also be obtained by index lowering (or raising): $g^{\mu\lambda}g_{\lambda\nu} = g^\mu{}_\nu = \delta^\mu{}_\nu$, which justifies our earlier assertion that $g^{\mu\nu}$ are the compoents of the inverse of $\mathfrak{g}$.

We also understand why, if $\delta^\mu{}_\nu$ are the components of the identity matrix, $\mathbf{I}$, $\delta_{\mu\nu} = g_{\mu\rho}\,\delta^\rho{}_\nu = g_{\mu\nu}$ will not in general be the entries of $\mathbf{I}$.

1.4.3 The metric

The inner product that we have introduced plays another extremely important rôle: it allows us to define distances and lengths in $\mathcal{V}^n$:

Definition 1.23. As a **metric tensor** (metric for short), $\mathfrak{g}$ tells us how to calculate lengths and distances in a vector space. The name is often extended (abusively) to its components $g_{\mu\nu}$.

$$\Delta s^2 \;=\; g_{\mu\nu}\,\Delta x^\mu\,\Delta x^\nu \tag{1.37}$$

gives the interval between two points labelled by x^μ and $x^\mu + \Delta x^\mu$.

In old-style notation, one often writes the metric in terms of an infinitesimal interval, or **line element**:

$$\mathrm{d}s^2 \;=\; g_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu$$

with the $\mathrm{d}x^\mu$ the components of an infinitesimal displacement. In modern notation, however, $\mathrm{d}x^\mu$ is interpreted quite differently, and one in fact identifies the bilinear form $\mathbf{d}\boldsymbol{s}^{\mathbf{2}}$ with $\mathfrak{g}$. The interval $\Delta s^2 =< \Delta\mathbf{x}, \Delta\mathbf{x} >$ is then the output of $\mathbf{d}\boldsymbol{s}^{\mathbf{2}}(\Delta\mathbf{x}, \Delta\mathbf{x}) = \mathfrak{g}(\Delta\mathbf{x}, \Delta\mathbf{x})$.

The metric, or line element, is said to define the geometry of the space. Two spaces of the same dimension may well have different geometries, eg. $\mathbb{R}^4$ can be endowed with a positive-definite ($\Delta s^2 > 0$) metric, but it would not be the metric of 4-dim spacetime.

Example 1.6. Consider the position vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ in $\mathbb{R}^3$, with components:

$$\mathbf{x}_1 \;\longmapsto\; (x_1, y_1, z_1)^{\mathrm{T}}\,, \qquad\qquad \mathbf{x}_2 \;\longmapsto\; (x_2, y_2, z_2)^{\mathrm{T}}$$

If we choose a positive-definite $\mathfrak{g}$ with matrix representation $\mathbf{g} = \mathbf{I}$:

$$\mathfrak{g}(\Delta\mathbf{x},\,\Delta\mathbf{x}) \;=\; g_{\mu\nu}\,\Delta\mathbf{x}^\mu\,\Delta\mathbf{x}^\nu \;=\; \begin{pmatrix} x_1 - x_2 & y_1 - y_2 & z_1 - z_2 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 - x_2 \\ y_1 - y_2 \\ z_1 - z_2 \end{pmatrix}$$

The result, $(\Delta s)^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 = (\Delta x)^2 + (\Delta y^2) + (\Delta z)^2$, is recognised to be the "Pythagorean" distance squared between two points: $|\mathbf{x}_1 - \mathbf{x}_2|^2$.

Example 1.7. In $\mathbb{R}^4$, let $\mathbf{x}_1$ and $\mathbf{x}_2$ be position vectors with components (ct_1, x_1, y_1, z_1) and (ct_2, x_2, y_2, z_2). Then take the *indefinite* $\mathfrak{g}$ with matrix representation:

$$g_{\mu\nu} \;=\; \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Therefore:

$$\mathfrak{g}(\Delta\mathbf{x},\,\Delta\mathbf{x}) \;=\; -c^2(t_1 - t_2)^2 + (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2$$

is the spacetime distance between two events in Einstein's Special Relativity, with c the speed of light. The covariant components of $\mathbf{x}_2$ are $(-ct_2, x_2, y_2, z_2)$.

Because $\mathfrak{g}$ is indefinite, there exist non-trivial null vectors (with zero norm) such that $\mathfrak{g}(\mathbf{x}, \mathbf{x}) = 0$.

Quite often, we will wish to work in bases other than coordinate bases. The formal properties of $\mathfrak{g}$ that we have reviewed still hold, but its *components* can be different, even in the same coordinate system:

Definition 1.24. A basis $\{\mathbf{e}_\mu\}$ such that $\mathfrak{g}(\mathbf{e}_\mu, \mathbf{e}_\nu) = \pm 1$ when $\mu = \nu$ and 0 otherwise is said to be **orthonormal**. A useful notation to distinguish it from a coordinate basis is $\{\mathbf{e}_{\hat{\mu}}\}$. This is an extension of the usual definition of orthonormality which admits only $+1$ and 0; it is useful in the case of indefinite metrics.

An orthonormal basis can be found *locally* (over a sufficiently small coordinate patch) for any metric, if necessary by using the Gram-Schmidt procedure, with some modifications if the metric is indefinite.

Definition 1.25. Let n_+ (n_-) denote the number of diagonal elements $\mathfrak{g}(\mathbf{e}_{\hat{\mu}}, \mathbf{e}_{\hat{\mu}})$ which are equal to $+1$ (-1). The **signature** of the metric is defined by $s = n_+ - n_-$. Since $n_+ + n_- = n$, the dimension of the space, we also have $s = n - 2n_-$, and $\det \mathbf{g} = (-1)^{n_-}$. n_+ and n_- are basis-independent, and therefore so is the signature.

The sign of the overall signature of an indefinite metric is arbitrary and must be set by convention, a source of much grief in relativity. Example 1.7 sets $s = +2$, a good choice of sign when the spatial indices 1, 2, and 3 are often raised or lowered. In the more general spacetimes of General Relativity, $s = -2$ is quite often (but not always...) used. Thus, beware!

Definition 1.26. A n-dim space endowed with a metric of signature $\pm n$ is called **Euclidean**. If $n_- = 1$ (or $n_- = n - 1$), the space is **pseudo-Euclidean**, aka **Minkowski** when $n = 4$.

Example 1.7 has a Minkowski metric in four-dimensional space.

Thanks to the metric, we recover the *vector* gradient of a function as defined in calculus. You may have noticed that throughout our discussion of manifolds and tangent spaces, no mention was made of an inner product on them, because none was needed—until now. An inner product, or metric $\mathfrak{g}$, pairs the 1-form $\mathbf{d}f$ with a vector, $\boldsymbol{\nabla} f$; indeed, from eq. (1.35):

$$< \boldsymbol{\nabla} f, \mathbf{v} > = \mathfrak{g}(\boldsymbol{\nabla} f, \mathbf{v}) = g_{\mu\nu} (\boldsymbol{\nabla} f)^\mu v^\nu = (\partial_\nu f)\, v^\nu \tag{1.38}$$

where $\mathbf{v}$ is an arbitrary vector. Therefore, the components of $\boldsymbol{\nabla} f$ in a coordinate basis are given by: $(\boldsymbol{\nabla} f)^\mu = g^{\mu\nu}\, \partial_\nu f$. Only in a Euclidean metric with a standard basis are the components of the vector gradient the same as those of the 1-form gardient.

Example 1.8. In a Minkowskian manifold with coordinates $(x^0 = ct, x^1, x^2, x^3)$, the matrix representing $\mathfrak{g}$ is (using the "mostly positive" sign convention) a diagonal matrix with elements $(-1, 1, 1, 1)$. Then:

$$\begin{aligned} \mathbf{d}f &= \partial_t f\, \mathbf{d}t + \partial_i f\, \mathbf{d}x^i \\ \boldsymbol{\nabla} f &= -\partial_{ct} f\, \boldsymbol{\partial}_{ct} + \partial^i f\, \boldsymbol{\partial}_i \qquad (\partial^i = g^{ij}\partial_j = \partial_i) \end{aligned}$$

There is something interesting about the determinant of the metric tensor which we find by writing the transformation law of the tensor as a matrix equation: $g'_{\mu\nu} = \partial_{\mu'} x^\alpha g_{\alpha\beta} \partial_{\nu'} x^\beta$, and taking the determinant. Defining $g = \det g_{\alpha\beta}$, we obtain:

$$g' = \left| \frac{\partial x}{\partial x'} \right|^2 g \tag{1.39}$$

where $|\partial x/\partial x'|$ is the **Jacobian** of the transformation matrix *from* x to x' coordinates. We discover that g does not not transform as a scalar, ie. it is not invariant.

Definition 1.27. A quantity that has extra powers of $|\partial x/\partial x'|$ as factors in its transformation law in addition to the usual $\partial_{\mu'} x^\alpha$ and/or $\partial_\alpha x^{\mu'}$ factors is called a **tensor density**. Thus, g is a scalar density.

What might seem no more than an exotic property becomes more relevant when we consider the n-dim volume element as usually written in an integral. This, as we know from calculus, transforms as : $\mathrm{d}^n x' = |\partial x'/\partial x| \mathrm{d}^n x$ (note the position of the prime in the Jacobian!), so is not an invariant. As a result, the volume integral of a scalar function is not an invariant, yet there should be no memory of the variables of integration left after integrating. Instead, transform $\sqrt{(-1)^{n_-} g}\, \mathrm{d}^n x$:

$$\sqrt{(-1)^{n_-} g'}\, \mathrm{d}^n x' = \left| \frac{\partial x}{\partial x'} \right| \sqrt{(-1)^{n_-} g} \left| \frac{\partial x'}{\partial x} \right| \mathrm{d}^n x = \sqrt{(-1)^{n_-} g}\, \mathrm{d}^n x$$

which is seen to be a scalar! Then integrals written as $\int \sqrt{|g|} f(\mathbf{x}) \mathrm{d}^n x$ *are* invariant, but the scalar function $f(\mathbf{x})$ we thought we were integrating is in fact the scalar density $\sqrt{|g|} f(\mathbf{x})$. This concept of tensor density as a *notational device* has been widely used in General Relativity, although post-1970 literature largely dispenses with it when p-forms are involved. Indeed, later in section 1.5.2, we shall introduce a deeper definition of the volume element.

1.5 Exterior Algebra

1.5.1 The exterior product

Definition 1.28. The **exterior** (**wedge**) product of two 1-forms is the antisymmetrised tensor product:

$$\boldsymbol{\sigma} \wedge \boldsymbol{\tau} \;\equiv\; \boldsymbol{\sigma} \otimes \boldsymbol{\tau} \;-\; \boldsymbol{\tau} \otimes \boldsymbol{\sigma}$$

Being skew-symmetric, the exterior product of two 1-forms can be called a **2-form**.

In general, p 1-forms can be used to construct a **simple** (or decomposable) p-form:

$$\boldsymbol{\sigma}^1 \wedge \ldots \wedge \boldsymbol{\sigma}^p \;=\; \delta^{1\,\ldots\,p}_{\mu_1\,\ldots\,\mu_p}\, \boldsymbol{\sigma}^{\mu_1} \otimes \cdots \otimes \boldsymbol{\sigma}^{\mu_p} \tag{1.40}$$

Needless to say, this expression can be directly applied to a basis of $\mathcal{V}^*$, $\{\mathbf{d}x^\mu\}$ ($\mu = 1, \ldots, p$):

$$\mathbf{d}x^{\rho_1} \wedge \cdots \wedge \mathbf{d}x^{\rho_p}(\boldsymbol{\partial}_{\mu_1}, \ldots, \boldsymbol{\partial}_{\mu_p}) \;=\; \delta^{\rho_1\,\ldots\,\rho_p}_{\nu_1\,\ldots\,\nu_p}\, \delta^{\nu_1}{}_{\mu_1} \cdots \delta^{\nu_p}{}_{\mu_p} \;=\; \delta^{\rho_1\,\ldots\,\rho_p}_{\mu_1\,\ldots\,\mu_p} \qquad (1 \le \rho_1 < \ldots < \rho_p \le n) \tag{1.41}$$

Thus, from $\{\mathbf{d}x^{\mu_i}\}$ ($1 \le i \le p$) a basis of $\bigwedge^p(\mathcal{V}^*)$ can be constructed which contains $n!/(p!(n-p)!)$ elements. In particular, a n-form on a n-dimensional space is a one-component object, in the sense that it must be a multiple of the *unique* basis element, $\mathbf{d}x^1 \wedge \mathbf{d}x^2 \wedge \ldots \wedge \mathbf{d}x^n$, with indices in increasing order.

The exterior product of a basis of $\bigwedge^p$ and a basis of $\bigwedge^q$ is a basis, $\mathbf{d}x^{\mu_1} \wedge \ldots \wedge \mathbf{d}x^{\mu_p} \wedge \mathbf{d}x^{\mu_{p+1}} \wedge \ldots \wedge \mathbf{d}x^{\mu_{p+q}}$, of $\bigwedge^{p+q}$, again with indices in increasing order.

By extension, a $(p+q)$-form can be constructed out of the antisymmetrised tensor product of $\boldsymbol{\sigma} \in \bigwedge^p$ and $\boldsymbol{\tau} \in \bigwedge^q$. In geometric and index notation:

$$\boldsymbol{\sigma} \wedge \boldsymbol{\tau}(\mathbf{u}_{\rho_1}, \ldots, \mathbf{u}_{\rho_{p+q}}) = \delta^{\mu_1\ldots\mu_p\nu_1\ldots\nu_q}_{\rho_1\ldots\rho_p\rho_{p+1}\ldots\rho_{p+q}}\, \boldsymbol{\sigma}(\mathbf{u}_{\mu_1} \ldots \mathbf{u}_{\mu_p})\, \boldsymbol{\tau}(\mathbf{u}_{\nu_1} \ldots \mathbf{u}_{\nu_q}) \qquad \mu_1 < \mu_2 \ldots < \mu_p, \quad \nu_1 < \ldots < \nu_q$$
$$(\boldsymbol{\sigma} \wedge \boldsymbol{\tau})_{\rho_1\ldots\rho_p\rho_{p+1}\ldots\rho_{p+q}} \;=\; \delta^{\mu_1\ldots\mu_p\nu_1\ldots\nu_q}_{\rho_1\ldots\rho_p\rho_{p+1}\ldots\rho_{p+q}}\, \sigma_{\mu_1\ldots\mu_p}\, \tau_{\nu_1\ldots\nu_q} \qquad \mu_1 < \mu_2 \ldots < \mu_p, \quad \nu_1 < \nu_2 \ldots < \nu_q \tag{1.42}$$

The wedge-product defines a so-called **exterior (Grassmann) algebra**.

If $p > n$, then at least two indices in the wedge product of the p basis 1-forms (or of the components of the p-form) must be identical. Therefore, skew-symmetry forces the maximum rank of a non-trivial p-form in n dimensions to be n.

The exterior product, in contrast to the vector ("cross") product of vector analysis which it generalises, is **associative**: $\boldsymbol{\sigma} \wedge (\boldsymbol{\tau} \wedge \boldsymbol{\theta}) = (\boldsymbol{\sigma} \wedge \boldsymbol{\tau}) \wedge \boldsymbol{\theta}$.

Another very important property of the exterior product of a p-form and a q form is that:

$$\boldsymbol{\sigma} \wedge \boldsymbol{\tau} \;=\; (-1)^{pq}\, \boldsymbol{\tau} \wedge \boldsymbol{\sigma} \tag{1.43}$$

This follows directly from eq. (1.42) by noting that it takes pq transpositions to get $\delta^{\nu_1\ldots\nu_q\mu_1\ldots\mu_p}_{\rho_1\ldots\rho_{p+q}}$ into $\delta^{\mu_1\ldots\mu_p\nu_1\ldots\nu_q}_{\rho_1\ldots\rho_{p+q}}$. This means that the exterior product commutes except when both forms have odd rank.

Eq. (1.42) is easier to use than it might seem. Here are three examples:

Example 1.9. Some people believe that we live in an 11-dimensional world. Let us work out one component of the 3-form that is the exterior product of a 2-form, $\boldsymbol{\sigma}$, and a 1-form, $\boldsymbol{\tau}$:

$$\begin{aligned}(\boldsymbol{\sigma} \wedge \boldsymbol{\tau})_{11\,36} \;&=\; \delta^{\mu\nu\lambda}_{11\,36}\, \sigma_{\mu\nu}\, \tau_\lambda \qquad \mu < \nu \\ &=\; \delta^{36\,11}_{11\,36}\, \sigma_{36}\tau_{11} \;+\; \delta^{3\,11\,6}_{11\,36}\, \sigma_{3\,11}\tau_6 \;+\; \delta^{6\,11\,3}_{11\,36}\, \sigma_{6\,11}\tau_3 \\ &=\; \sigma_{36}\tau_{11} \;-\; \sigma_{3\,11}\tau_6 \;+\; \sigma_{6\,11}\tau_3\end{aligned}$$

Example 1.10. In two dimensions, the exterior product of two 1-forms, $\boldsymbol{\sigma}^1$ and $\boldsymbol{\sigma}^2$, is:

$$\begin{aligned}\boldsymbol{\sigma}^1 \wedge \boldsymbol{\sigma}^2 &= (\sigma^1{}_1 \mathbf{d}x^1 + \sigma^1{}_2 \mathbf{d}x^2) \wedge (\sigma^2{}_1 \mathbf{d}x^1 + \sigma^2{}_2 \mathbf{d}x^2) \\ &= \sigma^1{}_1\sigma^2{}_2\, \mathbf{d}x^1 \wedge \mathbf{d}x^2 + \sigma^1{}_2\sigma^2{}_1\, \mathbf{d}x^2 \wedge \mathbf{d}x^1 = (\sigma^1{}_1\sigma^2{}_2 - \sigma^1{}_2\sigma^2{}_1)\, \mathbf{d}x^1 \wedge \mathbf{d}x^2 \\ &= \det \mathbf{S}\, \mathbf{d}x^1 \wedge \mathbf{d}x^2\end{aligned}$$

where $\mathbf{S}$ is the 2×2 matrix whose two rows are the components of $\boldsymbol{\sigma}^1$ and $\boldsymbol{\sigma}^2$, respectively.

Example 1.11. In three dimensions, $\{\mathbf{d}x^1 \wedge \mathbf{d}x^2,\ \mathbf{d}x^1 \wedge \mathbf{d}x^3,\ \mathbf{d}x^2 \wedge \mathbf{d}x^3\}$ forms a basis of the space of 2-forms, $\bigwedge^2(\mathcal{V})$. Therefore, the *most general* (not just simple) 2-form in a 3-dim dual space can be written as:

$$\boldsymbol{\sigma} = \sigma_{12}\, \mathbf{d}x^1 \wedge \mathbf{d}x^2 + \sigma_{23}\, \mathbf{d}x^2 \wedge \mathbf{d}x^3 + \sigma_{31}\, \mathbf{d}x^3 \wedge \mathbf{d}x^1 = \frac{1}{2}\, \sigma_{\mu\nu}\, \mathbf{d}x^\mu \wedge \mathbf{d}x^\nu \tag{1.44}$$

Now the summation on the right is unrestricted. In this last form, the dimension of $\mathcal{V}^*$ does not appear explicitly.

Three-dimensional *simple* 2-forms $\boldsymbol{\sigma}^1 \wedge \boldsymbol{\sigma}^2$, however, have the coordinate form (EXERCISE):

$$(\sigma^1{}_1\sigma^2{}_2 - \sigma^1{}_2\sigma^2{}_1)\, \mathbf{d}x^1 \wedge \mathbf{d}x^2 + (\sigma^1{}_3\sigma^2{}_1 - \sigma^1{}_1\sigma^2{}_3)\, \mathbf{d}x^3 \wedge \mathbf{d}x^1 + (\sigma^1{}_2\sigma^2{}_3 - \sigma^1{}_3\sigma^2{}_2)\, \mathbf{d}x^2 \wedge \mathbf{d}x^3 \tag{1.45}$$

In Euclidean $\mathbb{R}^3$ with Cartesian coordinates, the components would be those of the vector product of the two vectors associated with $\boldsymbol{\sigma}^1$ and $\boldsymbol{\sigma}^2$. It suggests that we could associate with a vector $\mathbf{v}$ not just a 1-form with components $g_{\mu\nu}v^\nu$, but a 2-form $\boldsymbol{\sigma}$ with *Cartesian* components $\sigma_{ij} \equiv \epsilon_{ijk}v^k$. $\bigwedge^{n-1}$ and $\bigwedge^1$ both have dimension n, which suggests that, eg., a simple 2-form $\boldsymbol{\sigma} \wedge \boldsymbol{\tau}$ might in some sense correspond to a 1-form. But the fact that the vector product is a *pseudo*-vector means that we are not quite ready for this; we still have to introduce another important idea, that of orientation, in section 1.5.4.

In four dimensions, the basis for $\bigwedge^2$ contains 6 elements. EXERCISE: What are the components of the exterior product of two 1-forms in three and four dimensions? (Given the basis elements $\mathbf{d}x^\mu \wedge \mathbf{d}x^\nu$, you should recognise that the components must be of the form (no pun intended) $\sigma^1_\mu\sigma^2_\nu - \sigma^2_\mu\sigma^1_\nu$.)

More generally, consider simple p–forms on a n-dimensional space. In terms of a basis $\{\mathbf{d}x^\nu\}$, we have for 1-forms $\boldsymbol{\sigma}^\mu$: $\boldsymbol{\sigma}^\mu = \sigma^\mu{}_\nu \mathbf{d}x^\nu$ (the superscripts on $\boldsymbol{\sigma}$ and $\mathbf{d}x$ being *labels* for the 1-forms). Thus:

$$\begin{aligned}\boldsymbol{\sigma}^1 \wedge \ldots \wedge \boldsymbol{\sigma}^p &= \sigma^1{}_{\nu_1} \mathbf{d}x^{\nu_1} \wedge \ldots \wedge \sigma^p{}_{\nu_p} \mathbf{d}x^{\nu_p} = \sigma^1{}_{\nu_1} \ldots \sigma^p{}_{\nu_p} \mathbf{d}x^{\nu_1} \wedge \ldots \wedge \mathbf{d}x^{\nu_p} \\ &= \left[\sigma^1{}_{\mu_1} \cdots \sigma^p{}_{\mu_p}\, \delta^{\mu_1 \ldots \mu_p}_{\nu_1 \ldots \nu_p}\right] \mathbf{d}x^{\nu_1} \wedge \ldots \wedge \mathbf{d}x^{\nu_p} \qquad \nu_1 < \ldots < \nu_p\end{aligned} \tag{1.46}$$

where the summation over each ν_i ($1 \le i \le p$) runs from 1 to n,

If we construct a $p \times n$ matrix $\mathbf{S}$ whose i^{th} row is the n components of the 1-form $\boldsymbol{\sigma}^i$, we may notice that the expression inside the square backets in eq. (1.46) is nothing but the determinant of the $p \times p$ submatrix extracted from *column* indices $\nu_1 \ldots \nu_p$ of $\mathbf{S}$, with $\nu_1 < \ldots < \nu_p$. Therefore, in eq. (1.46), each term in the sum over the ν indices has as coefficient a $p \times p$ determinant. Each row of a determinant contains p out of the n components of the 1-forms $\boldsymbol{\sigma}$, and these components, labelled by $\nu_1 < \ldots < \nu_p$, must be the same as the ones on the wedge product of basis covectors $\mathbf{d}x^{\nu_1} \wedge \ldots \wedge \mathbf{d}x^{\nu_p}$ in that term.

The output (a number!) resulting from inputting $\mathbf{u}_1, \ldots, \mathbf{u}_p$ into the simple p-form $\boldsymbol{\sigma}^1 \wedge \ldots \wedge \boldsymbol{\sigma}^p$ is:

$$\boldsymbol{\sigma}^1 \wedge \ldots \wedge \boldsymbol{\sigma}^p(\mathbf{u}_1, \ldots, \mathbf{u}_p) = \det\left[\boldsymbol{\sigma}^i(\mathbf{u}_j)\right] \tag{1.47}$$

ie. the determinant of the $p \times p$ matrix $\mathbf{S}$ whose entries are $S^i{}_j = \boldsymbol{\sigma}^i(\mathbf{u}_j)$. As we have seen, these are simply $\sigma^i_\mu u^\mu_j$, with μ running from 1 to n.

Example 1.12. For a 3-dim $\mathcal{V}^*$ of which the 2-forms $\mathbf{d}x^i \wedge \mathbf{d}x^j$ are basis elements, we have:

$$\mathbf{d}x^i \wedge \mathbf{d}x^j(\mathbf{u},\, \mathbf{v}) \;=\; \mathbf{d}x^i(\mathbf{u})\,\mathbf{d}x^j(\mathbf{v}) \;-\; \mathbf{d}x^j(\mathbf{u})\,\mathbf{d}x^i(\mathbf{v}) \;=\; \begin{vmatrix} u^i & u^j \\ v^j & v^j \end{vmatrix}$$

In $\mathbb{R}^n$ with Cartesian coordinates, we interpret this (up to a sign—see 1.5.2 below!) as the area of the parallelogram whose defining sides are the projections of $\mathbf{u}$ and $\mathbf{v}$ on the x^i-x^j plane.

When $p = n$, $\boldsymbol{\sigma}^1 \wedge \ldots \wedge \boldsymbol{\sigma}^n$ is a *one-component* object, because $\bigwedge^n(\mathcal{V}^*)$ has dimension 1. We can also see that the summation in the second line in eq. (1.46) can have only one term with the ν_i indices increasing from 1 to n:

$$\boldsymbol{\sigma}^1 \wedge \ldots \wedge \boldsymbol{\sigma}^n \;=\; \left[\sigma^1{}_{\mu_1} \cdots \sigma^n{}_{\mu_n}\, \delta^{\mu_1 \ldots \mu_n}_{1 \ldots n}\right] \mathbf{d}x^1 \wedge \ldots \wedge \mathbf{d}x^n \;=\; \det \mathsf{S}\, \mathbf{d}x^1 \wedge \ldots \wedge \mathbf{d}x^n \tag{1.48}$$

where the matrix S is now $n \times n$. Note that the indices $1 \ldots n$ are not summed over.

Example 1.13. There is another useful definition of the permutation symbol, $\delta^{\nu_1 \ldots \nu_n}_{\mu_1 \ldots \mu_n}$, equivalent to the one given by eq. (1.31), and which follows directly from (1.41):

$$\delta^{\nu_1 \ldots \nu_n}_{\mu_1 \ldots \mu_n} \;=\; \mathbf{d}x^{\nu_1} \wedge \ldots \wedge \mathbf{d}x^{\nu_n}(\boldsymbol{\partial}_{\mu_1},\, \ldots,\, \boldsymbol{\partial}_{\mu_n})$$

Then eq. (1.47) becomes:

$$\delta^{\nu_1 \ldots \nu_n}_{\mu_1 \ldots \mu_n} \;=\; \begin{vmatrix} \delta^{\nu_1}{}_{\mu_1} & \cdots & \delta^{\nu_1}{}_{\mu_n} \\ \vdots & & \vdots \\ \delta^{\nu_n}{}_{\mu_1} & \cdots & \delta^{\nu_n}{}_{\mu_n} \end{vmatrix} \tag{1.49}$$

Another application provides an easy test for the linear independence of p 1-forms: if their exterior product is different from zero, those p 1-forms are linearly independent. If they were not, one of them at least could be written as a linear combination of the others and antisymmetry would force the exterior product to vanish. Conversely, if the p 1-forms are linearly independent, $\boldsymbol{\sigma}^1 \wedge \cdots \wedge \boldsymbol{\sigma}^p$ icannot vanish.

Example 1.14. In general a p-form $\mathbf{F}$ is not simple. If it is, then the $2p$-form $\mathbf{F} \wedge \mathbf{F}$ must vanish by antisymmetry. But the converse can also hold: Let $\mathbf{F} \wedge \mathbf{F} = 0$, and take $\mathbf{F} \in \bigwedge^2$ in what follows. It can always be written as a simple form. For $n = 2$, $\mathbf{F} = F_{12}\mathbf{d}x^1 \wedge \mathbf{d}x^2$ is by inspection always simple, with no need to invoke $\mathbf{F} \wedge \mathbf{F} = 0$, which also always holds when $n = p = 2$.

For $n = 3$, $\mathbf{F} \wedge \mathbf{F}$ still trivially vanishes, and this time we need that fact to show that $\mathbf{F}$ is simple. We can always write:

$$\mathbf{F} \;=\; F_{12}\mathbf{d}x^1 \wedge \mathbf{d}x^2 \;+\; F_{13}\mathbf{d}x^1 \wedge \mathbf{d}x^3 \;+\; F_{23}\mathbf{d}x^2 \wedge \mathbf{d}x^3 \;\equiv \boldsymbol{\sigma} \;+\; \boldsymbol{\tau} \wedge \mathbf{d}x^3$$

where $\boldsymbol{\sigma}$ is a 2-form on a $n = 2$ subspace of the 3-dim space, and $\boldsymbol{\tau}$ is a 1-form on the same 2-dim subspace. Therefore, $\boldsymbol{\sigma}$ is simple. ie., $\boldsymbol{\sigma} = \boldsymbol{\alpha} \wedge \boldsymbol{\beta}$ ($\boldsymbol{\alpha},\ \boldsymbol{\beta} \in \bigwedge^1$) because $n = 2$.

Then $\mathbf{F} \wedge \mathbf{F} \;=\; \boldsymbol{\sigma} \wedge \boldsymbol{\sigma} + 2\boldsymbol{\sigma} \wedge \boldsymbol{\tau} \wedge \mathbf{d}x^3 \;=\; 0$. Also, $\boldsymbol{\sigma} \wedge \boldsymbol{\sigma} \;=\; 0$ because $\boldsymbol{\sigma}$ is simple. Then $\boldsymbol{\alpha} \wedge \boldsymbol{\beta} \wedge \boldsymbol{\tau} \wedge \mathbf{d}x^3 = 0$. But the only possible linear dependence between the four 1-forms is between $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ since none of them depends on $\mathbf{d}x^3$. Therefore, $\lambda_1\boldsymbol{\alpha} + \lambda_2\boldsymbol{\beta} + \lambda_3\boldsymbol{\tau} = 0$.

If $\lambda_3 = 0$, then $\boldsymbol{\beta}$ is a multiple of $\boldsymbol{\alpha}$, so $\boldsymbol{\sigma} = 0$, leaving $\mathbf{F} = \boldsymbol{\tau} \wedge \mathbf{d}x^3$, a simple form. If $\lambda_3 \neq 0$, $\boldsymbol{\tau} = a\boldsymbol{\alpha} + b\boldsymbol{\beta}$, and we obtain:

$$\mathbf{F} \;=\; \boldsymbol{\alpha} \wedge \boldsymbol{\beta} + (a\boldsymbol{\alpha} + b\boldsymbol{\beta}) \wedge \mathbf{d}x^3 \;=\; (\boldsymbol{\alpha} + \frac{b}{a}\boldsymbol{\beta}) \wedge (\boldsymbol{\beta} + a\mathbf{d}x^3)$$

which is a simple form. Thus, 2-forms on 3-dim space are always simple. EXERCISE: A 2-form in $n = 4$ is simple if, and only if, $\mathbf{F} \wedge \mathbf{F} = 0$. Also, when $\mathbf{F} \wedge \mathbf{F} \neq 0$ in $n = 4$, $\mathbf{F}$ can be written (EXERCISE) as the sum of two simple 2-forms.

1.5.2 Oriented manifolds, pseudo-vectors, pseudo-forms and the volume form

Definition 1.29. Two bases are said to have the same (opposite) **orientation** if the determinant of the matrix of the transformation between the two bases is positive (negative). Therefore, bases fall into two classes, or orientations. Orienting a manifold then means *arbitrarily* specifying one of the two orientations to be positive (**right-handed**), and the other negative (**left-handed**). Manifolds on which transport of a basis around some closed loop leads to a reversal of orientation are said to be **non-orientable** (eg. the Möbius strip).

In $\mathbb{R}^3$, for instance, $\mathbf{e}_x \wedge \mathbf{e}_y \wedge \mathbf{e}_z$, $\mathbf{e}_y \wedge \mathbf{e}_z \wedge \mathbf{e}_x$ and $\mathbf{e}_z \wedge \mathbf{e}_x \wedge \mathbf{e}_y$ can be transformed into one another by matrices of determinant $+1$. By convention, they are all taken to be right-handed. But $\mathbf{e}_y \wedge \mathbf{e}_x \wedge \mathbf{e}_z = -\mathbf{e}_x \wedge \mathbf{e}_y \wedge \mathbf{e}_z$ cannot be reached from $\mathbf{e}_x \wedge \mathbf{e}_y \wedge \mathbf{e}_z$ by a transformation with a matrix of positive determinant: it is an element of a left-handed basis.

Definition 1.30. An object that behaves in all respects as a vector or a p-form, except that its sign is reversed under a reversal of orientation of the manifold, is called a **pseudovector** or a **pseudoform**.

Example 1.15. Generalising example 1.12 above, the simple n-form $\mathbf{d}x^{\mu_1} \wedge \cdots \wedge \mathbf{d}x^{\mu_n}$, when acting on the vectors $\mathbf{v}_1, \ldots \mathbf{v}_n$ *in that order*, outputs a mumber of magnitude equal to the volume of the parallelopiped whose edges are the vectors $\mathbf{v}_1, \ldots \mathbf{v}_n$. Using eq. (1.48), this is readily computed as the determinant of all the vector components. But there is also a sign involved, with $+$ corresponding to the orientation defined by the vectors being the same as the orientation of the basis in which the vectors are written. We then say that this volume is oriented, in the sense that it changes sign under interchange of any two basis vectors, and we recognise it as a pseudoform.

Definition 1.31. In general coordinates u^i on a n-dim manifold, we define the **volume pseudoform**:

$$\mathbf{d}^n u \;\equiv\; \left|\frac{\partial x}{\partial u}\right| \mathbf{d}u^1 \wedge \cdots \wedge \mathbf{d}u^n \;=\; \sqrt{|g|}\, \mathbf{d}u^1 \wedge \cdots \wedge \mathbf{d}u^n$$

where the x^i form an *orthonormal* basis, usually Cartesian, and we have used eq. (1.39) with $|g| = 1$ for orthonormal bases. Both bases are connected by a transformation that preserves orientation.

1.5.3 The Levi-Civita pseudotensor

We have already remarked on the fact that the Levi-Civita *symbol* does not transform as a tensor. Consider, however, the volume pseudoform of definition 1.31. By inspection it is a n-form with the single independent component $(\mathbf{d}^n u)_{1\ldots n} = \sqrt{|g|}$. The other components are obtained by antisymmetrising with the Levi-Civita symbol, which we shall now denote by $[\mu_1 \ldots \mu_n]$ to avoid any confusion later. That is:

$$(\mathbf{d}^n u)_{\mu_1 \ldots \mu_n} \;=\; \sqrt{|g|}\, [\mu_1 \ldots \mu_n]$$

The objects on the right are themselves the components of a covariant pseudotensor, $\boldsymbol{\epsilon}$, of rank n. Henceforth, whenever we write components $\epsilon_{\mu_1 \ldots \mu_n}$, they are to be understood as $\sqrt{|g|}\, [\mu_1 \ldots \mu_n]$, so that $\epsilon_{1\ldots n} = \sqrt{|g|}$.

We obtain $\epsilon^{1 \ldots n}$ by raising the n indices of $\epsilon_{1 \ldots n}$ with $\mathfrak{g}$. In general coordinates:

$$\epsilon^{1\ldots n} = g^{1\mu_1} \cdots g^{n\mu_n} \sqrt{|g|}, \delta^{1 \ldots n}_{\mu_1 \ldots \mu_n} [1 \ldots n] \;=\; \det g^{\alpha\beta} \sqrt{|g|} \;=\; \frac{1}{(-1)^{n-}|g|}\sqrt{|g|} \;=\; \frac{(-1)^{n-}}{\sqrt{|g|}}$$

In *orthonormal* bases, this is simply: $\epsilon^{1 \ldots n} = (-1)^{n-} \epsilon_{1 \ldots n}$.

Both $\epsilon^{\nu_1 \ldots \nu_n}$ and $\epsilon_{\mu_1 \ldots \mu_n}$ being antisymmetric, we can relate the permutation symbol to the Levi-Civita pseudotensor:

$$\delta^{\nu_1 \ldots \nu_n}_{\mu_1 \ldots \mu_n} \;=\; a\, \epsilon^{\nu_1 \ldots \nu_n}\, \epsilon_{\mu_1 \ldots \mu_n}$$

To determine a, we note that $\epsilon^{1\,\dots\,n}\epsilon_{1\,\dots\,n} = (-1)^{n_-}$. The relation is then:

$$\epsilon^{\nu_1\,\dots\,\nu_n}\,\epsilon_{\mu_1\,\dots\,\mu_n} \;=\; (-1)^{n_-}\,\delta^{\nu_1\dots\nu_n}_{\mu_1\dots\mu_n} \;=\; (-1)^{n_-}\begin{vmatrix} \delta^{\nu_1}{}_{\mu_1} & \cdots & \delta^{\nu_1}{}_{\mu_n} \\ \vdots & & \vdots \\ \delta^{\nu_n}{}_{\mu_1} & \cdots & \delta^{\nu_n}{}_{\mu_n} \end{vmatrix} \tag{1.50}$$

For instance, in a Euclidean 3-dim space, $n_- = 0$, and the expanded product has six terms. When contracted over the last or first indices, we obtain (EXERCISE): $\epsilon^{ijk}\epsilon_{lnk} \;=\; \delta^i{}_l\delta^j{}_n \;-\; \delta^j{}_l\delta^i{}_n$. Other expressions for the product of Levi-Civita tensors in a 4-dim Minkowski space can be found in MTW, pp. 87-88.

1.5.4 The Hodge dual of a p-form

In section 1.5.1 we had pointed out that it might be possible to associate a 2-form with a vector. Now we can do this in all generality on n-dim spaces endowed with a metric structure. To the vector $\mathbf{v}$ corresponds a pseudoform of rank n-1:

$$\boldsymbol{\sigma} \;=\; v^\nu\,\epsilon_{\nu\mu_1\dots\mu_{n-1}}\,\mathbf{d}u^{\mu_1}\wedge\ldots\wedge\mathbf{d}u^{\mu_{n-1}} \qquad \mu_1<\ldots<\mu_{n-1} \tag{1.51}$$

which, like $\mathbf{v}$, has n (independent!) components. In 3-dim $\mathbb{R}^3$ we have the pseudo-2-form:

$$\boldsymbol{\sigma} \;=\; \sqrt{|g|}\,(v^3\,\mathbf{d}u^1\wedge\mathbf{d}u^2 \;+\; v^2\,\mathbf{d}u^3\wedge\mathbf{d}u^1 \;+\; v^1\,\mathbf{d}u^2\wedge\mathbf{d}u^3)$$

And there must also be a mapping between the 1-form dual to $\mathbf{v}$ and the (n-1)-form. We now generalise to a linear mapping between $\bigwedge^p$ and $\bigwedge^{n-p}$.

Definition 1.32. Let $\mathcal{V}^n$ be endowed with a metric and a basis $\{\boldsymbol{\partial}_\mu\}$. With $\boldsymbol{\epsilon}$ the Levi-Civita *pseudo-tensor*, the **Hodge dual**[†] maps a p-form $\boldsymbol{\sigma}$ to a (n-p)-form ${}^*\boldsymbol{\sigma}$ with components:

$$({}^*\boldsymbol{\sigma})_{\nu_1\dots\nu_{n-p}} \;=\; \frac{1}{p!}\,\sigma_{\mu_1\dots\mu_p}\,\epsilon^{\mu_1\dots\mu_p}{}_{\nu_1\dots\nu_{n-p}} \qquad \Big(\;=\;\sigma^{\mu_1\dots\mu_p}\,\epsilon_{\mu_1\dots\mu_p\nu_1\dots\nu_{n-p}} \qquad \mu_1<\ldots<\mu_p\Big) \tag{1.52}$$

from which it is obvious that the Hodge dual of a p-form is a *pseudo*-form, and vice-versa. It can be shown that, given a metric $\mathfrak{g}$, ${}^{**}\boldsymbol{\sigma} = (-1)^{n_-}(-1)^{p(n-p)}\boldsymbol{\sigma}$. So the mapping is idempotent in Euclidean spaces ($n_- = 0$) of odd dimension, such as $\mathbb{R}^3$. In a 4-dim Minkowski space ($n_- = 1$), Hodge duality is idempotent only on 1- and 3-forms.

One immediate application of eq. (1.52) is that the n-dim volume form is the Hodge dual of the 0-form 1:

$${}^*1 \;=\; \epsilon_{\mu_1\dots\mu_n}\,\mathbf{d}u^{\mu_1}\wedge\cdots\wedge\mathbf{d}u^{\mu_n} \qquad (\mu_1<\ldots<\mu_n)$$

A very important consequence of the fact that ${}^{**}\boldsymbol{\sigma} = \pm\boldsymbol{\sigma}$ is that a p-form and its Hodge dual contain *exactly the same information*! Thus, "dualising" a p-form (or an antisymmetric contravariant tensor) can remove some (or all!) the redundancy due to anrisymmetry while preserving its information. For instance, in 4-dim Minkowski space, one might be a bit intimidated by a 4-form with components $\sigma_{\mu\nu\lambda\rho}$ until one realises that it is dual to a pseudo-0-form, so one independent number instead of $4^4 = 256$. Or a 3-form with *a priori* $4^3 = 64$ components can be Hodge-dualised, that is, written in terms of only the four components of its dual pseudo-1-form, which are (up to $\sqrt{|g|}$) the only independent components of the 3-form.

Example 1.16. If $\mathbf{T}$ is a $(2,0)$ skew-symmetric tensor:

$$({}^*T)_\lambda \;=\frac{1}{2}\epsilon_{\mu\nu\lambda}\,T^{\mu\nu} \qquad \text{in Euclidean } \mathbb{R}^3$$

$$({}^*T)_{\lambda\rho} \;=\frac{1}{2}\epsilon_{\mu\nu\lambda\rho}\,T^{\mu\nu} \qquad \text{Minkowski } \mathbb{R}^4$$

[†]Here, the meaning of "dual" has no relation to its other use in "dual" space or basis.

In the first line,with $\sqrt{|g|} = 1$, it is not hard to work out that ${}^*T_1 = T^{23}$, ${}^*T_2 = T^{31}$, and ${}^*T_3 = T^{12}$, so that the tensor (1-form) dual to $\mathbf{T}$ contains only the three independent components of the original tensor.

In a previous example, we had noticed something special about the 3-dim Euclidean space of p-forms: both $\{\mathbf{d}x^1, \mathbf{d}x^2, \mathbf{d}x^3\}$ and $\{\mathbf{d}x^2 \wedge \mathbf{d}x^3,\ \mathbf{d}x^3 \wedge \mathbf{d}x^1,\ \mathbf{d}x^1 \wedge \mathbf{d}x^2\}$ are bases, the first being used to generate 1-forms and the second 2-forms. Now we know that the two bases are each other's Hodge dual. In fact, instead of working on the components, an equivalent way of Hodge-dualising is to work on the (co)basis: ${}^*(\mathbf{d}u^{\mu_1} \wedge \cdots \wedge \mathbf{d}u^{\mu_p}) = \epsilon^{\mu_1 \cdots \mu_p}{}_{\mu_{p+1} \cdots \mu_n}\, \mathbf{d}u^{\mu_{p+1}} \wedge \cdots \wedge \mathbf{d}u^{\mu_n}$, with $\mu_{p+1} < \cdots < \mu_n$ (or dividing by $(n-p)!$ if the summations are unrestricted). There are corresponding expressions for Hodge-dualising coordinate bases or the components of contravariant tensors, ie., the Hodge dual of a contravariant tensor (or its components) also exists, as the above example shows.

Example 1.17. If $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ are 3-dim 1-forms, the 2-form: $\boldsymbol{\sigma} \wedge \boldsymbol{\tau} = (\sigma_2\tau_3 - \sigma_3\tau_2)\,\mathbf{d}x^2 \wedge \mathbf{d}x^3 + (\sigma_3\tau_1 - \sigma_1\tau_3)\,\mathbf{d}x^3 \wedge \mathbf{d}x^1 + (\sigma_1\tau_2 - \sigma_2\tau_1)\mathbf{d}x^1 \wedge \mathbf{d}x^2$ has as its Hodge dual on a space with metric $\mathfrak{g}$ the pseudo-1-form:

$$ {}^*(\boldsymbol{\sigma} \wedge \boldsymbol{\tau}) = \sqrt{|g|}\left[(\sigma_2\tau_3 - \sigma_3\tau_2)\,\mathbf{d}x^1 + (\sigma_3\tau_1 - \sigma_1\tau_3)\,\mathbf{d}x^2 + (\sigma_1\tau_2 - \sigma_2\tau_1)\,\mathbf{d}x^3\right] $$

If $\boldsymbol{\sigma}$ corresponds to the vector $\mathbf{u}$ and $\boldsymbol{\tau}$ to $\mathbf{v}$ via the metric, this is the same as saying that ${}^*(\mathbf{u} \wedge \mathbf{v}) = \mathbf{u} \times \mathbf{v}$, or, according to eq. (1.52), $(\mathbf{u} \times \mathbf{v})^\mu = \frac{1}{2} g^{\mu\rho}\epsilon_{\nu\lambda\rho}(u^\nu v^\lambda - v^\lambda u^\nu) = g^{\mu\rho}\epsilon_{\rho\nu\lambda}u^\nu v^\lambda$. In other words, when calculating a vector product, one is implicitly taking a Hodge dual; this is the only way that the result can be a pseudo-vector.

It is easy to recover all the relations of vector analysis in Cartesian $\mathbb{R}^3$. For instance:

$$ \begin{aligned} \mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}) &= \epsilon_{\mu\nu\rho} u^\mu v^\nu w^\rho \\ &= w^\rho \epsilon_{\rho\mu\nu} u^\mu v^\nu \qquad \text{(cyclic permutation of indices on } \epsilon) \\ &= \mathbf{w} \cdot (\mathbf{u} \times \mathbf{v}). \end{aligned} $$

Why not try your hand at a few if you have never done it, if only to practice index manipulation.

1.6 Tensor Fields and how to Describe their Change

Definition 1.33. A **tensor field** $\mathbf{T}(p)$ on a manifold M^n is a function of points $p \in M^n$ whose components $T(p)$ are real-valued differentiable functions of coordinates on the manifold.

Examples: the coordinate vector field $\boldsymbol{\partial}_\mu$, the gravitational and electric fields, the metric tensor with components $<\boldsymbol{\partial}_\mu, \boldsymbol{\partial}_\mu>$ in a coordinate basis.

The components of a $(r,\ s)$ tensor field with respect to the coordinates x^μ in n dimensions are the n^{r+s} real-valued *functions* $\mathbf{T}(\mathbf{d}x^{\nu_1}, \ldots, \mathbf{d}x^{\nu_r}, \boldsymbol{\partial}_{\mu_1}, \ldots, \boldsymbol{\partial}_{\mu_s})$.

How do we describe the change of a tensor field at a point? More precisely, how do we differentiate it? We already know from section 1.2 how to take the directional derivative of a $(0,0)$ tensor, ie. a function. On a "flat" (without curvature) manifold, directional derivatives of tensor components, which are functions, can be calculated in the same way.

For general (r,s) tensors, however, because of the presence of bases, defining differentiation requires extra structure, called a **connection**, or **covariant derivative**. Raising this important issue is like opening Pandora's box (aka can of worms), because there are a lot of ways to construct such a connection. A few, however, have gained favour as "natural". We shall discuss a particular type of differentiation that offers a neat unification of the ideas of gradient, divergence and curl in vector calculus, and for which a connection is actually not needed.

1.6.1 Exterior derivative

We introduce the **exterior derivative operator**[†], $\mathbf{d}$, which acts on p-forms $\boldsymbol{\sigma} = \sigma_{|\mu_1\ldots\mu_p|}\mathbf{d}x^{\mu_1}\wedge\ldots\wedge\mathbf{d}x^{\mu_p}$ ($|\mu_1\ldots\mu_p| \equiv \mu_1 < \ldots < \mu_p$) defined over some manifold M^n to give $p+1$-forms, also defined on M^n. Let $\boldsymbol{\sigma}$ be a p-form and $\boldsymbol{\tau}$ a q-form. The operator satisfies the following properties:

(a) $\mathbf{d}\boldsymbol{\sigma} = \partial_{\mu_0}\sigma_{|\mu_1\ldots\mu_p|}\,\mathbf{d}x^{\mu_0}\wedge\mathbf{d}x^{\mu_1}\wedge\ldots\wedge\mathbf{d}x^{\mu_p} = \big(\mathbf{d}\sigma_{|\mu_1\ldots\mu_p|}\big)\wedge\mathbf{d}x^{\mu_1}\wedge\ldots\wedge\mathbf{d}x^{\mu_p}$
In terms of components:

$$(\mathbf{d}\boldsymbol{\sigma})_{|\mu_1\ldots\mu_{p+1}|} = \delta^{\nu_0\nu_1\ldots\nu_p}_{\mu_1\ldots\mu_{p+1}}\partial_{\nu_0}\sigma_{|\nu_1\ldots\nu_p|} = \partial_{\mu_1}\sigma_{\mu_2\ldots\mu_{p+1}} - \partial_{\mu_2}\sigma_{\mu_1\ldots\mu_{p+1}} + \ldots$$

(b) $\mathbf{d}(\boldsymbol{\sigma}+\boldsymbol{\tau}) = \mathbf{d}\boldsymbol{\sigma} + \mathbf{d}\boldsymbol{\tau}$ ($p=q$).

(c) If $\boldsymbol{\sigma}$ is a 0-form, ie. just a function, then $\mathbf{d}\boldsymbol{\sigma}$ is the 1-form gradient of that function.

(d) $\mathbf{d}(\boldsymbol{\sigma}\wedge\boldsymbol{\tau}) = \mathbf{d}\boldsymbol{\sigma}\wedge\boldsymbol{\tau} + (-1)^p\,\boldsymbol{\sigma}\wedge\mathbf{d}\boldsymbol{\tau}$ (aka the antiderivation property of $\mathbf{d}$ with respect to the exterior product).

(e) $\mathbf{d}^2\boldsymbol{\sigma} = 0$ (**Poincaré's lemma**).

We shall not prove the antiderivation property (you can do it as an EXERCISE), but Poincaré's lemma is so famous and important that it deserves some proof.

First, for an arbitrary function f (0-form):

$$\mathbf{d}^2 f = \mathbf{d}(\partial_\nu f\,\mathbf{d}x^\nu) = \partial_\mu\partial_\nu f\,\mathbf{d}x^\mu\wedge\mathbf{d}x^\nu = 0$$

since $\partial_\mu\partial_\nu$ is symmetric in μ and ν. If g is another function, $\mathbf{d}(\mathbf{d}f\wedge\mathbf{d}g) = \mathbf{d}^2 f\wedge\mathbf{d}g - \mathbf{d}f\wedge\mathbf{d}^2 g$ from the antiderivation property; this must vanish since $\mathbf{d}^2$ applied to any function gives zero. By extension, the exterior derivative of the exterior product of any number of arbitrary differential 1-forms also vanishes. Now, from properties (a) and (d) above:

$$\begin{aligned}\mathbf{d}^2\boldsymbol{\sigma} &= \mathbf{d}\big[\mathbf{d}\sigma_{|\mu_1\ldots\mu_p|}\wedge\mathbf{d}x^{\mu_1}\wedge\ldots\wedge\mathbf{d}x^{\mu_p}\big]\\ &= \mathbf{d}^2\big[\sigma_{|\mu_1\ldots\mu_p|}\big]\wedge\mathbf{d}x^{\mu_1}\wedge\ldots\wedge\mathbf{d}x^{\mu_p} + (-1)^{\mu_p+1}\mathbf{d}\sigma_{|\mu_1\ldots\mu_p|}\wedge\mathbf{d}\big[\mathbf{d}x^{\mu_1}\wedge\ldots\wedge\mathbf{d}x^{\mu_p}\big]\end{aligned}$$

The first term vanishes because we have already shown that $\mathbf{d}$ is nilpotent on functions; the second vanishes because we have also shown that the exterior derivative of a wedge product of differential 1-forms vanishes.

Example 1.18. As an example of the first property of $\mathbf{d}$, the exterior derivative of a 1-form $\boldsymbol{\sigma}$ is the 2-form:

$$\begin{aligned}\boldsymbol{\theta} = \mathbf{d}\boldsymbol{\sigma} &= \partial_\mu\sigma_\nu\,\mathbf{d}x^\mu\wedge\mathbf{d}x^\nu = \mathbf{d}\sigma_\nu\wedge\mathbf{d}x^\nu\\ \theta_{\mu\nu} &= \partial_\mu\sigma_\nu - \partial_\nu\sigma_\mu\end{aligned} \tag{1.53}$$

where we have used the fact that the exterior derivative of the 0-forms (functions) σ_ν are 1-forms. (EXERCISE: what would be the exterior derivative of a 2-form? what would its components be?)

[†]Some authors prefer the notation $\boldsymbol{\nabla}\wedge$ for the exterior derivative.

Example 1.19. In $\mathbb{R}^3$, with u, v, and w as *arbitrary* coordinates, the differential of a function f in the *coordinate* basis $\{\mathbf{d}u,\, \mathbf{d}v,\, \mathbf{d}w\}$ is the 1-form:

$$\mathbf{d}f \;=\; \partial_u f\, \mathbf{d}u \,+\, \partial_v f\, \mathbf{d}v \,+\, \partial_w f\, \mathbf{d}w \tag{1.54}$$

Only if the coordinates are Cartesian is this also valid for an *orthonormal* basis in the same coordinates. In a spherical coordinate basis $\{\mathbf{d}r,\, \mathbf{d}\,\theta,\, \mathbf{d}\,\phi\}$, for instance, $\mathbf{d}f$ would keep the above simple form. If instead we insist on using a basis whose components are normalised to unity, such as $\{\mathbf{d}\hat{r},\, \mathbf{d}\,\hat{\theta},\, \mathbf{d}\,\hat{\phi}\} = \{\mathbf{d}r,\, r\mathbf{d}\,\theta,\, r\,\sin\theta\,\mathbf{d}\,\phi\}$ — as is almost always the case in vector analysis applied to physics — consistency demands that we write:

$$\mathbf{d}f \;=\; \partial_r f\, \mathbf{d}\hat{r} \,+\, \frac{1}{r}\,\partial_\theta f\, \mathbf{d}\hat{\theta} \,+\, \frac{1}{r\sin\theta}\,\partial_\phi f\, \mathbf{d}\hat{\phi} \tag{1.55}$$

Example 1.20. What about the exterior derivative of a 1-form $\boldsymbol{\sigma} = \sigma_u \mathbf{d}u + \sigma_v \mathbf{d}v + \sigma_w \mathbf{d}w$ in $\mathbb{R}^3$? We get:

$$\begin{aligned}\mathbf{d}\boldsymbol{\sigma} \;&=\; \big(\partial_v\sigma_u\, \mathbf{d}v \,+\, \partial_w\sigma_u\, \mathbf{d}w\big) \,\wedge \mathbf{d}u \,+\, \big(\partial_u\sigma_v\, \mathbf{d}u \,+\, \partial_w\sigma_v\, \mathbf{d}w\big) \,\wedge \mathbf{d}v \,+\, \big(\partial_u\sigma_w\, \mathbf{d}u \,+\, \partial_v\sigma_w\, \mathbf{d}v\big) \,\wedge \mathbf{d}w \\ &=\; (\partial_v\sigma_w \,-\, \partial_w\sigma_v)\, \mathbf{d}v\wedge\mathbf{d}w \,+\, (\partial_w\sigma_u \,-\, \partial_u\sigma_w)\, \mathbf{d}w\wedge\mathbf{d}u \,+\, (\partial_u\sigma_v \,-\, \partial_v\sigma_u)\, \mathbf{d}u\wedge\mathbf{d}v\end{aligned} \tag{1.56}$$

Taking the Hodge dual gives the pseudo-1-form:

$${}^*\mathbf{d}\boldsymbol{\sigma} \;=\; \sqrt{|g|}\,\big[(\partial_v\sigma_w \,-\, \partial_w\sigma_v)\, \mathbf{d}u \,+\, (\partial_w\sigma_u \,-\, \partial_u\sigma_w)\, \mathbf{d}v \,+\, (\partial_u\sigma_v \,-\, \partial_v\sigma_u)\, \mathbf{d}w\big] \tag{1.57}$$

By analogy with tensor algebra results, we recover the covariant version of the components of the 3-dim curl of a vector, but *only in Cartesian coordinates*! Only in those coordinates is $\sqrt{|g|} = 1$, with covariant and contravariant components the same in Cartesian $\mathbb{R}^n$.

As we know all too well, the *vector* components of the curl of a *vector* in curvilinear coordinates can be quite complicated, but this complication is largely due to our insisting on working with objects which are less natural. Indeed, when a curl is called for, it always is the curl of a 1-form, and there is little need for its contravariant components which are sure to be more complicated because of the curvilinear metric being used to raise indices.

It is interesting that, in vector calculus with Cartesian coordinates, we could write $\boldsymbol{\sigma}$ as $\mathbf{A}\cdot\mathrm{d}\mathbf{x}$, with $\mathbf{A}$ a corresponding vector. Then the right-hand side of eq. (1.56) would correspond to $\boldsymbol{\nabla}\times\mathbf{A}\cdot\mathrm{d}\mathbf{S}$, where $\mathrm{d}\mathbf{S}$ is a surface element with Cartesian components $\mathbf{d}y\wedge\mathbf{d}z$, $\mathbf{d}z\wedge\mathbf{d}x$, and $\mathbf{d}x\wedge\mathbf{d}y$. Then we could write $\mathbf{d}\,(\mathbf{A}\cdot\mathrm{d}\mathbf{x}) = \boldsymbol{\nabla}\times\mathbf{A}\cdot\mathrm{d}\mathbf{S}$.

Example 1.21. Here is an intriguing example: the exterior derivative of a pseudo-2-form $\boldsymbol{\tau}$ in $\mathbb{R}^3$ with some metric $\mathfrak{g}$. Since this will be a pseudo-3-form, we expect it to be a one-component object. Indeed:

$$\begin{aligned}\mathbf{d}\boldsymbol{\tau} \;&=\; (\partial_u\tau_{vw}\, \mathbf{d}u) \,\wedge \mathbf{d}v\wedge\mathbf{d}w \,+\, (\partial_v\tau_{wu}\, \mathbf{d}v) \,\wedge \mathbf{d}w\wedge\mathbf{d}u \,+\, (\partial_w\tau_{uv}\mathbf{d}w) \,\wedge \mathbf{d}u\wedge\mathbf{d}v \\ &=\; (\partial_u\tau_{vw} \,+\, \partial_v\tau_{wu} \,+\, \partial_w\tau_{uv})\, \mathbf{d}u\wedge\mathbf{d}v\wedge\mathbf{d}w\end{aligned} \tag{1.58}$$

Now, in three-dimensions $\boldsymbol{\tau}$ can be viewed as the Hodge dual, $\boldsymbol{\tau} = {}^*\boldsymbol{\sigma}$, of the 1-form $\boldsymbol{\sigma} = \sigma_u\mathbf{d}u + \sigma_v\mathbf{d}v + \sigma_w\mathbf{d}w$. In terms of components, $\tau_{\mu\nu} = \epsilon_{\mu\nu\lambda}\sigma^\lambda$. Inserting and then taking the Hodge dual of the last expression, using ${}^*(\mathbf{d}u\wedge\mathbf{d}v\wedge\mathbf{d}w) = \epsilon^{123} = (-1)^{n_-}/\sqrt{|g|}$ from section 1.5.3, gives:

$$(-1)^{n_-}\,{}^*\mathbf{d}\,{}^*\boldsymbol{\sigma} \;=\; \frac{1}{\sqrt{|g|}}\,\partial_\mu(\sqrt{|g|}\,\sigma^\mu) \tag{1.59}$$

Extending to n dimensions, we call the right-hand side the **divergence**, $\mathbf{div}\,\mathbf{B}$, of the n-dim vector $\mathbf{B}$ with components $B^\mu = \sigma^\mu$. It is valid in general coordinates in any metric-endowed space. In vector calculus with Cartesian coordinates, $\boldsymbol{\tau} = B^1\mathbf{d}y\wedge\mathbf{d}z + B^2\mathbf{d}z\wedge\mathbf{d}x + B^3\mathbf{d}x\wedge\mathbf{d}y = \mathbf{B}\cdot\mathrm{d}\mathbf{S}$, and eq. (1.58) could be written as: $\mathrm{d}(\mathbf{B}\cdot\mathrm{d}\mathbf{S}) = \mathbf{div}\,\mathbf{B}\,\mathbf{d}^3x$.

The operator ${}^*\mathbf{d}\,{}^*$ sends a p-form into a (p-1)-form. In mathematical references, this operator is introduced (up to a sign!) as the **codifferential operator**, $\boldsymbol{\delta}$. We quote without proof the relation between them: When acting on a p-form in a Euclidean manifold, $\boldsymbol{\delta}\,\boldsymbol{\sigma} = (-1)^{n(p+1)+1}{}^*\mathbf{d}\,{}^*\boldsymbol{\sigma}$, and $\boldsymbol{\delta}\,\boldsymbol{\sigma} = (-1)^{n(p+1)}{}^*\mathbf{d}\,{}^*\boldsymbol{\sigma}$ in a pseudo-Euclidean manifold. Actually, these expressions happen to hold also in a Riemannian (curved) or pseudo-Riemannian manifold!

Like the exterior derivative, the codifferential operator is nilpotent. Indeed, $\boldsymbol{\delta}^2 = {}^*\mathbf{d}\,{}^*\mathbf{d}\,{}^* = \pm{}^*\mathbf{d}^2\,{}^* = 0$.

Definition 1.34. We shall define the **divergence** of any p-form, $\mathbf{div}\,\boldsymbol{\sigma}$, as $-\boldsymbol{\delta}\boldsymbol{\sigma} = (-1)^{n(p+1)+n_-}{}^*\mathbf{d}\,{}^*\boldsymbol{\sigma}$. This will ensure consistency between eq. (1.59) and the conversion between ${}^*\mathbf{d}\,{}^*$ and $\boldsymbol{\delta}$. We extend eq. (1.59) to the divergence of any p-form $\boldsymbol{\sigma}$ on a n-dim space:

$$(\mathbf{div}\,\boldsymbol{\sigma})_{\mu_1\ldots\mu_{p-1}} \;=\; \frac{1}{\sqrt{|g|}}\,\partial_\nu\big(\sqrt{|g|}\,\sigma^\nu{}_{\mu_1\ldots\mu_{p-1}}\big) \;=\; \frac{1}{\sqrt{|g|}}\,\partial_\nu\big(\sqrt{|g|}\,g^{\nu\rho}\sigma_{\rho\mu_1\ldots\mu_{p-1}}\big) \tag{1.60}$$

From eq. (1.59) also follows the definition of the 3-dim **Laplacian** of a scalar function f in coordinates u^i:

$$\nabla^2 f \;=\; \frac{1}{\sqrt{|g|}}\,\partial_i\big(\sqrt{|g|}\partial^i f\big) \;=\; \frac{1}{\sqrt{|g|}}\,\partial_i\big(\sqrt{|g|}\,g^{ij}\partial_j f\big) \tag{1.61}$$

1.6.2 Laplace-Beltrami operator, harmonic forms, and the Hodge decomposition

Definition 1.35. The **Laplace-Beltrami operator** is defined as $\boldsymbol{\Delta} = \boldsymbol{\delta}\,\mathbf{d} + \mathbf{d}\,\boldsymbol{\delta}$, sending p-forms to other p-forms of the same rank. When acting on a 0-form (scalar function), $\boldsymbol{\Delta} = \boldsymbol{\delta}\,\mathbf{d}$ always.

It is not hard to show that it reduces to the negative of the Laplacian operator of vector analysis, ie. $\boldsymbol{\Delta} = \boldsymbol{\delta}\,\mathbf{d} = -{}^*\mathbf{d}\,{}^*\mathbf{d} = -\partial_i\partial^i = -\nabla^2$, when acting on 0-forms on Euclidean $\mathbb{R}^3$ with Cartesian coordinates. We shall *define* ∇^2 so that $\nabla^2 = -\boldsymbol{\Delta}$ when acting on any p-form in Euclidean $\mathbb{R}^3$ equipped with a standard basis.

For instance, let it act on a 1-form $\boldsymbol{\sigma}$ in Euclidean $\mathbb{R}^3$. That is, take $\boldsymbol{\Delta}\boldsymbol{\sigma} = {}^*\mathbf{d}\,{}^*\mathbf{d}\boldsymbol{\sigma} - \mathbf{d}\,{}^*\mathbf{d}\,{}^*\boldsymbol{\sigma}$ using the conversion formula between δ and ${}^*\mathbf{d}\,{}^*$. Using eq. (1.57), the first term is the curl of a curl, whereas the second is the gradient of a divergence. Thus, we recover the expression well-known from vector calculus: $\boldsymbol{\nabla}^2\mathbf{A} = \boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\mathbf{A}) - \boldsymbol{\nabla}\times\boldsymbol{\nabla}\times\mathbf{A}$, where $\mathbf{A}$ is the vector associated with the 1-form $\boldsymbol{\sigma}$.

When acting on functions (0-forms) in Minkowski space, the Laplace-Beltrami operator becomes the negative of the d'Alembertian wave operator $\Box = \partial_\mu\partial^\mu$: $\Delta = -\Box$. In fact, this will *define* the d'Alembertian of any p-form in a Minkowski manifold.

Definition 1.36. A p-form $\boldsymbol{\sigma}$ is said to be **harmonic** if $\boldsymbol{\Delta}\boldsymbol{\sigma} = 0$. This generalises the notion of functions being called harmonic when they satisfy the Laplace equation.

Definition 1.37. A **closed** form is one whose exterior derivative vanishes. A p-form that can be written as the exterior derivative of a (p-1)-form is said to be **exact**.

Clearly, Poincaré's lemma states that an exact form is closed. But is a closed form exact, ie. if $\mathbf{d}\boldsymbol{\sigma} = 0$, does it follow that $\boldsymbol{\sigma} = \mathbf{d}\boldsymbol{\tau}$, with $\boldsymbol{\tau}$ uniquely determined? The answer is no, if only because one can always add the exterior derivative of an arbitrary (p - 2)-form $\boldsymbol{\theta}$ to $\boldsymbol{\tau}$ and still satisfy $\mathbf{d}\boldsymbol{\sigma} = 0$. Also, the converse of Poincaré's lemma (not proved) states that only in a submanifold in which all closed curves can be shrunk to a point does $\mathbf{d}\boldsymbol{\sigma} = 0$ entail the existence in that submanifold of a non-unique (p - 1)-form whose exterior derivative is $\boldsymbol{\sigma}$. In topology, we say that the submanifold must be **simply connected** (eg. no doughnuts!).

We quote without proof an important result: On finite-volume (compact) manifolds without boundaries, such as S^n, or on a torus, $\boldsymbol{\Delta}\boldsymbol{\sigma} = 0$ if, and only if, $\mathbf{d}\boldsymbol{\sigma} = 0$ and $\boldsymbol{\delta}\,\boldsymbol{\sigma} = 0$ (or ${}^*\mathbf{d}\,{}^*\boldsymbol{\sigma} = 0$). We then say that harmonic forms are both closed and co-closed. Usefully, this property also holds on open manifolds (eg. $\mathbb{R}^n$) if $\boldsymbol{\sigma}$ has **compact support** (it vanishes outside a bounded closed region), or if it goes to zero sufficiently fast at infinity. This is often the case for physically useful quantities.

Definition 1.38. Assuming a compact manifold without boundaries or, failing that, compact support or sufficiently fast fall-off at infinity, the **Hodge decomposition** of a p-form is:

$$\boldsymbol{\sigma} \;=\; \mathbf{d}\boldsymbol{\alpha} \,+\, \boldsymbol{\delta}\,\boldsymbol{\beta} \,+\, \text{harmonic form} \tag{1.62}$$

where $\boldsymbol{\alpha}$ is a (p-1)-form and $\boldsymbol{\beta}$ is a (p+1)-form. This decomposition is unique, in the sense that each term in the sum is unique, but not necessarily $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$ themselves. $\mathbf{d}\boldsymbol{\alpha}$, $\boldsymbol{\delta}\,\boldsymbol{\beta}$ and the harmonic p-form in the decomposition live in orthogonal subspaces of $\bigwedge^p$.

Example 1.22. In non-compact Euclidean $\mathbb{R}^3$, if $\mathbf{A}$ is a vector field with compact support, then the Hodge decomposition says that its associated 1-form can be written as the exterior derivative of a 0-form (ie. the gradient of a function), plus the divergence of a 2-form, $\boldsymbol{\beta}$, plus some harmonic 1-form. Now, since ${}^*\boldsymbol{\beta}$ is a pseudo-1-form in $\mathbb{R}^3$, $\boldsymbol{\delta}\,\boldsymbol{\beta} = {}^*\mathbf{d}\,{}^*\boldsymbol{\beta}$ is a 1-form, and we find from eq. (1.57) that this term in the Hodge decomposition corresponds to the curl of a vector. Therefore, we obtain *in terms of vectors* the **Helmholtz decomposition** for a vector field with compact support or sufficiently fast fall-off at infinity:

$$\mathbf{A} \;=\; \boldsymbol{\nabla}\phi \,+\, \boldsymbol{\nabla}\times\mathbf{M} \,+\, \mathbf{H} \tag{1.63}$$

where ϕ is a scalar field, $\mathbf{M}$ a vector field, and $\mathbf{H}$ another vector field which satisfies $\boldsymbol{\nabla}^2\mathbf{H} = 0$ everywhere. But if $\mathbf{H}$ vanishes at infinity in $\mathbb{R}^3$, then it must vanish everywhere.

The curl of the first term in the Helmholtz decomposition vanishes *identically*, and is often called the **longitudinal** projection of $\mathbf{A}$; the divergence of the second term vanishes *identically*, and we can call it the **transverse** projection of $\mathbf{A}$. It follows that $\boldsymbol{\nabla}\cdot\mathbf{A}$ contains no information whatsoever about the transverse part of $\mathbf{A}$, whereas $\boldsymbol{\nabla}\times\mathbf{A}$ knows nothing of its longitudinal part. This provides a very useful and powerful tool for analysing 3-dim first-order field equations (eg. Maxwell's equations) which are usually statements about the divergence and the curl of the fields. Conversely, if $\boldsymbol{\nabla}\cdot\mathbf{A} = 0$ *everywhere*, we can conclude that $\mathbf{A}$ is purely transverse, since then the 3-scalar field ϕ in eq. (1.63) satisfies the Laplace equation *everywhere*, which means it must vanish if it has compact support or fast enough fall-off.

1.6.3 Exterior derivative and codifferential operator of a 2-form in Minkowski spacetime

Let $\mathbf{F} \in \bigwedge^2$ on Minkowski (pseudo-Euclidean) $\mathbb{R}^4$. Demand that $\mathbf{F}$ be exact, with compact support or sufficiently fast fall-off at infinity. Then there exists a 1-form $\mathbf{A}$ such that $\mathbf{F} = \mathbf{d}\mathbf{A}$, and $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$, in any metric. This means that $\mathbf{d}\mathbf{F} = 0$. But this determines $\mathbf{A}$ only up to an additive term $\mathbf{d}f$, where f is an arbitrary scalar (0-form) field. It is also clear from Poincaré's lemma, $\mathbf{d}^2\mathbf{A} = 0\ \forall\ \mathbf{A}$, that any information about $\mathbf{F}$ contained in $\mathbf{d}\mathbf{F} = 0$, is of no use whatsoever in determining $\mathbf{A}$: we say that it is an identity on $\mathbf{A}$.

In addition, we give the exterior derivative of the Hodge dual of $\mathbf{F}$, the pseudo-2-form $\mathbf{d}\,{}^*\mathbf{F}$, as a "source" pseudo-3-form $\boldsymbol{\mathcal{J}}$, with corresponding Hodge dual 1-form $\mathbf{J} = {}^*\boldsymbol{\mathcal{J}}$. Then we have the inhomogeneous equation:

$$\mathbf{d}\,{}^*\mathbf{F} \;=\; 4\pi\,\boldsymbol{\mathcal{J}} \tag{1.64}$$

Interestingly enough, one can view eq. (1.64) as a statement that the source 3-form is exact, and therefore closed. This is a very important property because, if we take the exterior derivative of the equation, the left-hand side vanishes *identically*, and the right-hand side becomes: $\mathbf{d}\,\boldsymbol{\mathcal{J}} = 0$ or, equivalently, the better-known statement that the 4-divergence of the source 1-form vanishes: ${}^*\mathbf{d}\,{}^*\mathbf{J} = 0$. $\mathbf{d}\,\boldsymbol{\mathcal{J}} = 0$ is actually a (metric-independent) conservation law for the source!

What we have just seen is Maxwell's theory, with $\mathbf{F}$ the Faraday 2-form (tensor), $\mathbf{A}$ the electromagnetic potential 1-form, and $\mathbf{J}$ the electromagnetic 4-current.

The application to electromagnetism assumes a mostly positive metric, as in MTW or Griffith's *Introduction to Electrodynamics*. With a mostly negative metric, there would be a minus sign on the right-hand side of eq. (1.64).

Because the differential operator $\mathbf{d}$ is metric-independent, we have given both the homogeneous and inhomogeneous first-order Maxwell equations in terms of exterior derivatives of $\mathbf{F}$ and its dual ${}^*\mathbf{F}$. It is easy to convert the inhomogeneous equation to a divergence, simply by taking its Hodge dual:

$$ {}^*\mathbf{d}\,{}^*\mathbf{F} \;=\; 4\pi\,{}^*\boldsymbol{\mathcal{J}} \;=\; 4\pi\,\mathbf{J} \qquad\qquad \Longleftrightarrow \qquad\qquad \mathbf{div}\,\mathbf{F} \;=\; -\,4\pi\mathbf{J} \tag{1.65} $$

In terms of Cartesian components, this can be shown (EXERCISE) to be equivalent to[†]

$$ \partial^\mu F_{\mu\nu} \;=\; -\,4\pi\,J_\nu \qquad\qquad \Leftrightarrow \qquad\qquad \partial_\mu F^{\mu\nu} \;=\; -\,4\pi\,J^\nu $$

the latter form being more appropriate if we insist on thinking of the source term as a vector. I would argue, however, that the less conventional form eq. (1.64) is much the more natural. The exterior derivative is metric-independent, and its index form can be written entirely with covariant indices, the natural ones for p-forms. But to obtain its equivalent in divergence form, we have to Hodge-dualise the right-hand side, so that the vector $\mathbf{J}$ source depends on the metric (see the paragraph after eq. (1.52)), whereas its 3-form version does not. The price, of course, is that the 3-form version has 64 compnents.

As a 3-form, the homogeneous equation $\mathbf{dF} = 0$ also has a lot of components, and when it comes to solving the system, we may want to extract only the independent ones. Well, this equation is the same as $\mathbf{d}\,{}^*({}^*\mathbf{F}) = 0$ for which the Hodge dual is $\boldsymbol{\delta}{}^*\mathbf{F} = 0$. In other words, the divergence of the dual of $\mathbf{F}$ vanishes, which represents only four equations. Actually, this is a general, easily shown property: whenever the exterior derivative of a p-form in some manifold vanishes, so does the codifferential of its dual, and vice-versa. We note, however, that the homogoneous Maxwell equation expressed this way becomes metric-dependent, and its index form will not be the same in curved spacetime!

Another great advantage of writing Maxwell's equations in the geometric, index-independent formalism is that, provided the source is smoothly varying, they are formally the same in curved spacetime! All the extra terms due to the covariant derivatives that one has to introduce in General Relativity cancel out, essentially because they are symmetric in the same indices in which the p-form equations are antisymmetric. In index notation, however, only the equations involving no Hodge dual remain the same.

And nothing prevents us from constructing an extended Maxwell-like theory (which does not describe electromagnetism) involving $\mathbf{F}$ as a 3-form instead of a 2-form. In the past few decades it has received a good deal of attention in some quarters.

1.7 Integrals of Differential (Pseudo)Forms

In the course of figuring out the meaning of $\int \boldsymbol{\sigma}^p$, where we now use the notation $\boldsymbol{\sigma}^p$ to show explicitly the rank of a p-form, we shall discover that pretty much any integral that makes sense in n-dim calculus can be written as the integral of some p-form or pseudo-p-form.

1.7.1 Integrals of (pseudo)p-forms over a p-dim submanifold

As a warm-up, consider the integral of the Hodge dual of a scalar function f, $\int {}^*f$, over a n-dim region V in $\mathbb{R}^n$ (eg., over some volume in $\mathbb{R}^3$). The Hodge dual of a scalar function, of course, is a n-form. Then:

$$ \int_V {}^*f \;=\; \int_V f(\mathbf{u})\,\sqrt{|g|}\,\mathbf{d}u^1 \wedge \cdots \wedge \mathbf{d}u^n \;=\; \int_V f(\mathbf{x})\,\mathbf{d}x^1 \wedge \cdots \wedge \mathbf{d}x^n \;=\; \int_V f(\mathbf{x})\,\mathbf{d}^n x $$

where u are general coordinates and $\mathbf{d}^n x$ is in the volume pseudo-n-form in Cartesian coordinates. If the coordinate basis $\{\boldsymbol{\partial}_\mu\}$ $(1 \le \mu \le n)$ has the same orientation as the (arbitrarily chosen) orientation of the region, then we define:

[†] Again, with a mostly negative metric, such as in Jackson's *Classical Electrodynamics*, there would be no minus sign on the right-hand side. This is because $\mathbf{F}$ has opposite sign between the two conventions so as to obtain the same relations between the electric and magnetic fields and the vector and scalar potentials.

Definition 1.39.

$$\int_V f(\mathbf{x})\,\mathbf{d}x^1 \wedge \cdots \wedge \mathbf{d}x^n \;\coloneqq\; \int_V f(\mathbf{x})\,\mathrm{d}x^1 \cdots \mathrm{d}x^n \;=\; \int_V f(\mathbf{x})\,\mathrm{d}^n x \tag{1.66}$$

ie. the ordinary multiple integral of a *scalar* function of n variables in n dimensions.

When a p-dim region R is embedded in a n-dim manifold, it will be described with some coordinates $\mathbf{u}(\mathbf{x})$, that is, n functions u^i of the p Cartesian coordinates x^j that parametrise $\mathbb{R}^p$. Also, an orientation can be defined for the region. What is the meaning of the integral of a p-form over such a region? We give two examples in $\mathbb{R}^3$.

Example 1.23. Integral of a 1-form over a curve or "line integral"

As we have seen before, a curve C can be parametrised in terms of some parameter $t \in \mathbb{R}$, with d_t defining the positive orientation of the parameter space $\mathbb{R}$. Then, if $\boldsymbol{\alpha}$ is a 1-form on $\mathbb{R}^3$, eq. (1.66) and the chain rule yield:

$$\int_C \boldsymbol{\alpha} \;=\; \int_C \alpha_i \mathbf{d}u^i \;=\; \int_a^b \alpha_i[\mathbf{u}(t)]\,\mathrm{d}_t u^i\,\mathrm{d}t \;=\; \int_a^b \boldsymbol{\alpha}(\mathrm{d}_t\mathbf{u})\,\mathrm{d}t$$

Only if $\mathbb{R}^3$ is given a metric and the curve parametrised with Cartesian coordinates is this the usual integral of a *vector* $\mathbf{A}$ along the curve, $\int \mathbf{A}\cdot\mathrm{d}\mathbf{x}$. In general, to integrate a vector along a curve, a metric *must* be introduced so as to transform the vector components into its associated 1-form's components: $\int \mathbf{A}\cdot\mathrm{d}\mathbf{u} = \int g_{ij}A^j\mathrm{d}_t u^i\,\mathrm{d}t$. But no metric is needed to integrate a 1-form along a curve, and this is the simpler and more natural operation.

If $\boldsymbol{\alpha}$ is exact, then we immediately have the fundamental theorem of calculus:

$$\int_C \mathbf{d}f \;=\; \int_a^b \partial_{u^i} f\,\mathrm{d}_t u^i\,\mathrm{d}t \;=\; \int_a^b \mathrm{d}f \;=\; f(b) - f(a) \;=\; \int_{\partial C} f$$

where ∂C is the boundary, ie. the end-points, of the curve.

Example 1.24. Integral of a 2-form over a surface

Let S be some surface embedded in $\mathbb{R}^3$ and described with three coordinate functions $u^i(x^1, x^2)$. The surface is parametrised with $(x^1,\, x^2) \in \mathbb{R}^2$, with basis vectors $\boldsymbol{\partial}_{x^1} \equiv \boldsymbol{\partial}_1$ along the x^1 direction and $\boldsymbol{\partial}_{x^2} \equiv \boldsymbol{\partial}_2$ along the x^2 direction, for which some orientation has been defined as positive. What meaning can we give to the integral of a 2-form $\boldsymbol{\beta}$ over S? From the chain rule and eq. (1.66) we find:

$$\int_S \boldsymbol{\beta} \;=\; \int_S \beta_{jk}\,\mathbf{d}u^j \wedge \mathbf{d}u^k \;=\; \int \beta_{jk}[\mathbf{u}(x^1,x^2)]\,(\partial_1 u^j\,\partial_2 u^k \;-\; \partial_2 u^j\,\partial_1 u^k)\,\mathrm{d}x^1\,\mathrm{d}x^2 \qquad (j<k)$$

The integrals in $\mathbb{R}^2$ on the right are over a rectangular region in parameter space. The two coordinate vectors (see section 1.2.3), $\partial_1\mathbf{u}$ and $\partial_2\mathbf{u}$, are tangent to the surface at every point, and are usually linearly independent, so form a basis for the space of (pseudo)vectors tangent to the surface. Here again, no metric is required.

Now the Hodge dual of $\boldsymbol{\beta}$, which is a pseudo-1-form, has an associated pseudo-vector $\mathbf{B}$ with, as components, the contravariant components of the Hodge dual, $B^i = \epsilon^{ijk}\beta_{jk}$ $(j<k)$, eg., $B^1 = \beta_{23}/\sqrt{|g|}$, etc. Then:

$$\beta_{jk}\,(\partial_1 u^j \partial_2 u^k \;-\; \partial_2 u^j\,\partial_1 u^k) \;=\; \epsilon_{ijk}\,B^i\,(\partial_1 u^j \partial_2 u^k \;-\; \partial_2 u^j\,\partial_1 u^k) \;=\; \sqrt{|g|}\begin{vmatrix} B^1 & B^2 & B^3 \\ \partial_1 u^1 & \partial_1 u^2 & \partial_1 u^3 \\ \partial_2 u^1 & \partial_2 u^2 & \partial_2 u^3 \end{vmatrix}$$

From eq. (1.47), we recognise the last member of the equality as the output obtained from inserting the three vectors whose components are the rows of the determinant, into the three input slots of a simple

3-form—more accurately, a pseudo-3-form which, from definition (1.31) can be identified with the volume pseudo-form $\mathbf{d}^3u$. Then our integral can be written:

$$\int_S \boldsymbol{\beta} \;=\; \int \mathbf{d}^3u(\mathbf{B}, \partial_1\mathbf{u}, \partial_2\mathbf{u})\,\mathrm{d}x^1\,\mathrm{d}x^2$$

This makes it obvious that the integral is independent of the orientation of $\mathbb{R}^3$, since switching it flips the sign of both $\mathbf{B}$ and $\mathbf{d}^3u$. At every point on the surface, we can choose the unit $\mathbf{n}$ normal to the surface so that $\mathbf{n}$ and the vectors $\partial_1\mathbf{u}$ and $\partial_2\mathbf{u}$ tangent to the surface form a right-handed (positive orientation) system. We also note that only the normal component of $\mathbf{B}$ can contribute to the integral (WHY?).

Then the scalar functionl $\mathbf{d}^3u(\mathbf{B}, \partial_1\mathbf{u}, \partial_2\mathbf{u})$ is the normal component of $\mathbf{B}$ multiplied by the surface of the parallelogram defined by the coordinate vectors. Defining the surface element $\mathrm{d}S \equiv |\partial_1\mathbf{u} \times \partial_2\mathbf{u}|$, there comes:

$$\int_S \boldsymbol{\beta} \;=\; \int B_\mathrm{n}\,\mathrm{d}S \;=\; \int \mathbf{B}\cdot\mathrm{d}\mathbf{S} \tag{1.67}$$

where the often used last expression is called the **flux** of the pseudo-vector $\mathbf{B}$ through the surface S. It does not depend on the parametrisation chosen for S which is integrated out. The same result holds if $\boldsymbol{\beta}$ is a pseudo-2-form, with $\mathbf{B}$ now a vector.

1.7.2 Stokes Theorem

This famous theorem, which we shall not prove, equates the integral of the exterior derivative of a differentiable (pseudo)p-form, $\boldsymbol{\omega}$, over a bounded region V in a manifold to the integral of $\boldsymbol{\omega}$ over the boundary ∂V of V. A technicality is that both V and ∂V must have compatible orientations.

$$\int_V \mathbf{d}\boldsymbol{\omega} \;=\; \int_{\partial V} \boldsymbol{\omega} \tag{1.68}$$

The boundary need not be connected, and it can be broken up into non-overlapping parts when it cannot be covered by a single coordinate patch. Then we simply sum the integrals over each part.

Example 1.25. At the end of example 1.23 we had already worked out an application when $\boldsymbol{\omega}$ is a 0-form: the fundamental theorem of calculus. When $\boldsymbol{\omega}$ is a 1-form and V a 2-dim surface in Euclidean $\mathbb{R}^3$ parametrised with Cartesian coordinates and bounded by a closed curve C, the same example gives immediately: $\int_{\partial V}\boldsymbol{\omega} = \oint_C \mathbf{A}\cdot\mathrm{d}\mathbf{u}$, From eq. (1.56) and example 1.24, $\int_S \mathbf{d}\boldsymbol{\omega} = \int_S \boldsymbol{\nabla}\times\mathbf{A}\cdot\mathrm{d}\mathbf{S}$, and we recover the well-known Kelvin-Stokes formula. Only if we want to work with that usual vector calculus relation is a metric required.

Finally, when $\boldsymbol{\omega}$ is a pseudo-2-form in Euclidean $\mathbb{R}^3$, V a volume and S a surface enclosing the volume, we recover the formula $\int_V \boldsymbol{\nabla}\cdot\mathbf{B}\,\mathrm{d}V = \oint_S \mathbf{B}\cdot\mathrm{d}\mathbf{S}$ from examples 1.21 and 1.24. This time, a metric is needed.

1.8 Maxwell Differential Forms in Three + One Dimensions

With $\mathbf{F}$ the Maxwell 2-form, define two 3-dim p-forms: an electric field strength 1-form $\boldsymbol{\mathcal{E}}$ and a magnetic field strength 2-form $\boldsymbol{\mathcal{B}}$, by:

$$\mathbf{F} \;=\; F_{|\mu\nu|}\mathrm{d}x^\mu\wedge\mathbf{d}x^\nu \;=\; \boldsymbol{\mathcal{E}}\wedge\mathbf{d}t \,+\, \boldsymbol{\mathcal{B}} \tag{1.69}$$

where:

$$\boldsymbol{\mathcal{E}} \;:=\; F_{10}\mathbf{d}x^1 \,+\, F_{20}\mathbf{d}x^2 \,+\, F_{30}\mathbf{d}x^3 \qquad \boldsymbol{\mathcal{B}} \;:=\; F_{12}\mathbf{d}x^1\wedge\mathbf{d}x^2 \,+\, F_{31}\mathbf{d}x^3\wedge\mathbf{d}x^1 \,+\, F_{23}\mathbf{d}x^2\wedge\mathbf{d}x^3 \tag{1.70}$$

Now, formally, $\mathbf{d} = \vec{\mathbf{d}} + \mathbf{d}t \wedge \partial_t$. where $\vec{\mathbf{d}}$ denotes the *3-dimensional* exterior derivative. Then the homogeneous Maxwell equation, $\mathbf{dF} = 0$, becomes:

$$\begin{aligned}\left[\vec{\mathbf{d}} + \mathbf{d}t \wedge \partial_t\right]\left[\boldsymbol{\mathcal{E}} \wedge \mathbf{d}t + \boldsymbol{\mathcal{B}}\right] &= \vec{\mathbf{d}}\boldsymbol{\mathcal{E}} \wedge \mathbf{d}t + \vec{\mathbf{d}}\boldsymbol{\mathcal{B}} + \mathbf{d}t \wedge \partial_t\boldsymbol{\mathcal{B}} \\ &= \left(\vec{\mathbf{d}}\boldsymbol{\mathcal{E}} + \partial_t\boldsymbol{\mathcal{B}}\right) \wedge \mathbf{d}t + \vec{\mathbf{d}}\boldsymbol{\mathcal{B}} \\ &= 0\end{aligned}$$

In the second line, the plus sign in the bracket results from the double transposition needed to bring $\mathbf{d}t$ to the right through the two basis 1-forms of $\boldsymbol{\mathcal{B}}$. We conclude that in three dimensions, the homogeneous Maxwell equation gives rise to:

$$\vec{\mathbf{d}}\boldsymbol{\mathcal{B}} = 0 \qquad\qquad \vec{\mathbf{d}}\boldsymbol{\mathcal{E}} + \partial_t\boldsymbol{\mathcal{B}} = 0 \tag{1.71}$$

Expand the potential 1-form $\mathbf{A}$ as $-\phi\,\mathbf{d}t + \boldsymbol{\mathcal{A}}$, where $\phi \in \bigwedge^0$ and $\boldsymbol{\mathcal{A}} \in \bigwedge^1$, both in *three* dimensions. Expanding $\mathbf{F} = \mathbf{dA}$, or using the Poincaré lemma for $\vec{\mathbf{d}}$, one obtains (EXERCISE):

$$\boldsymbol{\mathcal{E}} = -\mathbf{d}\phi - \partial_t\boldsymbol{\mathcal{A}}, \qquad\qquad \boldsymbol{\mathcal{B}} = \vec{\mathbf{d}}\boldsymbol{\mathcal{A}} \tag{1.72}$$

I wish to reiterate that eq. (1.71) and (1.72) are *metric-independent*, and will thus hold in flat or curved spacetime, so long as we use a coordinate basis.

The Hodge duals of eq. (1.71) can be written as:

$$\mathrm{div}\,{}^*\boldsymbol{\mathcal{B}} = 0 \qquad\qquad {}^*\vec{\mathbf{d}}\boldsymbol{\mathcal{E}} + \partial_t{}^*\boldsymbol{\mathcal{B}} = 0$$

If we identify the contravariant components of the pseudo-1-form ${}^*\boldsymbol{\mathcal{B}}$ with the usual components of the magnetic-field pseudo-vector, and use eq. (1.57), we see that these are equivalent to the homogeneous Maxwell equations in their vector-calculus form: $\boldsymbol{\nabla} \cdot \mathbf{B} = 0$ and $\boldsymbol{\nabla} \times \mathbf{E} + \partial_t\mathbf{B} = 0$.

After Hodge-dualising the second equation (1.72), these are readily identified with the standard expressions for the fields in terms of the scalar and vector potential ${}^*\vec{\mathbf{d}}\boldsymbol{\mathcal{A}}$. But we also see that it is much more natural to view the 3-dim magnetic field as a 2-form which is the exterior derivative of a 1-form, than as a pseudo-vector which is the curl of another vector. In the same vein, I contend that it is more natural to describe the electric field strength with the 1-form $\boldsymbol{\mathcal{E}}$ than with the vector $\mathbf{E}$. It is consistent with force and momentum also being more naturally 1-forms (consider $\mathrm{e}^{\mathrm{i}p_\mu x^\mu}$!).

The inhomogeneous (source) Maxwell equation requires much more care, because the Hodge dual inevitably involves a metric, and because a 4-dim Hodge dual is not necessarily like a 3-dim Hodge dual! First, we must derive an expansion of ${}^*\mathbf{F}$ in terms of $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{\mathcal{B}}$. A safe, if somewhat inelegant, method is to expand it in terms of the components of $\mathbf{F} = \frac{1}{2}F_{\mu\nu}\mathbf{d}x^\mu \wedge \mathbf{d}x^\nu$:

$$\begin{aligned}{}^*\mathbf{F} &= \frac{1}{4}F^{\mu\nu}\,\epsilon_{\mu\nu\alpha\beta}\mathbf{d}x^\alpha \wedge \mathbf{d}x^\beta \\ &= -\sqrt{|g|}\,\left[F^{10}\mathbf{d}x^2 \wedge \mathbf{d}x^3 + F^{20}\mathbf{d}x^3 \wedge \mathbf{d}x^1 + F^{30}\mathbf{d}x^1 \wedge \mathbf{d}x^2 + \left(F^{12}\mathbf{d}x^3 + F^{31}\mathbf{d}x^2 + F^{23}\mathbf{d}x^1\right) \wedge \mathbf{d}t\right]\end{aligned}$$

Now we must write this in terms of the *covariant* components of $\mathbf{F}$, and this is where the metric must come in, since $F^{\mu\nu} = g^{\mu\alpha}g^{\nu\beta}F_{\alpha\beta}$:

$$F^{i0} = (g^{00}g^{ij} - g^{i0}g^{0j})F_{j0} + g^{ij}g^{0k}F_{jk}, \qquad\qquad F^{ij} = (g^{i0}g^{jl} - g^{il}g^{j0})F_{l0} + g^{ik}g^{jl}F_{kl}$$

We know that F_{j0} and F_{jk} are the components of the 3-dim p-forms $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{\mathcal{B}}$, respectively. If $g^{0i} \neq 0$, each contravariant component of $\mathbf{F}$ will involve *both* $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{\mathcal{B}}$, which will lead to very complicated results. When $g^{0i} = 0$, however, we are left with $F^{i0} = g^{00}g^{ij}F_{j0}$, and $F^{ij} = g^{ik}g^{jl}F_{kl}$, and lowering the spatial components of $\mathbf{F}$ involves *only the spatial sector of the metric* (ignoring the g^{00} factor), the same sector that is used to raise indices on the Levi-Civita tensor. Also, if we take $g^{00} = -1$ (mostly positive) Minkowski metric, the $\sqrt{|g|}$ factor

is the same for the three-dimensional metric determinant as for the 4-dim one. Because of all this, we can now write:

$$ {}^*\mathbf{F} \;=\; -\left[\frac{1}{2}\epsilon_{ijk}F^{i0}\mathbf{d}x^j\wedge\mathbf{d}x^k \;+\; \frac{1}{2}\epsilon_{ijk}F^{ij}\mathrm{d}x^k\wedge\mathbf{d}t\right] $$

where the roman indices run from 1 to 3. Now we can relate the two terms to $\mathcal{E}$ and $\mathcal{B}$:

$$ \frac{1}{2}\,\epsilon_{ijk}F^{i0}\mathbf{d}x^j\wedge\mathbf{d}x^k \;=\; \frac{1}{2}\,\epsilon_{ijk}g^{00}g^{il}F_{l0}\mathbf{d}x^j\wedge\mathbf{d}x^k \;=\; \frac{1}{2}\,g^{00}\epsilon_{ijk}\mathcal{E}^i\mathbf{d}x^j\wedge\mathbf{d}x^k \;=\; g^{00}\,{}^*\boldsymbol{\mathcal{E}} \;=\; -{}^*\boldsymbol{\mathcal{E}} $$

Also:

$$ \frac{1}{2}\,\epsilon_{ijk}F^{ij}\mathbf{d}x^k \;=\; {}^*\boldsymbol{\mathcal{B}} $$

with no assumption needed for the spatial part of the 4-dim metric. Then our expansion is ${}^*\mathbf{F} = -{}^*\boldsymbol{\mathcal{B}}\wedge\mathbf{d}t + {}^*\boldsymbol{\mathcal{E}}$ where it is understood that, on the right-hand side only, the *3-dim* Hodge dual is taken. It is not difficult to show (EXERCISE) that: $\mathbf{d}\,{}^*\mathbf{F} = -(\vec{\mathbf{d}}\,{}^*\boldsymbol{\mathcal{B}} - \partial_t\,{}^*\boldsymbol{\mathcal{E}})\wedge\mathbf{d}t + \vec{\mathbf{d}}\,{}^*\boldsymbol{\mathcal{E}}$.

We define the Maxwell source pseudo-3-form as the expansion:

$$ \boldsymbol{\mathcal{J}} \;\equiv \boldsymbol{\rho} \;-\; \mathbf{j}\wedge\mathbf{d}t \;\equiv\; \rho\,\epsilon_{ijk}\,\mathbf{d}x^i\wedge\mathbf{d}x^j\wedge\mathbf{d}x^k \;-\; {}^*\mathbf{J}\wedge\mathbf{d}t \qquad (i<j<k) $$

where ρ is the charge scalar density, $\boldsymbol{\rho}$ the *three-dim* charge-density pseudo-3-form and $\mathbf{J}$ the *3-dim* current density 1-form. Inserting these expansions in eq. (1.64) yields the two 3-dim Maxwell field equations:

$$ \vec{\mathbf{d}}\,{}^*\boldsymbol{\mathcal{E}} \;=\; 4\pi\boldsymbol{\rho}, \qquad\qquad \vec{\mathbf{d}}\,{}^*\boldsymbol{\mathcal{B}} \;=\; \mathbf{j} \;+\; \partial_t\,{}^*\boldsymbol{\mathcal{E}} \tag{1.73} $$

Taking the 3-dim Hodge dual of these equations recovers the vector-calculus form of Gauss's law for electricity and the Ampère-Maxwell equation.

2 MODULE II — GROUP THEORY 1: Discrete Groups

One of the most beautiful and useful concepts in physics and, dare I say, mathematics, is that of **symmetry**. Loosely speaking, it recognises the existence of patterns or, mathematically-speaking, a characteristic behaviour of objects under transformations. When a symmetry in a problem is detected, the problem is amenable to much simplification and might even be solvable. Useful information can be recovered even if the symmetry is only approximate, or is "broken" in a way that can be understood. Equally important, a symmetry often signals the existence of a **conserved** quantity. For instance, from space-translation invariance (aka homogeneity of space) follows linear-momentum conservation, whereas time-translation invariance generates energy conservation; isotropy of space (its invariance under rotations) leads to angular-momentum conservation. Conservation of electric charge is embodied in the **gauge invariance** of Maxwell's equations, which itself can be seen as generating constraints that automatically suppress potentially disastrous negative-energy modes.

In modern mathematics, the language of **group theory** provides a unified and systematic framework for classifying and describing symmetries. In part because it is jargon-heavy, group theory is often relegated to the fringes of most physicists' training. Yet much insight can be gained from at least a modicum of familiarity with it, and this is what we shall now attempt.

2.1 Groups: Definitions, Taxonomy and Examples (BF 10.1)

As much of what we are going to discuss involves mappings, it is worth saying a few words right at the outset about what kind of mappings will be of interest to us.

Most important, we will want to be able to **compose** mappings. For instance, if f and g are two mappings, we want to be able to apply one, and then the other; we shall denote[†] a binary composition by $f \circ g$ or $f \star g$, with the understanding that g is applied first. In order that a string of such mappings be uniquely defined, they must be **associative**, and thus reducible to a succession of binary mappings. Moreover, we should like to be able to retrace any step we take in a unique way; this can be achieved by demanding that any mapping f we consider be **invertible** with a mapping f^{-1}. Therefore, there must exist an identity mapping, e, such that $f \star f^{-1} = f^{-1} \star f = e$.

2.1.1 Basic definitions and nomenclature

The mappings of interest to us belong to a more general type of collection of objects called a **group**:

> **Definition 2.1.** Let G be a set of distinct objects endowed with an associative binary composition law or operation such that:
>
> - $\forall\, a,\, b \in G,\ \ a \star b \in G$;
> - there exists a unique element $e \in G$ such that, $\forall\, a \in G$, $e \star a = a \star e = a$;
> - $\forall\, a \in G$, there exists a unique element $a^{-1} \in G$ such that $a^{-1} \star a = a \star a^{-1} = e$;
>
> then we say that G is a **group**.

Two remarks are in order. The composition law is often called **group multiplication**, a term we shall try to avoid because it almost irresistibly evokes ordinary multiplication. Also, although a group composition law definitely meets the requirements set out above for mappings, group elements themselves may be mappings, or transformations (eg. translations, rotations, permutations), but also numbers, vectors, matrices, etc.

The composition of two elements of G is in general noncommutative: $a \star b \neq b \star a$.

> **Definition 2.2.** When $a \star b$ is commutative, $\forall\, a,\, b \in G$, we call G an **Abelian** group.

> **Definition 2.3.** A group of n elements ($n < \infty$) is said to be **finite** and **of order** n.

[†] Since a binary composition is not necessarily a multiplication, we avoid the potentially misleading $f \cdot g$ notation of Byron and Fuller (and many others).

Definition 2.4. A group is **discrete** when it contains a countable number of elements. All finite groups are discrete, but infinite discrete groups exist. Infinite groups which are not discrete are called **continuous**.

Definition 2.5. A group of the form $\{e,\, a,\, a^2,\, \ldots,\, a^{n-1}\}$, where $a^i \equiv a \star a \star \ldots$ i times, and where n is the smallest integer for which $a^n = e$, is called Z_n (sometimes C_n), the **cyclic group** of order n.

All a^i in Z_n are distinct for, supposing $a^i = a^j$, we would have $a^{i-j} = e$, with $i - j < n$, and n would not be the smallest number of compositions of a that yields e.

Given any element a_i of a finite group, there will be an integer m such that $a_i^m = e$. Then we say that a_i itself is **of order** m. If m is smaller than the order of the group, the group is not cyclic. One shows (EXERCISE) that groups whose elements all have order 2 are Abelian.

Here are a few straightforward examples of groups: $\mathbb{C}$ under addition, with $e = 0$ and $a^{-1} = -a$; $\mathbb{C} - \{0\}$ under multiplication, with $e = 1$ and $z^{-1} = 1/z$; the set of complex $n \times n$ matrices with non-zero determinant, $GL(n,\, \mathbb{C})$, under *matrix* multiplication, with $e = \mathbf{I}$; the set of the n complex roots of 1 under multiplication, with $e = 1$.

It is important to keep in mind that a given set may be a group under one operation, but not under another. Thus, the set $\mathbb{Z}$ is not a group under multiplication (and neither is $\mathbb{Z} - \{0\}$); but is is a group under addition with $e = 0$ and $a^{-1} = -a$.

Exercise: spot any discrete and cyclic groups in these examples.

2.1.2 Cayley tables

Let a_i $(i = 1, \ldots, n)$ be an element of a finite group. By convention, $a_1 = e$. We can construct a $n \times n$ **composition table**, or **Cayley table**, whose ij^{th} element is $a_i \star a_j$. Then the first row and the first column must be $\{e,\, a_2,\, \ldots,\, a_n\}$. They are sometimes omitted by authors who are not nice to their readers.

A constraint (and check!) is that any column or row of the composition table of a finite group must contain all elements of the group, and only once. Indeed, suppose that in the row corresponding to a_i we had $a_i \star a_j = a_i \star a_k$ for some $j,\ k$. But since a_i must have a unique inverse, this would force $a_j = a_k$. A similar argument can be made for columns.

Another constraint is that $a_i \star a_j = a_j$ only if $a_i = e$, and $a_i \star a_j = a_i$ only if $a_j = e$, which occurs only in the first row and the first column, respectively. It follows that $a_i \star a_j = a_j$ and $a_i \star a_j = a_i$ are ruled out in the other entries.

Constructing Cayley tables for cyclic groups is easy. Let us do it for $n = 2,\ 3$, and 4:

e	a
a	e

$\{e,\, a\}$

e	a	b
a	b	e
b	e	a

$\{e\,, a,\, b = a^2\}$

e	a	b	c
a	b	c	e
b	c	e	a
c	e	a	b

$\{e\,, a,\, b = a^2, c = a^3\}$

Notice that, for $n = 2$ and 3, these tables are the only ones we can construct that meet the two constraints mentioned just above. Therefore, finite groups of order 1, 2 and 3 are cyclic. The $n = 4$ case, however, would seem to open up more possibilities: if $a \star a = b$, the constraints determine the other entries, and we obtain a table for the cyclic group Z_4. But we could also take $a \star a = c$ and $a \star a = e$: In the former case, the constraints determine the rest of the table; in the latter case, we can choose $b \star b = a$ or $b \star b = e$, yielding:

e	a	b	c
a	c	e	b
b	e	c	a
c	b	a	e

e	a	b	c
a	e	c	b
b	c	a	e
c	b	e	a

e	a	b	c
a	e	c	b
b	c	e	a
c	b	a	e

By relabelling $b \longleftrightarrow c$ in the first table, and $a \longleftrightarrow b$ in the second, and re-ordering the rows and columns, we obtain tables which are identical to the cyclic table, and we conclude that in they are really those of the cyclic group of order 4.

The last table is genuinely different. it corresponds to a group of order 4, called the **four-group**—aka Felix Klein's Vierergruppe V—in which every element is its own inverse (and thus of order 2), with the fourth element constructed out of the other two non-identity elements (otherwise V would be cyclic!): $\{e,\, a,\, b,\, a\star b\}$. An example is the symmetry group, D_2, of a rectangular solid centered on the origin: the identity, and rotations by π about the x, y and z axes.

EXERCISE: Is it possible to construct the Cayley table of a group of order 5 from the requirement that all its elements be of order 2?

2.1.3 Generators of a group

Cyclic groups and the four-group illustrate a very useful feature of some groups: the fact that all their elements can be generated by a subset of those elements. More precisely:

> **Definition 2.6.** A set of **generators** of a group G is a subset of G from which all other elements of G can be obtained by repeated compositions of the generators among themselves.

Any element g of a finite group generates a cyclic group Z_n, where n is the order of g, whereas the four-group is obtained from two generators. EXERCISE: construct the Cayley table for the group of order 6: $\{e,\, a,\, b,\, b^2,\, a\star b,\, b\star a\}$.

As another example, we can think of a rotation by $\pi/6$ about the z axis as the generator of the finite group of rotations by $k\pi/6$ ($k = 0,\, \ldots,\, 11$).

2.1.4 Isomorphisms

We have just been introduced to the important idea that groups which look different may in some sense be the same because their composition tables are identical or can be made to be identical by relabelling. We now formalise this idea:

> **Definition 2.7.** If there exists a one-to-one correspondence between all the elements of one group $\{G, \circ\}$ and all the elements of another group $\{H, \star\}$ such that under this mapping the two groups have identical composition tables, then the mapping is said to be an **isomorphism**, and G and H are **isomorphic**: $G \cong H$.

Another definition can be more helpful when, as happens with continuous groups, there is no composition table as such:

> **Definition 2.8.** If there exists a one-to-one mapping f between all the elements of one group $\{G, \circ\}$ and all the elements of another group $\{H, \star\}$ such that under this mapping, $f(a),\ f(b) \in H$ and $f(a \circ b) = f(a) \star f(b)\ \forall\, a,\, b \in G$, then f is an isomorphism of G onto H, and $G \cong H$.

Other examples of isomorphic groups are

- the group of permutations of two objects (S_2), the group of rotations by π around the z axis, the group $\{1,\, -1\}$ (under multiplication);
- the group of complex numbers and the group of vectors in a plane, both under addition;
- the continuous groups $\{\mathbb{R},\, +\}$ and $\{\mathbb{R}^+,\, \times\}$ with the exponential as the isomorphism. Later we will see that because $\mathrm{e}^x\mathrm{e}^y = \mathrm{e}^{x+y}$, $\mathrm{e}^x \in \{\mathbb{R}^+,\, \times\}$ provides a one-dimensional matrix representation of $\{\mathbb{R},\, +\}$.

> **Definition 2.9.** A **homomorphism**, like an isomorphism, preserves group composition, but it is not one-to-one (eg. it could be many-to-one). .

2.2 Special Subsets of a Group (BF10.3)

There are a number of useful ways to classify the elements of a given group.

2.2.1 Special Ternary Compositions: Conjugacy Classes

Definition 2.10. Given $a \in G$, any element $b \in G$ which can be obtained as $b = x \circ a \circ x^{-1}$, where $x \in G$, is called the **conjugate of** a by x. This **conjugation** operation, which consists of two binary compositions, is analogous to similarity transformations on matrices.

Now conjugacy has the following properties:

- Reflexivity: $a = e \circ a \circ e^{-1}$, or a is self-conjugate.
- Symmetry: let $b = x \circ a \circ x^{-1}$. Then $a = y \circ b \circ y^{-1}$, with $y = x^{-1} \in G$.
- Transitivity: let $b = x \circ a \circ x^{-1}$ and $a = y \circ c \circ y^{-1}$. Then:

$$b \;=\; x \circ a \circ x^{-1} \;=\; x \circ y \circ c \circ y^{-1} \circ x^{-1} \;=\; (x \circ y) \circ c \circ (x \circ y)^{-1}$$

and since $x \circ y \in G$, b is conjugate to c.

This leads to the definition:

Definition 2.11. The subset of elements of a group which are conjugate to one another form a **conjugacy**, or **equivalence**[†], **class**, often abbreviated to just a **class**.

The systematic way of constructing the class for any element a_i of a group is to form:

$$(e \circ a_i \circ e^{-1},\, a_1 \circ a_i \circ a_1^{-1},\, \ldots,\, a_{i-1} \circ a_i \circ a_{i-1}^{-1}, a_{i+1} \circ a_i \circ a_{i+1}^{-1},\, \ldots)$$

From this it follows that e is always in a class by itself, and that each element of an Abelian group must also be in a class by itself. Since the cyclic group of order n is Abelian, each of its elements is its own class, and the same goes for the four-group.

Classes are disjoint: they have no element in common (EXERCISE: show this).

Elements in the same class share some properties. In particular, they must all be of the same order (proof on p. BF594). In a particularly important type of group, matrix groups, conjugate matrices are similar to one another; they could represent the same "thing" in different bases.

EXERCISE: obtain the classes for the group: $\{e,\, a,\, b,\, b^2,\, a \star b,\, b \star a\}$.

Note that there is another way to form interesting sets of conjugates: $\forall a_i \in G$, form set $\{x \circ a_i \circ x^{-1}\}$ for a *given* $x \in G$.

2.2.2 Subgroups

Definition 2.12. A subset H of a group G that behaves as a group in its own right, and under the same composition law as G is said to be a **subgroup** of G: $H \subseteq G$. A subgroup H of G is **proper** if it is **non-trivial** (ie. not e) and if $H \subset G$ (ie. $H \neq G$). Unlike classes, the subgroups of a group are not dsjoint, and can have more elements than e in common.

The four-group V has the (non-disjoint) proper subgroups: $\{e,\, a\}$, $\{e,\, b\}$, and $\{e,\, c = a \star b\}$. By inspection, the group of order 6 $\{e,\, a,\, b,\, b^2,\, a \star b,\, b \star a\}$ contains the proper cyclic subgroup $\{e,\, b,\, b^2\}$ of order 3.

[†] Actually, conjugacy is only a particular type of equivalence.

A Small Zoo of Famous (and Important!) Subgroups:

- The set of even integers is a subgroup of $\{\mathbb{Z}, +\}$, which is itself a subgroup of $\{\mathbb{R}, +\}$, which is a subgroup of $\{\mathbb{C}, +\}$. All those subgroups are proper.
- The **special** subset $SL(n, \mathbb{C})$ of the General Linear group $GL(n, \mathbb{C})$ of transformations represented by $n \times n$ invertible complex matrices **which have determinant one**.
- The subset $U(n)$ of the General Linear group $GL(n, \mathbb{C})$ of transformations represented by $n \times n$ complex matrices which are **unitary** ($\mathbf{U}^\dagger \mathbf{U} = \mathbf{I}$), and whose determinant has the form $\mathrm{e}^{\mathrm{i}\theta}$. $SU(n)$ is the **special** subgroup of $U(n)$ matrices with determinant 1.
- $GL(n, \mathbb{R})$, which has $SL(n, \mathbb{R})$ as a proper subgroup. is a proper subgroup of $GL(n, \mathbb{C})$,
- The subset $O(n)$ of the General Linear group $GL(n, \mathbb{R})$ of transformations represented by $n \times n$ real matrices which are **orthogonal** ($\mathbf{O}^{\mathrm{T}} \mathbf{O} = \mathbf{I}$), and whose determinant is 1 or -1. $SO(n)$ is the subgroup of $O(n)$ matrices with unit determinant, isomorphic to the **group of rotations** on $\mathbb{R}^n$.

Definition 2.13. Let H be a subgroup of G. For a given $x \in G$, the set $\{\, x \circ a_i \circ x^{-1} \,\}$, $\forall\, a_i \in H$, form a subgroup of G (EXERCISE) called the x**-conjugate** of H. Note that this is not the same as conjugacy classes!

It is straightforward to check that the subgroups of the four-group are self-conjugate for any x in the group. Conjugate subgroups are isomorphic (under conjugation). Indeed, if f in definition 2.8 is x-conjugation, then:

$$x \star (a_1 \star a_2) \star x^{-1} \;=\; (x \star a_1 \star x^{-1}) \;\star\; (x \star a_2 \star x^{-1})$$

As an example, we have the group of rotations of a body around, say, the x axis, and the group of its rotations around, say, the y axis. Both are subgroups of $SO(3)$ and are conjugate to one another. More precisely, any rotation around the x axis is equivalent to some rotation $R \in SO(3)$, followed by a rotation around the y axis, followed by R^{-1}.

New Notation:

1. It is now time to wean ourselves from the often cumbersome star (circle) notation for group composition. From now on, we shall omit them whenever there is little risk of confusion with ordinary multiplication.
2. If H and H' are two subsets of $\{G, \star\}$, often we shall write $H\,H' = H \star H'$ for $\{h\,h' = h \star h'\}$ $\forall\, h \in H,\ h' \in H'$.

Let us try out our new notation on the following definition:

Definition 2.14. A subgroup $N \subseteq G$ is called **invariant** (or **normal**) if $N = G\,N\,G^{-1}$. More precisely, if $g\,h\,g^{-1} \in N$ $\forall\, h \in N$ and $\forall\, g \in G$, N is an invariant subgroup of G. Alternate notation: $N \lhd G$, $G \rhd N$.

From this definition, we see that any invariant subgroup must be self-conjugate. As well, the identity and G itself are always invariant subgroups. EXERCISE: Show that the set $\{g_i^{-1} g_j^{-1} g_i g_j\}$ $\forall g_i,\, g_j \in G$, keeping only distinct objects, forms an invariant subgroup of G.

Definition 2.14 is sometimes written $G\,N = N\,G$, but it does not mean that an invariant subgroup must be Abelian (although it *can* be). Rather, it means that if you compose any element of N with any group element from the left, there is *some* element of N which, when composed with the group element from the right, gives the same result.

Definition 2.15. The **centre** of a group is the set of *all* its elements that commute with all elements of the group.

Example 2.1. From the composition table for the four-group, it is straightforward to work out that each of its non-trivial subgroups, $\{e,\, a\}$, $\{e,\, b\}$, $\{e,\, a\,b\}$, are invariant. Indeed, the four-group is Abelian, and all the subgroups of an Abelian group are invariant. Like all Abelian groups, the four-group is its own (trivial) centre.

A very powerful method for findin invariant subgroups is based on the following lemma : Let $H \subset G$. Then H *is invariant if and only if it contains complete classes*, ie. if it is a union of classes of G. Indeed, if H is invariant, *all* the conjugates (elements in the same class) of any $h \in H$ are also in H, and this holds for all classes, which are disjoint; so only complete classes can be in H. Conversely, if $\forall\, h \in H$ all the elements of the class of h are also in H, then, $\forall\, g \in G$, $g\, h\, g^{-1} \in H$, which is the definition of an invariant subgroup.

Definition 2.16. A **simple group** has no invariant subgroup other than itself and the identity.

2.2.3 Cosets and factor groups (BF 10.3)

Definition 2.17. Let H be a subgroup of G, and let $g \in G$. Then $g\,H$ is a **left coset** of H for a given g, and $H\,g$ is a **right coset** of H. The set of all left (right) cosets of H is called the **left (right) coset space** of H.

Every coset $g\,H$ of H must contain the same number of elements, equal to the order of H.

If H is invariant, to any ot its left cosets corresponds an identical right coset, and vice-versa, as follows immediately from Def. 2.14. In particular, the right and left cosets of any Abelian subgroup are identical.

Example 2.2. Let $G = \mathbb{R}^3$ under addition, and H be a plane containing the origin. For a given vector $\mathbf{a}$, $\mathbf{a} + H \in H$ if $\mathbf{a} \in H$; otherwise, $\mathbf{a} + H$ is another plane H' parallel to H, and we would say in this language that it is a left (or right) coset of the plane through the origin. And H itself would also be a coset.

The most important property of cosets is that they are either disjoint or else identical.

Indeed, let $g_1\, h_1 = g_2\, h_2$ for some $h_1,\ h_2 \in H$ and $g_1,\ g_2 \in G$. Therefore, $g_1 = g_2\, h_2\, h_1^{-1}$. Now consider some other element of the same coset, $g_1\, h_3$ ($h_3 \in H$); then $g_1 h_3 = g_2\,(h_2\, h_1^{-1}\, h_3) = g_2\, h_4$, where $h_4 = h_2\, h_1^{-1}\, h_3 \in H$. That is, if two elements of different cosets are the same, then any other element, say $g\, h_3$, in the first coset, must be equal to some element of the second coset. Since the same argument holds when we switch g_1 and g_2, we conclude that if $g_1\, H$ and $g_2\, H$ have one element in common, they have all their elements in common and are thus identical. The same proof applies to right cosets. Thus, we can say that the coset space of a subgroup $H \subset G$ provides a **partition** of G.

An immediate consequence of this property is that since $e\, H = H$ is a left coset of a subgroup H, that coset contains e and is the only coset that is a group. No other coset of H can be a group since it cannot contain e. For any other element h of subgroup H that is not the identity, we still have $h\, H = H$ because H is a group, and all cosets $h\, H = H$ are in fact the same one, ie. H itself.

If $H \subset G$, every element of G must occur either in H or one (and only one) of its other cosets.This forms the foundation of the proof of **Lagrange's Theorem**: *The order n of a finite group is an integer multiple of the order m of any of its subgroups.* Indeed, since every element of the group is either in the subgroup or in one of its other k distinct cosets, each with m elements, $(k+1)m = n$. The ratio n/m is called the **index** of the subgroup.

Let $a \in G$. Clearly, it *generates* a cyclic subgroup of G of order m: $\{e,\, a,\, a^2,\, \ldots,\, a^{m-1}\}$, where $m \leq n$ is the order of a. Therefore, the order n of G must be an integer multiple of the order m of any of its elements. If n is prime, $m = n$ or $m = 1$, and we have proved that the only non-trivial finite group of order prime is the cyclic group.

Now consider the set whose elements are the subgroup *as a whole* and all its other cosets, each also as a whole:

Definition 2.18. A **factor** space for $H \subset G$ is the set of all left cosets of H. Note that the elements of this space are the cosets themselves, each considered as a whole.

Factor spaces of a subgroup H are not necessarily groups; but there is one important exception:

Definition 2.19. To an invariant subgroup N of G is associated a **factor group** of G, G/N, whose elements are N and all its cosets. Again, N and its cosets themselves are elements of the factor group, not the elements of N or of the cosets. The order of G/N is the order of G divided by the order of N, ie. the index of N, hence the name **quotient group** often applied to G/N.

To show that the factor space of an invariant subgroup is a group, we note that for any coset $g\,N$,

$$(g\,N)\,N \;=\; g\,N\,N \;=\; g\,N; \qquad\qquad N\,(g\,N) \;=\; g\,N\,N \;=\; g\,N$$

where we have used the associativity of the group product and the invariant nature of N. This establishes that N must be the identity of the factor group. The composition law follows from:

$$(g_1\,N)\,(g_2\,N) \;=\; g_1\,g_2\,N\,N \;=\; (g_1\,g_2)\,N$$

since $g\,N = N\,g \;\; \forall\, g \in G$, Lastly,

$$(g\,N)\,(g^{-1}\,N) \;=\; g\,g^{-1}\,N\,N \;=\; N \;=\; e$$

So $g^{-1}\,N$ is the inverse of $g\,N$.

Factor groups can be useful when, for one reason or another, we do not need to distinguish between the elements of subgroups of a group.

2.2.4 Direct Products

Definition 2.20. Let H_1 and H_2 be subgroups of G which have only the identity e as common element, and let $h_1\,h_2 \;=\; h_2\,h_1 \;\forall\, h_1 \in H_1, \forall\, h_2 \in H_2$. If it is *possible* to write $g \;=\; h_1\,h_2 \;\forall\, g \in G$, then $G \equiv H_1 \times H_2$ is said to be the **internal direct product** of its subgroups H_1 and H_2. The latter are invariant (EXERCISE).

Example 2.3. We can think of $O(3)$ as the direct product of $SO(3)$ and the subgroup consisting of the identity matrix and the reflection matrix $-\mathbf{I}$. This latter group is obviously Abelian and, since its elements commute with all 3-dim rotations (ie. the elements of $SO(3)$), it is an invariant subgroup and the centre of $O(3)$. Also, the four-group introduced in section 2.1.2 can be seen as $Z_2 \times Z_2$, or $\{e,\,a\} \times \{e,\,b\} = \{e,\,a,\,b,\,a\,b\}$. However, $Z_4 \neq Z_2 \times Z_2$, even though Z_2 is an invariant subgroup of Z_4, and thus $Z_2 = Z_4/Z_2$!.

Another well-known way of constructing a (this time, **external**) direct product of, say, two a priori unrelated matrix groups with elements $\mathbf{A} \in H_1$ and $\mathbf{B} \in H_2$ would be:

$$\begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix}$$

Or we could *construct* $\{1,\,-1\} \times \{1,\,-1\} = \{(1,\,1),\,(1,\,-1),\,(-1,\,1),\,(-1,\,-1)\}$. the external direct product of Z_2 with itself, in this realisation. This, of course, is the four-group.

2.3 The Mother of All Finite Groups: the Group of Permutations

2.3.1 Definitions, cycles, products

The most important finite group is the **group of permutations** of n objects, S_n, aka the **symmetric group**, which contains $n!$ elements corresponding to the $n!$ possible rearrangements of the objects. A permutation is by definition a bijective mapping. Following a standard convention, we notate, with $1 \leq k \leq n!$:

Definition 2.21.

$$\pi_k = \begin{pmatrix} 1 & 2 & 3 & \dots & n \\ \pi_k(1) & \pi_k(2) & \pi_k(3) & \dots & \pi_k(n) \end{pmatrix}$$

The horizontal ordering of the initial objects is immaterial. Also as a matter of convention, we agree that it is the objects in the slots which are rearranged, not the slots. It is quite possible to define $\pi_k(j)$ as the mapping of *whatever* object happens to sit in slot j; but, again following standard usage, eg. in BF, we define $\pi_k(j)$ as the mapping, *in a given slot*, of the *object* labelled by j to another object labelled by a number between 1 and n. Finally, we do not have to use numbers as labels, but they offer the greatest range.

In a permutation, an object i may be mapped into itself, ie. it stays in the same slot. But more typically object i is mapped to j, while j is mapped to k; and so on along a chain that ends back at object a after l steps. When this occurs, we speak of a l**-cycle**. More precisely:

Definition 2.22. Let $\pi_k \in S_n$, and let l be the smallest integer for which $[\pi_k(j)]^l = j$, for some $1 \leq j \leq n$. Then the sequence of objects in $[\pi_k(j)]^l$ is called a l-cycle (sometimes a r-cycle...).

This suggests a much more compact notation for π_k, one in which we bother to write only the l-cycles ($l > 1$), and consider a given permutation as the product of simpler permutations.

As an example, we write:

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 5 & 4 & 2 & 3 & 1 & 6 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 5 & 2 & 3 & 4 & 1 & 6 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 4 & 2 & 3 & 5 & 6 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 2 & 3 & 4 & 5 & 6 \end{pmatrix} \equiv (1\,5)\,(2\,4\,3)$$

It is easy to see the advantages of the cycle notation introduced at the end of the line! *Note that the cycles are disjoint*. Any permutation can be, and usually is, represented by a sequence of disjoint cycles.

Insight: if one imagines n states of a system being evolved through discrete time increments by successive applications of S_n, the cyclic structure means that not all states are accessible from a given initial state, only those allowed by the cycle to which the initial state belongs. The time evolution of the system is therefore constrained. We could assign some quantity that takes some value corresponding to each particular cycle. Then we could say that this quantity is conserved, in the sense that the time evolution of an initial state is constrained to states that have the same value of the conserved quantity, because they are the only ones accessible from the initial state. Perhaps this gives us an inkling of a connection between groups and conservation laws.

Any $\pi_k \in S_n$ can always be written as the product[†] of **transpositions**, or two-cycles. Indeed, a l-cycle may always be decomposed as a product of $l - 1$ transpositions, but these are not disjoint.

Definition 2.23. A permutation is **even** (**odd**) if it is equivalent to an even (odd) number of transpositions, or switches; thus, a l-cycle which contains an even number of symbols is equivalent to an odd permutation, and vice-versa. An even permutation is said to have **parity** 1, and an odd permutation parity -1. We expect that parity will put strong constraints on the group product table of S_n.

Transpositions always have odd parity. The mapping from S_n to the parities $\{1, -1\}$ is a nice example of a homomorphism.

[†] Since there is little scope for confusion in the context of S_n, we replace "group composition" with "group product".

Definition 2.24. A permutation with only one cycle of length $l > 1$ is a **cyclic permutation** of length l.

In cycle notation, $S_2 = \{e,\ (1\,2)\}$ and $S_3 = \{e,\ (1\,2),\ (1\,3),\ (2\,3),\ (1\,3\,2),\ (1\,2\,3)\} \equiv \{\pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6\}$, are the smallest non-trivial symmetric groups. For S_3, note the three-cycles $(1\,2\,3) = \left(\begin{smallmatrix}1&2&3\\2&3&1\end{smallmatrix}\right)$ and $(1\,3\,2) = \left(\begin{smallmatrix}1&2&3\\3&1&2\end{smallmatrix}\right)$. I have deliberately changed the order of the latter from what it is in BF, but if you write out the corresponding permutation in full notation for BF's $(3\,2\,1)$, you will see that it is identical to mine. So long as we cycle through in the same direction (here, to the right), the order of the elements in a cycle does not matter!

Warning: do not confuse the symbols in a 3-cycle with the *outcome* of a permutation in S_3!

2.3.2 Subgroups of S_n

One obvious subgroup of S_n is the so-called **alternating group**, A_n, of all its even permutations. Odd permutations cannot form a group, because their product is always an even permutation. Another, less obvious, but important subgroup of S_n is the cyclic group of order n, generated by the permutation $(1\,2\,3\,\ldots\,n)$, which returns the initial state to itself after n products.

Now for subgroups of S_3: Lagrange's Theorem allows only non-trivial proper subgroups of order 2 or 3. The alternating subgroup A_3 is read off the list of the elements of S_3: $\{e,\ (1\,3\,2),\ (1\,2\,3),\}$ which must be cyclic because all groups of order 3 are isomorphic to Z_3. Note: this is not a general feature as the cyclic subgroups of higher order generated by odd permutations in $S_{n>3}$ contain permutations of both even and odd parity.

The product of a transposition by itself is the identity, so $\pi_2\,\pi_2 = \pi_3\,\pi_3 = \pi_4\,\pi_4 = e$. Then the other (isomorphic!) subgroups of S_3 are $\{e, \pi_2\}$, $\{e, \pi_3\}$, and $\{e, \pi_4\}$. The centre of S_3 —and of $S_{n>3}$ — is just e.

2.3.3 Cayley table of S_n

The group-product table of S_3 contains 36 entries, "only" 25 of which are non-trivial, from which we have just found three. But I claim that no more than one other product needs to be worked out with the explicit form of the permutations; the rest can all be found by a bit of astute reasoning.

The entries of the 2×2 sub-table for rows and columns corresponding to π_5 and π_6 are determined by the fact that they must be even permutations (since they are the group product of even permutations). The diagonals cannot be e; if they were, the other entry in the same row would be the same as the the first entry of its column. Alternatively, π_5 and π_6 being the only non-trivial elements in A_3, they must be each other's inverse.

Next, the non-diagonal elements of rows and columns corresponding to π_2, π_3 and π_4 must be either π_5 or π_6, the only even permutations other than e. To fill in this sector only requires calculating one group product, say, $\pi_2\,\pi_3$:

$$\pi_2\,\pi_3 \;=\; \begin{pmatrix}1&2&3\\2&1&3\end{pmatrix}\begin{pmatrix}1&2&3\\3&2&1\end{pmatrix} \;=\; \begin{pmatrix}1&2&3\\3&2&1\\3&1&2\end{pmatrix} \;=\; \begin{pmatrix}1&2&3\\3&1&2\end{pmatrix} \;=\; \pi_5$$

The other unfilled entries in rows and columns for π_5 and π_6 must be either π_2, π_3, or π_4. For columns π_5 and π_6, applying π_2 to $\pi_2\,\pi_3$ gives $\pi_3 = \pi_2\,\pi_5$, which determines the rest from the general table-building rules. Similarly, $\pi_2\,\pi_3\,\pi_3 = \pi_2 = \pi_5\,\pi_3$, and the other entries in the π_5 and π_6 rows are determined. Here is the final result in two equivalent forms:

e	π_2	π_3	π_4	π_5	π_6
π_2	e	π_5	π_6	π_3	π_4
π_3	π_6	e	π_5	π_4	π_2
π_4	π_5	π_6	e	π_2	π_3
π_5	π_4	π_2	π_3	π_6	e
π_6	π_3	π_4	π_2	e	π_5

$\equiv$

e	π_5	π_6	π_2	π_3	π_4
π_5	π_6	e	π_4	π_2	π_3
π_6	e	π_5	π_3	π_4	π_2
π_2	π_3	π_4	e	π_5	π_6
π_3	π_4	π_2	π_6	e	π_5
π_4	π_2	π_3	π_5	π_6	e

2.3.4 Cayley's Theorem

Why is S_n so important? As so often, the group-product table of a group G of order n, gives the key to the answer. $\forall\, a_i \in G$, the row $\{a_i\ a_j\}$ $(1 \le j \le n)$ is merely *a bijective rearrangement of* $\{a_i\}$, that is:

$$a_i \longmapsto \pi_{a_i} = \begin{pmatrix} a_1 & a_2 & \dots & a_n \\ a_i\,a_1 & a_i\,a_2 & \dots & a_i\,a_n \end{pmatrix}, \qquad a_i\,a_j \longmapsto \pi_{a_i a_j} = \begin{pmatrix} a_1 & a_2 & \dots & a_n \\ a_i\,a_j\,a_1 & a_i\,a_j\,a_2 & \dots & a_i\,a_j\,a_n \end{pmatrix}$$

But we can also write:

$$\begin{aligned} \pi_{a_i} &= \begin{pmatrix} a_1 & a_2 & \dots & a_n \\ a_i\,a_1 & a_i\,a_2 & \dots & a_i\,a_n \end{pmatrix} = \begin{pmatrix} a_j\,a_1 & a_j\,a_2 & \dots & a_j\,a_n \\ a_i\,(a_j\,a_1) & a_i\,(a_j\,a_2) & \dots & a_i\,(a_j\,a_n) \end{pmatrix} \\ \Longrightarrow \qquad \pi_{a_i}\,\pi_{a_j} &= \begin{pmatrix} a_j\,a_1 & a_j\,a_2 & \dots & a_j\,a_n \\ a_i\,a_j\,a_1 & a_i\,a_j\,a_2 & \dots & a_i\,a_j\,a_n \end{pmatrix} \begin{pmatrix} a_1 & a_2 & \dots & a_n \\ a_j\,a_1 & a_j\,a_2 & \dots & a_j\,a_n \end{pmatrix} \\ &= \begin{pmatrix} a_1 & a_2 & \dots & a_n \\ a_i\,a_j\,a_1 & a_i\,a_j\,a_2 & \dots & a_i\,a_j\,a_n \end{pmatrix} \end{aligned}$$

What we have shown is that $\pi_{a_i}\,\pi_{a_j} = \pi_{a_i a_j}$; in other words, by definition 2.8, permutations preserve the group product of G, and we have **Cayley's Theorem**:

Every group of order n is isomorphic to a subgroup of S_n whose elements (except for e) shuffle ***all*** *objects in the set on which it acts*. By Lagrange's theorem, the index of a finite group as a subgroup of S_n is $n!/n = (n-1)!$.

We have already seen an example of this: the cyclic group of order 3 is a subgroup of S_3, and there is only one instance of $Z_3 \subset S_3$. EXERCISE: How many distinct instances of $Z_4 \subset S_4$ are there? How many of the four-group?

Another interesting example of a relation between a group and S_n is a permutation $\pi \in S_3$ which defines an action of S_3 on $\mathbb{R}^3$ which permutes the coordinates x^i of a vector, $x^i\pi = x^{\pi i}$. Letting π act from the right, we have, for instance, $(x_1,\ x_2,\ x_3) \star \pi_5 = (x_3,\ x_1,\ x_2)$. Since we act from the right[†], the vectors are columns; with implicit summation over j:

$$x^i\,\pi_5 = T^i{}_j(\pi_5)\,x^j \qquad T^i{}_j(\pi_5) = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \qquad x^i\,\pi_6 = T^i{}_j(\pi_6)\,x^j \qquad T^i{}_j(\pi_6) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

Note that $\pi_5\,\pi_6 = e$, so that π_5 and π_6 are the inverse of one another. This is reflected in the T matrices: they are each other's inverse and transpose, and they have determinant $+1$, so they are elements of $SO(3)$. EXERCISE: Is this true of $T^i{}_j(\pi_{2,3,4})$?

2.3.5 Conjugates and Classes of S_n

To find the classes of S_n, we must form, for each $\pi_i \in S_n$, all its conjugates $\pi_j\,\pi_i\,\pi_j^{-1}$. This seemingly daunting task can actually be performed fairly easily, thanks to the nature of S_n. To keep the following manipulations as uncluttered as possible, let us write $\pi_i = a$ and $\pi_j = b$, with $a = \left(\begin{smallmatrix} 1 & 2 & \dots & n \\ a_1 & a_2 & \dots & a_n \end{smallmatrix}\right)$ and $b = \left(\begin{smallmatrix} 1 & 2 & \dots & n \\ b_1 & b_2 & \dots & b_n \end{smallmatrix}\right)$. Then:

$$\begin{aligned} b\,a\,b^{-1} &= \begin{pmatrix} 1 & 2 & \dots & n \\ b_1 & b_2 & \dots & b_n \end{pmatrix} \begin{pmatrix} 1 & 2 & \dots & n \\ a_1 & a_2 & \dots & a_n \end{pmatrix} \begin{pmatrix} b_1 & b_2 & \dots & b_n \\ 1 & 2 & \dots & n \end{pmatrix} = \begin{pmatrix} 1 & 2 & \dots & n \\ b_1 & b_2 & \dots & b_n \end{pmatrix} \begin{pmatrix} b_1 & b_2 & \dots & b_n \\ a_1 & a_2 & \dots & a_n \end{pmatrix} \\ &= \begin{pmatrix} a_1 & a_2 & \dots & a_n \\ b_{a_1} & b_{a_2} & \dots & b_{a_n} \end{pmatrix} \begin{pmatrix} b_1 & b_2 & \dots & b_n \\ a_1 & a_2 & \dots & a_n \end{pmatrix} \\ &= \begin{pmatrix} b_1 & b_2 & \dots & b_n \\ b_{a_1} & b_{a_2} & \dots & b_{a_n} \end{pmatrix} \end{aligned} \tag{2.1}$$

[†]We are getting a little bit ahead of ourselves here. If we acted from the left, the vectors would be row-vectors, and we would have $x^i\pi = x^j\,T_j{}^i(\pi)$, with π_5 and π_6 interchanged. The difference between the right and left actions of a group will be discussed in the next section.

How did we obtain $\begin{pmatrix} a_1 & a_2 & \dots & a_n \\ b_{a_1} & b_{a_2} & \dots & b_{a_n} \end{pmatrix}$ in the second line from $\begin{pmatrix} 1 & 2 & \dots & n \\ b_1 & b_2 & \dots & b_n \end{pmatrix}$ in the last member of the first line? By noting that a_1 must occur in some slot on the top line of the latter, and using the arbitrariness in the order of the columns in the permutation to move that column to first position, and then renaming the upper element a_1. Then we do the same for $2 \to a_2$, etc. The bottom elements are then the outcome of permuting a_i with permutation b to get b_{a_i}.

Something very important can be deduced from this result: *All permutations in a class have the same cycle structure, not only for S_n, but for all finite groups because of Cayley's theorem*. Since classes are disjoint, classifying the elements of S_n according to their cycle structure also uniquely yields the classes of S_n! In groups other than S_n, although all elements in a class have the same cycle structure, *elements with the same cycle structure may belong to different classes* (eg. $A_4 \subset S_4$)...

Take S_3 as a simple example. As classes we only have $\mathcal{C}_1 = \{e\}$, $\mathcal{C}_2 = \{(1\,2),\,(1\,3),\,(2\,3)\}$, and $\mathcal{C}_3 = \{(1\,3\,2),\,(1\,2\,3)\}$. Thus, because it is the only subgroup of S_3 (apart form $\{e\}$ and S_3 itself) that is the sum of *complete* classes, and whose order divides the order of S_3 (6), $\mathcal{C}_1 + \mathcal{C}_3 = \{e,\,(1\,2\,3),\,(1\,3\,2)\} = A_3$ is the only invariant[†] proper subgroup of S_3.

Now consider S_4. There are two other permutations with the same cycle structure as $(1\,2)(3\,4)$: $(1\,3)(2\,4)$ and $(1\,4)(2\,3)$. Apart from this and the separate class $\{e\}$, the other classes of S_4 are easily obtained as $(1\,2)$ and its 5 conjugate transpositions, $(1\,2\,3)$ and its seven conjugates, and $(1\,2\,3\,4)$ and its five conjugates. All five classes are of course disjoint.

In the literature, classes of S_n are routinely identified by **partitions** of n reflecting their cycle structure. Thus, a given class will be written $(i^{\alpha_i}\ \dots\ j^{\alpha_j})$, with $(1 \le (i,\ j) \le n)$, where α_i indicates the number of i-cycles for the class.

Start with e, which is always a class of S_n. Its cycle structure can be written as a product of n 1-cycles: $e = (1)\,(2)\,\cdots\,(n)$. So the class would be denoted in this notation by 1^n. A transposition must have one 2-cycle and $n-2$ 1-cycles, and S_n must contain $n(n-1)/2$ of them (eg., six for S_4 as above); it is denoted by $(2\,1^{n-2})$. An arbitrary permutation involves α_i i-cycles, and $\sum_i^n i\,\alpha_i = n$. It is in that sense that we say that the cycle structure of a class corresponds to a partition of n.

Once we have noticed this correspondence, it becomes rather easy to find how many classes there are in S_n and what their cycle structure is. We adopt the usual convention that represents the cycle structure of a class by $(\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n)$, where the λ_i must sum up to n. Thus, the only possible partitions of S_3 would correspond to classes (1^3), $(2\,1)$, and (3), ie. a class with three 1-cycles (the identity), a class with one 2-cycle and one 1-cycle (the transpositions), and a class with one 3-cycle. As for S_4, the only possible partitions of 4 give rise to the five classes (1^4), $(2\,1^2)$, (2^2), $(3\,1)$, and (4).

The number of elements in a class of S_n is the number of *distinct* ways of partitioning n numbers into the cycle structure of the class:

$$\frac{n!}{\alpha_1!\dots\alpha_n!\,1^{\alpha_1}\dots n^{\alpha_n}} \tag{2.2}$$

where $\alpha_i!$ is the number of non-distinct ways of ordering α_i commuting cycles, and i^{α_i} is the number of non-distinct orderings of the symbols inside an i-cycle. From this expression it should be easy to recover the number of elements in each class of S_4 as given above.

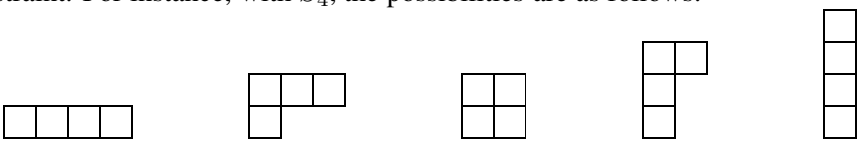
Now we can identify (EXERCISE) the invariant subgroups of S_4 without ever writing down its 24×24 composition table!

2.3.6 Graphical representation of classes: Young frames

A graphical way of representing the classes of S_n is to take n squares and arrange them in rows and columns so that each column corresponds to an i-cycle and the number of boxes cannot increase from one column to the next

[†]Note that this subgroup being Abelian is not sufficient to establish that it is invariant; it must be self-conjugate with respect to *all* elements in S_3.

on the right, and from one row to the one next below. The game then consists in building all possible arrangements that satisfy this constraint. For instance, with S_4, the possibilities are as follows:



Then we just read off the cycle structure for each: (1^4), $(2\,1^2)$, (2^2), $(3\,1)$, and (4), respectively. Suddenly, finding the classes of such monsters as, say S_8, no longer seems so intimidating. These diagrams are known as **Young frames**.

2.3.7 Cosets of S_n

Now that we know the Cayley table of S_3, finding the left cosets of its subgroups is as easy as reading column elements in the table. For the subgroup $\{e,\,\pi_2\}$, the cosets are $\pi_k\,\{e,\,\pi_2\} = \{\pi_k,\,\pi_k\,\pi_2\}$ $(1 \le k \le 6)$. So each coset consists of the first two entries in each row of the table. There are thus only three *distinct* cosets: $\{e,\,\pi_2\}$, $\{\pi_3,\,\pi_6\}$, $\{\pi_4,\,\pi_5\}$. As stated earlier in Definition 2.18, this set of cosets is the factor space for $\{e,\,\pi_2\} \subset S_3$, which can be written as the partition $e\,\{e,\,\pi_2\} + \pi_3\,\{e,\,\pi_2\} + \pi_4\,\{e,\,\pi_2\}$. But the right cosets (EXERCISE) are not the same as the left cosets, since none of these three subgroups is invariant, ie. $g\,H\,g^{-1} \neq H\;\forall\, g \in S_3$ since here H contains one odd permutation and g can be even.

Turn now to the remaining non-trivial proper subgroup of S_3, $A_3 = \{e,\,\pi_5,\,\pi_6\}$, of all even permutations in S_3. Its left cosets are $\{\pi_k, \pi_k\,\pi_5, \pi_k\,\pi_6\}$. For instance, by inspection of the group-product table of S_3, $\pi_2\,\{e,\,\pi_5,\,\pi_6\} = \{\pi_2,\,\pi_3,\,\pi_4\}$, which is identical to the other cosets $\pi_3\,\{e,\,\pi_5,\,\pi_6\}$ and $\pi_4\,\{e,\,\pi_5,\,\pi_6\}$. Also, $e\,\{e, \pi_5, \pi_6\} = \pi_5\{e, \pi_5, \pi_6\} = \pi_6\{e, \pi_5, \pi_6\}$, as expected. So another partition of S_3 is provided by $\{e,\,\pi_5,\,\pi_6\} + \pi_2\,\{e,\,\pi_5,\,\pi_6\}$. Note that the left and right cosets are now identical, another way of saying that $\{e,\,\pi_5,\,\pi_5\}$ is invariant, as we had found through simpler means. Then $\big\{\,\{e,\,\pi_5,\,\pi_6\}, \{\pi_2,\,\pi_3,\,\pi_4\}\,\big\}$ is the factor group of S_3 . From the composition table for S_3, we see that the element $\{e,\,\pi_5,\,\pi_6\}$ is the identity, and that this factor group S_3/A_3 is in fact isomorphic to Z_2 and can be identified with it. It is easy to show that Z_2 is a factor group of $S_n\;\forall\, n$. Equivalently, A_n is always a normal subgroup of S_n.

2.4 Representations of Groups

We have already mentioned that groups can be associated with symmetries, but we have to make this connection explicit in the language of group theory. Doing so will help us flesh out the rather abstract ideas and tools we have introduced. We shall find that linear operators on vector spaces (most often, operators on a Hilbert space) provide us with this connection.

2.4.1 What is a symmetry?

Let G be a group of linear transformations that act on some $\mathbf{x} \in \mathbb{R}^n$. Let us also give ourselves functions $f(\mathbf{x})$ that are square-integrable, ie., that live in a Hilbert space, $\mathcal{L}^2 \equiv \mathcal{H}$, which we call the **carrier** space.

> **Definition 2.25.** Let $\mathrm{g} \in G$. We define an **action from the left**, $[T_\mathrm{g} f](\mathbf{x}) = f(\mathrm{g}^{-1}\mathbf{x})$, and an **action from the right**, $[T_\mathrm{g} f](\mathbf{x}) = f(\mathbf{x}\,\mathrm{g})$, $\forall\, f$. The set of operators $\{T_\mathrm{g}\}$ introduced in this way and which act on the *functions* themselves, is also a group.

Why did we define the left action of a group G as $\mathrm{g}^{-1}\,\mathbf{x}$, and not $\mathrm{g}\,\mathbf{x}$? Surely, since g and its inverse are both elements of the same group, it should not matter which we use. "Left" really only makes sense in relation to "right", so that when we define the right action of a group, we have to use the inverse of the element as written in the left action. But we could always use g for any one action so long as we use g^{-1} for the other... couldn't we?

Well, let us check whether the $\{T_\mathrm{g}\}$ do form a group with, say, group action from the right defined as $\mathbf{x}\,\mathrm{g}$, or, alternatively, as $\mathbf{x}\,\mathrm{g}^{-1}$. Let us denote by $T_{\mathrm{g}_i\mathrm{g}_j}$ the transformation associated with the group product pair $\mathrm{g}_i\,\mathrm{g}_j \in G$.

Then, with $\mathrm{g}_i = i$ and $\mathrm{g}_j = j$ in subscripts so as to lighten up the formalism:

$$\big[T_{ij}\, f\big](\mathbf{x}) \;=\; f(\mathbf{x}\,\mathrm{g}_i\,\mathrm{g}_j) \;=\; \big[T_j\, f\big](\mathbf{x}\,\mathrm{g}_i) \;=\; \big[T_i T_j\, f\big](\mathbf{x})$$

which means that the T operators form a group; but what if instead:

$$\big[T_{ij}\, f\big](\mathbf{x}) \;=\; f(\mathbf{x}\,(\mathrm{g}_i\,\mathrm{g}_j)^{-1}) \;=\; f(\mathbf{x}\,\mathrm{g}_j^{-1}\,\mathrm{g}_i^{-1}) \;=\; \big[T_i\, f\big](\mathbf{x}\,\mathrm{g}_j^{-1}) \;=\; \big[T_j T_i\, f\big](\mathbf{x})$$

Something awkward has happened: if we *write* the right action as $\mathbf{x}\,\mathrm{g}^{-1}$, the associated transformations do *not* form a group! And, as you should verify, neither do they if we write the left action as $\mathrm{g}\,\mathbf{x}$.

So, as a matter of *notational* consistency, we should always write $\mathbf{x}\,\mathrm{g}$ and $\mathrm{g}^{-1}\,\mathbf{x}$, which is indeed what BF do (without much explanation).

Now let there be a linear operator $A_{\mathbf{x}}$ such that, $\forall\, f \in \mathcal{H}$, $[A_{\mathbf{x}}\, f](\mathbf{x}) \;=\; h(\mathbf{x})$, where $h \in \mathcal{H}$. Then we transform $A_{\mathbf{x}}$ under G in the following way: $T_{\mathrm{g}}\, A_{\mathbf{x}}\, T_{\mathrm{g}}^{-1}$.

Definition 2.26. When $T_{\mathrm{g}}\, A_{\mathbf{x}}\, T_{\mathrm{g}}^{-1} \;=\; A_{\mathbf{x}}$, $\forall\, \mathrm{g} \in G$, $A_{\mathbf{x}}$ is said to be **invariant under the action of the group**. If also $T_{\mathrm{g}} f(\mathbf{x}) = f(\mathbf{x})$, we often say that f is invariant under G itself as well.

Since the condition for invariance can also be written as $T_{\mathrm{g}}\, A_{\mathbf{x}} = A_{\mathbf{x}}\, T_{\mathrm{g}}$, $\forall\, \mathrm{g} \in G$, then an operator that is invariant under a group of operators must commute with all the operators in that group.

Example 2.4. The group of translations

Let $\mathbf{a} \in \mathbb{R}^3$. Then $\{g = T_{\mathbf{a}}\}$, where $T_{\mathbf{a}}\mathbf{x} = \mathbf{x} - \mathbf{a}$, form an Abelian group since $T_{\mathbf{a}}\, T_{\mathbf{b}} = T_{\mathbf{b}}\, T_{\mathbf{a}} = T_{\mathbf{a}+\mathbf{b}}$. Also, let f be an analytic function. Then:

$$f(T_{\mathbf{a}}^{-1}\mathbf{x}) \;=\; f(\mathbf{x}+\mathbf{a}) \;=\; \sum_{n=0}^{\infty} \frac{1}{n!} \big[\mathbf{a}\cdot\boldsymbol{\nabla}\big]^n f(\mathbf{x}) \;=\; [\mathrm{e}^{\mathbf{a}\cdot\boldsymbol{\nabla}}\, f](\mathbf{x})$$

We identify the translation operator: $\mathcal{T}_{\mathbf{a}} = \mathrm{e}^{\mathbf{a}\cdot\boldsymbol{\nabla}}$; indeed, $[\mathcal{T}_{\mathbf{a}} f](\mathbf{x}) = f(T_{\mathbf{a}}^{-1}\mathbf{x}) = f(\mathbf{x}+\mathbf{a})$. The operator $\boldsymbol{\nabla}$ is called the **infinitesimal generator** of translations.

In its usable Taylor-expansion form, $\mathrm{e}^{\mathbf{a}\cdot\boldsymbol{\nabla}}$ plainly commutes with the Laplacian ∇^2, because all derivatives commute. So the Laplacian is invariant under the group of translations.

If $A_{\mathbf{x}}$ has eigenvalues and eigenfunctions. and if it is invariant under G, then there should exist a set of functions $\{f^i\}$ such that:

$$\big[A_{\mathbf{x}}\, T_{\mathrm{g}}\, f^i\big](\mathbf{x}) \;=\; \big[T_{\mathrm{g}}\, A_{\mathbf{x}}\, f^i\big](\mathbf{x}) \;=\; \lambda_{\mathrm{g}}\, [T_{\mathrm{g}}\, f^i](\mathbf{x})$$

which says that if f is an eigenfunction of $A_{\mathbf{x}}$, so is $T_{\mathrm{g}}\, f^i$, with *the same eigenvalue*. If the eigenvalue is non-degenerate, ie. if f is unique, then $T_{\mathrm{g}}\, f^i$ must be proportional to f^i, ie., f^i is also an eigenfunction of T_{g}, but with some a priori different eigenvalue also depending on g. In the degenerate case, however, given N eigenfunctions, all we can say is that the $T_g\, f^i$ are linear combinations of $\{f^j\}$:

$$T_{\mathrm{g}}\, f^i \;=\; f^j\, D_j{}^i(\mathrm{g}) \tag{2.3}$$

with summation over repeated indices implied.

2.4.2 Matrix representations of a group (BF10.4)

Definition 2.27. A **representation** D of a group G is a homomorphic mapping onto a set of *finite*-dimensional invertible matrices such that $\mathsf{D}(e) = \mathsf{I}$, the identity matrix, and $\mathsf{D}(\mathrm{g}_i)\,\mathsf{D}(\mathrm{g}_j) = \mathsf{D}(\mathrm{g}_i\,\mathrm{g}_j)$, in the sense that matrix multiplication preserves the group composition law.

If the homomorphism is one-to-one, a representation is **faithful**. The dimension of the representation is the dimension of its matrices or, equivalently, the dimension of the carrier space on which it acts.

Whenever we find a set of degenerate eigenfunctions for some operator that is invariant under a group G, we expect to be able to connect these functions to a representation of the group.

Matrix representations arise in a much more general context than symmetry. The matrices $GL(n,\mathbb{C})$ of rank n can be thought of as the set of all invertible linear transformations of a vector space of complex-valued functions $\mathcal{V} = \{f(\mathbf{x})\}$, where $\mathbf{x} \in \mathbb{R}^n$. If $\{\mathbf{e}_\mathrm{i}\}$ is a basis for $\mathcal{V}$, then $\mathbf{x} = x^i\,\mathbf{e}_\mathrm{i}$, where the x^i are the components of $\mathbf{x}$ in the basis, and the subscript i on the basis vectors specifies which vector, not which component of the vector.

Now let us simplify things a bit by taking $f(\mathbf{x}) = \mathbf{x}$. Then the **left** action of an element $\mathrm{g} \in G$, expressed in terms of the linear transformations T_g, *must* be written as:

$$T_\mathrm{g}(\mathbf{x}) \;=\; \mathrm{g}^{-1}\,\mathbf{x} \;=\; x^i\,\mathrm{g}^{-1}\,\mathbf{e}_\mathrm{i} \;=\; x^i\left(\mathbf{e}_\mathrm{j}\,D^j{}_i(\mathrm{g}^{-1})\right) \;=\; \mathbf{e}_\mathrm{j}\left(D^j{}_i(\mathrm{g}^{-1})x^i\right) \tag{2.4}$$

Only this exact definition of the associated D matrices preserves the group product of G. Indeed:

$$\mathrm{g}_1^{-1}\,\mathrm{e}_\mathrm{i} \;=\; \mathrm{e}_\mathrm{j}\,D^j{}_i(\mathrm{g}_1^{-1})$$

$$\mathrm{g}_2^{-1}\,\mathrm{g}_1^{-1}\,\mathbf{e}_\mathrm{i} \;=\; \mathrm{g}_2^{-1}\,\mathbf{e}_\mathrm{j}\,D^j{}_k(\mathrm{g}_1^{-1}) \;=\; \mathbf{e}_\mathrm{k}\,D^k{}_j(\mathrm{g}_2^{-1})\,D^j{}_i(\mathrm{g}_1^{-1})$$

Comparing with $\mathrm{g}_2^{-1}\,\mathrm{g}_1^{-1}\,\mathrm{e}_\mathrm{i} \;=\; \mathbf{e}_\mathrm{k}\,D^k{}_i(\mathrm{g}_2^{-1}\,\mathrm{g}_1^{-1})$, we see that $\mathsf{D}(\mathrm{g}_2^{-1}\,\mathrm{g}_1^{-1}) = \mathsf{D}(\mathrm{g}_2^{-1})\,\mathsf{D}(\mathrm{g}_1^{-1})$, or $\mathsf{D}(\mathrm{g}_1\,\mathrm{g}_2) = \mathsf{D}(\mathrm{g}_1)\,\mathsf{D}(\mathrm{g}_2)$, as required for the D matrices to have the same product rule as the group. This is perfectly consistent with eq. (2.3) above, but now we know that eq. (2.3) corresponded to the left action of the group, $\mathrm{g}^{-1}\,f^i$, which was not so obvious because of the use of the T_g operators which always act from the left.

It is an instructive exercise to show that the proper way of expressing the right action of the same group, $\mathbf{x}\,\mathrm{g}$, in terms of its (right) representation D matrices is:

$$\mathbf{x}\,\mathrm{g} \;=\; \mathbf{e}_\mathrm{i}\,x^i\,\mathrm{g} \;=\; \left(D^i{}_j(\mathrm{g})x^j\right)\mathbf{e}_\mathrm{i} \tag{2.5}$$

in which D acts on the the x^i written as a *column* vector. Because of this, some people see the right action as the more "natural" one. For a given g, the right D matrices are in general different than the left ones.

2.4.3 Non-unicity of group representations

One might hope to define an algorithm that would churn out *the* representation of a group. But there is no such thing as a unique representation! Indeed, suppose we have a set of n-dimensional matrices which represent a group. It is always possible to obtain another representation, also of dimension n, by mapping these matrices to the identity matrix. This is called the **identity representation**, and it always exists. Also, the homomorphic map of the same matrices to their determinant preserves the group product (since $\det(AB) = (\det A)(\det B)$, which provides another representation which this time is one-dimensional. Of course, nobody claims that such representations are faithful…

Also, we can make a change of basis: $\mathbf{e}'_i \;=\; \mathbf{e}_j\,S^j{}_i$, or $\mathbf{e}_i \;=\; \mathbf{e}'_j\,(S^{-1})^j{}_i$. Then we have the **similarity** transformation: $\mathsf{D}'(\mathrm{g}) \;=\; \mathsf{S}\,\mathsf{D}(\mathrm{g})\,\mathsf{S}^{-1}$, and the D' obey the same product rules as the D matrices.

> **Definition 2.28.** Representations connected by a similarity transformation are said to be **equivalent**. They differ only by a choice of basis.

Example 2.5. Consider the continuous group, called $SO(2)$, of rotations around the z axis embedded into the group of three-dimensional rotations. We focus on its left action and look for representations.

We parametrise a rotation by $g = R_\alpha$ such that $R_\alpha\,\phi = \phi - \alpha$. This would correspond to rotating the standard basis in $\mathbb{R}^3$ by α (often called a passive transformation), with the coordinates of a vector characterised by angle ϕ being mapped to the same coordinates of that vector rotated by $-\alpha$ in the initial basis. One method for finding representations is to use eq. (2.4), which becomes in this context:

$$\big[\mathcal{R}_\alpha f_i\big](\phi) \;=\; f_i\big[R_\alpha^{-1}\,\phi\big] \;=\; f_i(\phi+\alpha) \;=\; D_i{}^j(-\alpha)\,f_j(\phi)$$

We want to find a set of functions which transform into linear combinations of themselves under $\mathcal{R}_\alpha$. Try $f_1 = \cos\phi$, $f_2 = \sin\phi$. Then:

$$\begin{aligned}\big[\mathcal{R}_\alpha f_1\big](\phi) \;&=\; \cos(\phi+\alpha) \;=\; (\cos\alpha)\,\cos\,\phi \,-\, (\sin\alpha)\,\sin\,\phi \;=\; (\cos-\alpha)\,f_1(\phi) \,+\, (\sin-\alpha)\,f_2(\phi)\\ \big[\mathcal{R}_\alpha f_2\big](\phi) \;&=\; \sin(\phi+\alpha) \;=\; (\sin\alpha)\,\cos\,\phi \,+\, (\cos\alpha)\,\sin\,\phi \;=\; -\,(\sin-\alpha)\,f_1(\phi) \,+\, (\cos-\alpha)\,f_2(\phi)\end{aligned}$$

Compare this with $D_i{}^j(-\alpha)\,f_j(\phi)$, and switch the sign of α to obtain the $\mathbf{D}(\alpha)$ matrix:

$$\mathbf{D}^{(1)}(R_\alpha) \;=\; \begin{pmatrix}\cos\alpha & \sin\alpha\\ -\sin\alpha & \cos\alpha\end{pmatrix}$$

Well, that's one two-dimensional representation for $SO(2)$, and it is probably the most often used. But it is not the only one! If we had instead chosen $f_1 = \mathrm{e}^{\mathrm{i}\phi}$, $f_2 = \mathrm{e}^{-\mathrm{i}\phi}$, going through the same procedure would yield another matrix:

$$\mathbf{D}^{(2)}(R_\alpha) \;=\; \begin{pmatrix}\mathrm{e}^{\mathrm{i}\alpha} & 0\\ 0 & \mathrm{e}^{-\mathrm{i}\alpha}\end{pmatrix}$$

so here is another different two-dimensional representation. Or is it different? In fact, no, because the transformation $\mathbf{S}^{-1}\,\mathbf{D}^{(1)}\,\mathbf{S}$, with the *single* matrix $\mathbf{S} = \begin{pmatrix}1 & \mathrm{i}\\ \mathrm{i} & 1\end{pmatrix}/\sqrt{2}$, diagonalises $\mathbf{D}^{(1)}$ into $\mathbf{D}^{(2)}$ for any angle α, ie. for all elements of the rotation group.

But there are more: *each* of the linearly independent functions $\mathrm{e}^{\mathrm{i}\alpha}$ and $\mathrm{e}^{-\mathrm{i}\alpha}$ is also a perfectly acceptable one-dimensional representation of $SO(2)$! Both $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$ can be viewed as a joining of these one-dimensional representations, which we shall call $\mathbf{D}^{(3)}$ and $\mathbf{D}^{(4)}$ Since $\mathrm{e}^{\mathrm{i}\alpha}$ is a representation of $U(1)$, this establishes the isomorphism of $SO(2)$ and $U(1)$ via the mapping $\mathrm{e}^{\mathrm{i}\alpha}$. Obviously, there must be something special about $\mathrm{e}^{\pm\mathrm{i}\alpha}$. Before we discover what it is, let us look at another instructive example.

Example 2.6. Let us work out a three-dimensional representation of the *left* action of S_3, $\pi^{-1}\,\mathbf{x}$, on $\mathbb{R}^3$. Since S_n merely shuffles the components of $\mathbf{x}$ it preserves its length, which is the definition of orthogonal matrices whose transpose is their inverse. In fact, $S_n \subset O(n)$! Then, from eq. (2.4), $\pi_k^{-1}\mathbf{x} = x^i\mathbf{e}_j D^j{}_i(\pi_k^{-1}) = \big(x^i D_i{}^j(\pi_k)\big)\mathbf{e}_j$ so as to view the permutations as a reshuffling of the *components* (written as row vectors!) of $\mathbf{x}$, and we obtain:

$$\mathbf{D}^{(1)}(\pi_1) \;=\; \begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}, \qquad \mathbf{D}^{(1)}(\pi_2) \;=\; \begin{pmatrix}0&1&0\\1&0&0\\0&0&1\end{pmatrix}, \qquad \mathbf{D}^{(1)}(\pi_3) \;=\; \begin{pmatrix}0&0&1\\0&1&0\\1&0&0\end{pmatrix},$$

$$\mathbf{D}^{(1)}(\pi_4) \;=\; \begin{pmatrix}1&0&0\\0&0&1\\0&1&0\end{pmatrix}, \qquad \mathbf{D}^{(1)}(\pi_5) \;=\; \begin{pmatrix}0&1&0\\0&0&1\\1&0&0\end{pmatrix}, \qquad \mathbf{D}^{(1)}(\pi_6) \;=\; \begin{pmatrix}0&0&1\\1&0&0\\0&1&0\end{pmatrix}$$

This so-called left **defining (fundamental) representation** of S_3 is faithful: it is one-to-one with the $\pi_i \in S_3$. An analogous n-dim defining representation can be constructed for any S_n.

Now, I claim that there exists another (two-dimensional!) representation of S_3:

$$\mathbf{D}^{(2)}(\pi_1) \;=\; \mathbf{D}^{(2)}(\pi_5) \;=\; \mathbf{D}^{(2)}(\pi_6) \;=\; \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\mathbf{D}^{(2)}(\pi_2) \;=\; \mathbf{D}^{(2)}(\pi_3) \;=\; \mathbf{D}^{(2)}(\pi_4) \;=\; \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

One immediately checks that products of these matrices are consistent with the group products of S_3 as given earlier in its composition table. So they do form a representation of S_3, albeit not a faithful one.

Even less faithful, but no less acceptable, is the one-dimensional representation obtained by mapping the π_i to their parity values:

$$\mathbf{D}^{(3)}(\pi_1) \;=\; \mathbf{D}^{(3)}(\pi_5) \;=\; \mathbf{D}^{(3)}(\pi_6) \;=\; 1$$

$$\mathbf{D}^{(3)}(\pi_2) \;=\; \mathbf{D}^{(3)}(\pi_3) \;=\; \mathbf{D}^{(3)}(\pi_4) \;=\; -1$$

And, of course, we can always map all the π_i to 1 and get another (trivial) representation!

On the other hand, we could join $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$ into a $\mathbf{D}^{(4)} = \mathbf{D}^{(1)} \oplus \mathbf{D}^{(2)}$ (direct sum) representation whose six matrices are 5-dimensional and block-diagonal, each with the submatrices on the diagonal taken, one from $\mathbf{D}^{(2)}$ (the upper one, say), and the other from $\mathbf{D}^{(1)}$, for a *given* permutation π_i.

A less trivial object, the **direct product** $\mathbf{D}^{\alpha\otimes\beta} = \mathbf{D}^{(\alpha)} \otimes \mathbf{D}^{(\beta)}$ of two representations $\mathbf{D}^{(\alpha)}$ and $\mathbf{D}^{(\beta)}$ is the matrix formed from all the products of their elements:

$$(\mathbf{D}^{\alpha\otimes\beta})^{ik}{}_{jl} \;=\; (\mathbf{D}^{(\alpha)})^{i}{}_{j}\,(\mathbf{D}^{(\beta)})^{k}{}_{l} \tag{2.6}$$

2.4.4 The regular representation of finite groups

Definition 2.29. The **regular** representation of the*left* action of a finite group is the set of matrices $\mathbf{D}(\mathrm{g})$ such that if $\mathrm{g} \in G$, then:

$$\mathbf{D}(\mathrm{g})\,\mathrm{g}_i \;=\; \mathrm{g}_j\,D^{j}{}_{i}(\mathrm{g}) \quad \forall\,\mathrm{g}_i \in G \qquad\qquad D^{j}{}_{i}(\mathrm{g}) \;=\; \begin{cases} 1 & \mathrm{g}\,\mathrm{g}_i \;=\; \mathrm{g}_j \\ 0 & \mathrm{g}\,\mathrm{g}_i \;\neq\; \mathrm{g}_j \end{cases}$$

The dimension of the regular representation is equal to N_{G}, the order of the group. It is seen to be closely related to the group-composition table of the group. We can also see that $D^{j}{}_{i}(e) = \delta^{j}{}_{i}$, ie. $\mathbf{D}(e) = \mathbf{I}$. Also, the other matrices in the representation must have a 1 as their $(ji)^{\mathrm{th}}$ element and 0 for all other elements in row j and column i; by inspection, this 1 is never on the diagonal.

Similarly, we can define a regular representation for the right action of a group.

Another way of thinking about the regular representation is to consider the group elements as forming an othonormal basis, $\{\mathrm{g}_i\}$, for a vector space. Then:

$$D_{ij}(\mathrm{g}) \;=\; \langle \mathrm{g}_j \,|\, \mathbf{D}(\mathrm{g}) \,|\, \mathrm{g}_i \rangle \;=\; \langle \mathrm{g}_j \,|\, \mathrm{g}\,\mathrm{g}_i \rangle \tag{2.7}$$

The regular-representation matrices can thus be viewed as composed of the matrix elements of the operators $\mathbf{D}(\mathrm{g})$.

A note of caution: do not confuse the dimension of a representation, which is the dimension of the carrier space (the space of *functions* on which group operators act), with the dimension of the coordinate space on which these functions act.

2.4.5 Invariant Spaces and Kronecker sum

To understand what relationship may exist between all those representations, it is now time to bring in another very useful concept:

Definition 2.30. Let $\{f^{(i)}\}$ be a subspace $\mathcal{H}^{(i)}$ of the space of functions on which the linear transformations $\{T_{\mathrm{g}}\}$ associated with a group G act. If, $\forall\, f^{(i)} \in \mathcal{H}^{(i)}$ and $\forall\, \mathrm{g} \in G$, $T_{\mathrm{g}}\, f^{(i)} \in \mathcal{H}^{(i)}$, the subspace is **invariant under** G.

Definition 2.31. Let $\mathcal{H}^{(1)}$ and $\mathcal{H}^{(2)}$ be subspaces of a Hilbert space $\mathcal{H}$ such that $\mathcal{H}$ is the sum of the two subspaces with the zero function as the only overlap between the two. Then, if any function in $\mathcal{H}$ can be written uniquely as the sum of a function in $\mathcal{H}^{(1)}$ and another in $\mathcal{H}^{(2)}$, $\mathcal{H}$ is called the **Kronecker (or direct) sum** of $\mathcal{H}^{(1)}$ and $\mathcal{H}^{(2)}$, written $\mathcal{H} = \mathcal{H}^{(1)} \oplus \mathcal{H}^{(2)}$. The dimension of $\mathcal{H}$ is the sum of the dimensions of $\mathcal{H}^{(1)}$ and $\mathcal{H}^{(2)}$.

2.4.6 Reducible and irreducible representations (BF10.5)

Definition 2.32. If some function space $\mathcal{H}$ has a *proper* subspace invariant under G, in the sense that the action of any $\mathbf{D}(\mathrm{g})$ on any element of the subspace does not go out of the subspace, then the representation consisting of the matrices $\mathbf{D}(\mathrm{g})$ ($\forall\, \mathrm{g} \in G$) is said to be **reducible**.

Definition 2.33. If a space $\mathcal{H}$ has no proper subspace invariant under a group G, the representations of G are said to be **irreducible**.

Definition 2.34. If, $\forall\, \mathrm{g} \in G$, the matrices in a representation $\mathbf{D}(\mathrm{g})$ can be brought into diagonal block form by the *same* similarity transformation, then the representation is reducible to lower-dimensional representations whose elements are the block matrices.

If there is another level of invariant subspaces, so that any or all of these block matrices can themselves be written in diagonal block form, and so on, until we are left with only irreducible representations, then one arrives at the **fully reducible** representation $\mathbf{D}(\mathrm{g})$:

$$\mathbf{D}(\mathrm{g}) \;=\; a_1\, \mathbf{D}^{(1)}(\mathrm{g}) \,\oplus\, a_2\, \mathbf{D}^{(2)}(\mathrm{g}) \,\oplus\, \cdots \oplus\, a_N\, \mathbf{D}^{(N)}(\mathrm{g}) \tag{2.8}$$

where a_i is the number of times the irreducible representation $\mathbf{D}^{(i)}(\mathrm{g})$ occurs in the Kronecker (direct) sum.

When the n-dimensional function space $\mathcal{H}$ has proper invariant subspaces, it means that while $\mathcal{H}$ may have a set of linearly independent functions, there are at least two subspaces in $\mathcal{H}$, each of which has its own smaller set of linearly independent functions which transform among themselves. Indeed, let $\mathcal{H}^A$ be an invariant subspace of dimension d, and let $\{\mathbf{e}_1, \ldots, \mathbf{e}_d, \ldots\}$ be a basis of $\mathcal{H}$ with $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$ a basis of $\mathcal{H}^A$. We write vectors of functions in $\mathcal{H}$ in block form $\begin{pmatrix} A \\ B \end{pmatrix}$, where $A \in \mathcal{H}^A$ has dimension d, and B belongs to the complement subspace $\mathcal{H}^B$, of dimension n-d. When the complement subspace is also invariant, as it always is in cases of interest to physics (see section 2.4.7 just below), then a representation matrix:

$$\mathbf{D}(\mathrm{g}) \;=\; \begin{pmatrix} \mathbf{D}^A(\mathrm{g}) & 0 \\ 0 & \mathbf{D}^B(\mathrm{g}) \end{pmatrix} \tag{2.9}$$

maps vectors $\begin{pmatrix} A \\ B \end{pmatrix}$ into other vectors $\begin{pmatrix} A' \\ B' \end{pmatrix}$ where $A' \in \mathcal{H}^A$ and $B' \in \mathcal{H}^B$. The submatrix $\mathbf{D}^A(\mathrm{g})$ has dimension d, and $\mathbf{D}^B(\mathrm{g})$ has dimension n-d. Also, since:

$$\begin{pmatrix} \mathbf{D}^A(\mathrm{g}) & 0 \\ 0 & \mathbf{D}^B(\mathrm{g}) \end{pmatrix} \begin{pmatrix} \mathbf{D}^A(\mathrm{g}') & 0 \\ 0 & \mathbf{D}^B(\mathrm{g}') \end{pmatrix} = \begin{pmatrix} \mathbf{D}^A(\mathrm{g})\,\mathbf{D}^A(\mathrm{g}') & 0 \\ 0 & \mathbf{D}^B(\mathrm{g})\,\mathbf{D}^B(\mathrm{g}') \end{pmatrix}$$

$\mathbf{D}^A(\mathrm{g})$ and $\mathbf{D}^B(\mathrm{g})$ do preserve the group product, eg: $\mathbf{D}^A(\mathrm{g})\mathbf{D}^A(\mathrm{g}') = \mathbf{D}^A(\mathrm{g}\,\mathrm{g}')$, as they should.

Going back to $SO(2)$, the $\mathbf{D}^{(2)}$ representation we have obtained is clearly fully reducible as it is written in block-diagonal form, with 1×1 submatrices $\mathbf{D}^{(3)} = \mathrm{e}^{\mathrm{i}\alpha}$ and $\mathbf{D}^{(4)} = \mathrm{e}^{-\mathrm{i}\alpha}$ (ie. it is diagonal), and we can write it as $\mathbf{D}^{(2)} = \mathbf{D}^{(3)} \oplus \mathbf{D}^{(4)}$, where both $\mathbf{D}^{(3)}$ and $\mathbf{D}^{(4)}$ are one-dimensional and therefore, irreducible.

As for S_3, the 5-dimensional representation we have constructed, $\mathbf{D}^{(4)}$, is (by construction) reducible since it is in block-diagonal form, and we have $\mathbf{D}^{(4)} = \mathbf{D}^{(1)} \oplus \mathbf{D}^{(2)}$. What about the two-dimensional representation, composed of $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$? Try to find invariant subspaces of the space of 2-dim vectors, that is, a subset of 2-dim vectors which are sent into themselves by both matrices. Well, the identity matrix is already in block-diagonal form. As for the other matrix, we diagonalise it to find its eigenvectors: $\left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right)$, corresponding to eigenvalue $+1$, and $\left(\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right)$, corresponding to eigenvalue -1. The representation is reducible since a similarity transformation exists that transforms $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ into $\left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$, Thus, the $\mathbf{D}^{(2)}$ representation gets reduced to the two *inequivalent* 1-dim irreducible representations $\mathbf{D}^{(5)} = 1$, $\mathbf{D}^{(6)} = -1$. We can then write the $\mathbf{D}^{(4)}$ representation of S_3 as the direct sum: $\mathbf{D}^{(4)} = \mathbf{D}^{(1)} \oplus \mathbf{D}^{(5)} \oplus \mathbf{D}^{(6)}$. As for $\mathbf{D}^{(1)}$, it is itself reducible (EXERCISE) to two irreducible representations, $\mathbf{D}^{(5)}$, and a set of six 2-dim matrices. Therefore, the 5-dim $\mathbf{D}^{(4)}$ is fully reducible, since it can be written as the direct sum of four representations, one 2-dim, one 1-dim, and two 1-dim (the parity of the elements), all irreducible. Can you see why these irreducible representations could not all be one-dimensional?

So this algorithm works, but it would be nice not to have to rely on looking for invariant subspaces and similarity transformations, which can get quite involved.

2.4.7 Unitary representations (BF10.6)

A representation $\mathbf{D}(\mathrm{g})$ is unitary if $\mathbf{D}^\dagger(\mathrm{g}) = \mathbf{D}(\mathrm{g}^{-1})$, $\forall\, \mathrm{g} \in G$. In terms of matrix elements, $D_{ij}(\mathrm{g}^{-1}) = D^*_{ji}(\mathrm{g})$.

Notice that both $\mathbf{D}^{(3)}$ and $\mathbf{D}^{(4)}$ for $SO(2)$ are unitary, since their complex conjugate is their inverse. So is $\mathbf{D}^{(2)}$.

Now, if $\mathbf{D}(\mathrm{g})$ is not already unitary, we can always find a similarity transformation matrix $\mathbf{S}$, the Hermitian square root of $\mathbf{S}^2 = \sum_g \mathbf{D}^\dagger(g)\mathbf{D}(g)$, such that $\mathbf{D}'(\mathrm{g}) = \mathbf{S}\,\mathbf{D}(\mathrm{g})\,\mathbf{S}^{-1}$ is unitary (EXERCISE—first, show that $\mathbf{D}^\dagger(g')\mathbf{S}^2\mathbf{D}(g') = \mathbf{S}^2$). *Any representation of a finite group is equivalent to a unitary representation*, ie. one whose matrices are unitary. This is also true for certain infinite (continuous) groups, such as compact Lie groups. Also, it can be shown that if a subspace is invariant under a unitary representation, the complement of that subspace must also be invariant. And it can also be shown that every representation of a *finite* group is fully reducible.

2.5 Schur's Lemmas and the Classification of Group Representations (BF10.6)

We now present two fundamental results of group theory which provide useful criteria for the irreducibility of representations and from which can be derived relations that are of tremendous help in classifying these representations.

2.5.1 Schur's Lemmas

Schur's First Lemma: The only matrix $\mathbf{M}$ that commutes with the matrices of a given *irreducible* representation $\mathbf{D}(\mathrm{g})$, $\forall \mathrm{g} \in G$, is a multiple of the identity matrix.

Indeed, since any $\mathbf{M} \in GL(n, \mathbb{C})$ has at least one (perhaps complex) non-zero eigenvector, let $\mathbf{M}\,\mathbf{A} = \lambda\,\mathbf{A}$. And, assuming that $\mathbf{M}$ and $\mathbf{D}$ commute, we have $\mathbf{M}\,(\mathbf{D}\,\mathbf{A}) = \mathbf{D}\,\mathbf{M}\,\mathbf{A} = \lambda\,(\mathbf{D}\,\mathbf{A})$. This means that the $\mathbf{D}(\mathrm{g}_p)\,\mathbf{A}$

are also eigenvectors of $\mathbf{M}\ \forall\, \mathrm{g}_p \in G$; since they span the same vector space as the eigenvectors of $\mathbf{M}$, that space is invariant under G.

By definition 2.33 of irreducibility, however, that space $\mathcal{H}$ on which $\mathbf{D}$ acts has no *proper* subspace itself invariant under G. Therefore, the subspace spanned by the eigenvectors associated with this one eigenvalue is the whole of $\mathcal{H}$. In other words, $\forall\, \psi \in \mathcal{H}$, $\psi = a_p\, \mathbf{A}^{(p)}$, where $\mathbf{A}^{(p)} = \mathbf{D}(\mathrm{g}_p)\mathbf{A}$, and:

$$\mathbf{M}\,\psi = a_i \mathbf{M}\,\mathbf{A}^{(i)} \;=\; \lambda\, a_i\, \mathbf{A}^{(i)} \;=\; \lambda\,\psi$$

so that *all* vectors in $\mathcal{H}$ are eigenvectors of $\mathbf{M}$, with the *same* eigenvalue λ. This can happen only if $\mathbf{M} = \lambda\,\mathbf{I}$.

As a corollary, if a matrix can be found which is not a multiple of $\mathbf{I}$ and yet commutes with all matrices in a representation, that representation must be reducible.

Two important consequences of Schur's First Lemma

From Schur's First Lemma follows an immediate consequence for Abelian groups. These have commuting elements, and so any matrix $\mathbf{D}(\mathrm{g})$ in a given representation commutes with the matrices associated with *all* other group elements in the representation. Assume that the representation is irreducible; then the Lemma requires that $\mathbf{D}(\mathrm{g}) = \lambda_\mathrm{g}\,\mathbf{I}$, and this for all $\mathrm{g} \in G$. But the $n \times n$ identity matrix, which is diagonal, cannot be irreducible if it represents *all* group elements, contradicing our assumption. Instead, $\mathbf{D}$ is fully reducible to n irreducible representations 1. We conclude that *all irreducible representations of an Abelian group are one-dimensional*.

Here is another neat result that can be derived from Schur's First Lemma. Let $\mathcal{C}$ be a class of G, and let $\mathbf{D}(\mathrm{g}_p)$ be a matrix in a given *irreducible* representation of G. Construct a matrix by summing all the matrices $\mathbf{D}(\mathrm{g}_p)$ in the representation that correspond to elements g_p of $\mathcal{C}$:

$$\mathbf{M} \;=\; \sum_p \mathbf{D}(\mathrm{g}_p) \qquad\qquad \forall\, \mathrm{g}_p \in \mathcal{C} \tag{2.10}$$

Now, if g' is some arbitrary element of G, we have:

$$\mathbf{D}(\mathrm{g}')\,\mathbf{M}\,\mathbf{D}(\mathrm{g}'^{-1}) \;=\; \sum_p \mathbf{D}(\mathrm{g}')\,\mathbf{D}(\mathrm{g}_p)\,\mathbf{D}(\mathrm{g}'^{-1}) \;=\; \sum_p \mathbf{D}(\mathrm{g}'\,\mathrm{g}_p\,\mathrm{g}'^{-1}) \;=\; \mathbf{M}$$

where the next-to-last equality expresses the fact that the matrices preserve the group product of G, and the last equality results from the fact that, since $\mathrm{g}'\,\mathrm{g}_p\,\mathrm{g}'^{-1} \in \mathcal{C}$, the left-hand side of the last equality is just a rearrangement of the previous equation. Thus, $\mathbf{D}(\mathrm{g})\,\mathbf{M} = \mathbf{M}\,\mathbf{D}(\mathrm{g})\ \forall\, \mathrm{g} \in G$, and, from Schur's First Lemma, $\mathbf{M} = \lambda\,\mathbf{I}$, with λ a constant that depends on the class and on the representation. Thus, the trace of $\mathbf{M}$ is $n\lambda$, where n is the dimension of the representation.

Before proceeding to find λ, we establish an interesting fact: In a given representation, *all matrices associated with elements of the same class have the same trace.* Recall that the class to which g belongs is made of $\{\mathrm{g}'\,\mathrm{g}\,\mathrm{g}'^{-1}\}\ \forall\, \mathrm{g}' \in G$. Then the trace of $\mathbf{D}(\mathrm{g}'\,\mathrm{g}\,\mathrm{g}'^{-1})$ is equal[†] to the trace of $\mathbf{D}(\mathrm{g})$, which we denote by χ.

Using this fact and the definition of $\mathbf{M}$, we have: $\mathbf{Tr}\, M = n_\mathrm{c}\chi$, where n_c is the number of elements in the class. Since that trace is also $n\lambda$, we find:

$$\mathbf{M} \;=\; \frac{n_\mathrm{c}}{n}\,\chi\,\mathbf{I} \tag{2.11}$$

Schur's Second Lemma: If a non-zero matrix $\mathbf{M}$ exists such that $\mathbf{D}^{(\alpha)}(\mathrm{g})\mathbf{M} = \mathbf{M}\mathbf{D}^{(\beta)}(\mathrm{g})\ \forall\, \mathrm{g} \in G$, then $\mathbf{D}^{(\alpha)}$ and $\mathbf{D}^{(\beta)}$ must be equivalent irreducible representations. If $\mathbf{D}^{(\alpha)}$ and $\mathbf{D}^{(\beta)}$ are inequivalent, $\mathbf{M} = 0$.

This lemma is often proved (pp. BF615–617) by assuming that the representations are unitary. This makes for no loss of generality when G is finite or is a compact Lie group, since these (eg. $O(n)$) have finite-dimensional representations.

[†]This is because $\mathbf{Tr}\, AB = A_i{}^j\, B_j{}^i = B_j{}^i\, A_i{}^j = \mathbf{Tr}\, BA$.

2.5.2 An orthogonality relation (BF10.6)

Another important consequence of Schur's Lemmas is the fact that the matrix elements of all the inequivalent irreducible representations of a finite group, or those for infinite groups that have finite-dimensional representations, form a set of *orthogonal* functions of the elements of the group. More specifically, if $\{\mathbf{D}(\mathrm{g}_p)\}$ is the set of all matrices $\mathbf{D}_p$ in an *irreducible* representation $\mathbf{D}$, then, for two such representations labelled by α and β::

$$\sum_{p=1}^{N_\mathrm{G}} \left(D_p^{(\alpha)}\right)^i{}_k \left(D^{(\beta)}_{\mathrm{g}_p^{-1}}\right)_j{}^l \;=\; \frac{N_\mathrm{G}}{n_\alpha}\,\delta^i{}_j\,\delta_k{}^l\,\delta_{\alpha\beta} \tag{2.12}$$

where N_G is the order of the group and n_α is the dimension of the representation $\mathbf{D}^{(\alpha)}$. The expression on the left-hand side is not matrix multiplication! Each term is the product of *any* two matrix entries of $D_p^{(\alpha)}$ and $D^{(\beta)}_{\mathrm{g}_p^{-1}}$.

In the case of unitary representations, this relation simplifies to:

$$\sum_{p=1}^{N_\mathrm{G}} \left(D_p^{(\alpha)}\right)^i{}_k \left(D_p^{(\beta)\,*}\right)^l{}_j \;=\; \frac{N_\mathrm{G}}{n_\alpha}\,\delta^i{}_j\,\delta_k{}^l\,\delta_{\alpha\beta} \tag{2.13}$$

These relations set powerful constraints on the matrix elements of representations.

Eq. (2.12) is so important that it deserves a proof. Fortunately, this proof is fairly easy. Define a matrix $\mathbf{M}$ such that:

$$\mathbf{M} \;=\; \sum_{p=1}^{N_\mathrm{G}} \mathbf{D}_p^{(\alpha)}\,\mathbf{X}\,\left[\mathbf{D}_p^{(\beta)}\right]^{-1} \tag{2.14}$$

where $\mathbf{D}^{(\alpha)}$ and $\mathbf{D}^{(\beta)}$ are irreducible matrix representations of G, and X is any arbitrary operator represented by a matrix $\mathbf{X}$. Note that the sum runs over all the group elements, each one labelled by a value of p. Then, for some $\mathrm{g}_{p'} \in G$,

$$\mathbf{D}_{p'}^{(\alpha)}\,\mathbf{M}\,\left[\mathbf{D}_{p'}^{(\beta)}\right]^{-1} \;=\; \sum_{p=1}^{N_\mathrm{G}} \mathbf{D}_{p'p}^{(\alpha)}\,\mathbf{X}\,\left[\mathbf{D}_{p'p}^{(\beta)}\right]^{-1}$$

The sum on the right-hand side is just a different rearrangement of the sum that defines $\mathbf{M}$, so that:

$$\mathbf{M} \;=\; \mathbf{D}_{p'}^{(\alpha)}\,\mathbf{M}\,\left[\mathbf{D}_{p'}^{(\beta)}\right]^{-1}$$

and therefore $\mathbf{M}$ satisfies the criterion specified in Schur's Lemmas: $\mathbf{D}^{(\alpha)}(\mathrm{g})\,\mathbf{M} = \mathbf{M}\,\mathbf{D}^{(\beta)}(\mathrm{g})\;\forall\,\mathrm{g} \in G$. There are two cases to consider:

- When $\alpha \neq \beta$ (ie. for different *inequivalent* irreducible representations), $\mathbf{M} = 0$ from the Second Lemma. In index notation, eq. (2.14) becomes:

  $$M^i{}_j \;=\; \sum_{p=1}^{N_\mathrm{G}} \left(\mathbf{D}_p^{(\alpha)}\right)^i{}_m X^m{}_n \left(\mathbf{D}^{(\beta)}_{\mathrm{g}_p^{-1}}\right)^n{}_j \;=\; 0$$

  Because $\mathbf{X}$ is arbitrary, this greatly restricts the $\mathbf{D}$ matrices. Let us choose $\mathbf{X}$ to be a matrix whose only non-zero element, 1, is its $(kl)^{\mathrm{th}}$ entry. We can write this formally as: $X^m{}_n = \delta^m{}_k\,\delta_n{}^l$ to obtain:

  $$\sum_{p=1}^{N_\mathrm{G}} \left(\mathbf{D}_p^{(\alpha)}\right)^i{}_k \left(\mathbf{D}^{(\beta)}_{\mathrm{g}_p^{-1}}\right)^l{}_j \;=\; 0$$

- When $\alpha = \beta$, Schur's First Lemma requires that $\mathbf{M} = \lambda_{\mathbf{X}}\,\mathbf{I}$, ie.:

$$M^i{}_j \;=\; \sum_{p=1}^{N_G} \left(\mathbf{D}_p^{(\alpha)}\right)^i{}_m \, X^m{}_n \left(\mathbf{D}_{g_p^{-1}}^{(\alpha)}\right)^n{}_j \;=\; \lambda_{\mathbf{X}}\,\delta^i{}_j$$

  Choosing the same $\mathbf{X}$ as before this time leads to:

$$\sum_{p=1}^{N_G} \left(\mathbf{D}_p^{(\alpha)}\right)^i{}_k \left(\mathbf{D}_{g_p^{-1}}^{(\alpha)}\right)^l{}_j \;=\; \lambda_k{}^l\,\delta^i{}_j$$

  Taking the trace (setting $i = j$ with summation over repeated indices) and interchanging the $\mathbf{D}$ factors to get a matrix product, there comes:

$$\sum_{p=1}^{N_G} \left(\mathbf{D}_{g_p^{-1}}^{(\alpha)}\right)^l{}_j \left(\mathbf{D}_p^{(\alpha)}\right)^j{}_k \;=\; \lambda_k{}^l\,n_\alpha$$

  The product inside the sum now implies matrix multiplication and, because the $\mathbf{D}$ matrices preserve the group product of G, we have:

$$\sum_{p=1}^{N_G} \left(\mathbf{D}_{g_p^{-1}}^{(\alpha)}\right)^l{}_j \left(\mathbf{D}_p^{(\alpha)}\right)^j{}_k \;=\; \sum_{p=1}^{N_G} \left(\mathbf{D}_{g_p g_p^{-1}}^{(\alpha)}\right)^l{}_k \;=\; N_G\,\delta^l{}_k$$

  Comparing, we get: $\lambda^l{}_k = N_G \delta^l{}_k / n_\alpha$.

There only remains to collect the results for $\alpha = \beta$ and $\alpha \neq \beta$ into the compact forms (2.12) or (2.13).

This means that the matrix elements $\sqrt{\frac{n_\alpha}{N_G}}\,D_{ij}^{(\alpha)}$ of a unitary irreducible representation must be orthonormal functions of the group elements g, and therefore linearly independent over the space spanned by the group elements. Furthermore, they form a complete set.

EXERCISE: Show that for a finite group the sum over all elements g_p of the matrix elements $(D_p)^i{}_j$ (i and j fixed) of an *irreducible* representation is zero. This property can provde a useful.check.

2.5.3 Characters of a representation (BF10.7); orthogonality of the characters

Definition 2.35. The **character** of a representation $\mathbf{D}(\mathrm{g})$ of a group G is defined as a map from G to $\mathbb{C}$:

$$\chi(\mathrm{g}) \;=\; \mathbf{Tr}\,\mathbf{D}(\mathrm{g})$$

Therefore the character of a representation for $\mathrm{g} \in G$ is just the trace of its representing matrix $\mathbf{D}(\mathrm{g})$. Characters of reducible representations are **compound**; those of irreducible representations are called **simple**.

Matrices for equivalent representations all have the same character; in other words, any statement about characters is basis-independent! Also, if g and g′ are conjugate to one one another, there must exist an element g″ of G such that

$$\mathbf{D}(\mathrm{g}') \;=\; \mathbf{D}(\mathrm{g}'')\,\mathbf{D}(\mathrm{g})\,\mathbf{D}^{-1}(\mathrm{g}'')$$

Therefore, *all the matrices representing group elements in the same class and representation have the same character*. But, if they are to be inequivalent, two representations must have at least one different character for some class,

Now we can set $k = i$ and $l = j$ in eq. (2.13):

$$\sum_{p=1}^{N_G} \left(D_p^{(\alpha)}\right)^i{}_i \left(D_p^{(\beta)*}\right)^j{}_j \;=\; \frac{N_G}{n_\alpha}\,\delta^i{}_j\,\delta_i{}^j\,\delta_{\alpha\beta} \;=\; \frac{N_G}{n_\alpha}\,\delta^i{}_i\,\delta_{\alpha\beta}$$

where repeated indices are summed over. Since $\delta^i{}_i = n_\alpha$, this can be rewritten as:

$$\sum_{p=1}^{N_\mathrm{G}} \chi_p^{(\alpha)}\, \chi_p^{(\beta)*} \;=\; N_\mathrm{G}\, \delta_{\alpha\beta} \tag{2.15}$$

This is our first orthogonality relation between the characters of irreducible representations.

Some of the terms in this sum will be identical since they correspond to group elements in the same class. So, instead of thinking of this sum as over all the elements of the group, we can collect all terms belonging to the same class, which we label with k, and sum over the classes, ie. over k:

$$\sum_{k=1}^{N_\mathrm{c}} n_k\, \chi_k^{(\alpha)}\, \chi_k^{(\beta)*} \;=\; N_\mathrm{G}\, \delta_{\alpha\beta} \tag{2.16}$$

where n_k is the number of elements in class k, and N_c is the number of classes in the group. This looks for all the world like an orthogonality relation between two vectors, $\sqrt{n_k/N_\mathrm{G}}\,\chi^{(\alpha)}$ and $\sqrt{n_k/N_\mathrm{G}}\,\chi^{(\beta)}$, each of dimension N_c.

For a given irreducible representation, eq. (2.16) becomes:

$$\sum_{k=1}^{N_\mathrm{c}} n_k \left|\chi_k^{(\alpha)}\right|^2 \;=\; N_\mathrm{G} \tag{2.17}$$

This is a necessary and sufficient condition for the representation to be irreducible!

> **Example 2.7.** Take for instance the 3×3 representation of S_3 found in section 2.4.3. There is one element, the identity, with trace 3, in its own class, three elements (the transpositions) in another class with trace 1, and two elements (the cyclic permutations) with trace 0. Eq. (2.17) gives: $c_1\chi_1^2 + c_2\chi_2^2 + c_3\chi_3^2 = 1\,(3)^2 + 3\,(1)^2 + 2\,(0)^2 = 12$. This is not equal to 6, the number of elements in S_3, therefore the representation must be reducible.

According to eq. (2.16), the character "vectors" of the N_r different irreducible representations are orthogonal. There are N_r such orthogonal vectors, and their number may not exceed the dimensionality of the space, N_c, so that $N_\mathrm{r} \le N_\mathrm{c}$. We will need this result a little later.

Now consider the decomposition of a fully reducible representation into a direct sum of irreducible ones, given in eq. (2.8). If we take its trace, we get an equation for the compound character $\chi(\mathrm{g})$: $\chi(\mathrm{g}) = a_\alpha\, \chi^{(\alpha)}(\mathrm{g})$, where the (direct) sum runs over the N_r irreducible representations of the group. The compound character is seen to be a linear combination of simple characters with positive coefficients (the number of times a given irreducible representation appears in the decomposition).

Multiplying this relation by $\chi^{\star\,(\beta)}(\mathrm{g})$ and summing over group elements now labelled by an index p, we find from eq. (2.15) which expresses the orthogonality of the characters:

$$\sum_{p}^{N_\mathrm{G}} \chi_p\, \chi_p^{\star\,(\beta)} \;=\; a_\alpha \sum_{p}^{N_\mathrm{G}} \chi_p^{(\alpha)}\, \chi_p^{*\,(\beta)} \;=\; a_\alpha\, N_\mathrm{G}\, \delta_{\alpha\beta} \;=\; a_\beta\, N_\mathrm{G}$$

Thus, the number of times—**multiplicity**—each irreducible representation of a group occurs in the direct-sum decomposition of a reducible representation is readily calculated in terms of characters, often making this direct sum easy to obtain:

$$a_\alpha \;=\; \frac{1}{N_\mathrm{G}} \sum_{p}^{N_\mathrm{G}} \chi_p\, \chi_p^{*\,(\alpha)} \tag{2.18}$$

On the other hand, given some representation with characters $\chi(\mathrm{g})$, we can easily find whether it is irreducible by computing (EXERCISE): $\sum_k n_k \chi_k \chi_k^* = N_\mathrm{G} \sum_\alpha a_\alpha^2$. If the sum on the left turns out to be equal to N_G, then the sum on the right must be 1, and only one α contributes to it. Thus, the representation is irreducible; if the sum is greater than N_G, it is reducible.

2.5.4 Multiplicity of irreducible representations and a sum rule for their dimension

We are now ready to exploit the regular representation to obtain other *general* results for irreducible representations.

As we have seen in section 2.4.4, the entries of the matrices of the regular representation can only be 1 or 0. Since only the identity will map a group element to itself, the only matrix with 1 anywhere on the diagonal is the identity matrix. Therefore, in the regular representation, the characters $\chi(\mathrm{g})$ all vanish except for $\chi(e) = N_{\mathrm{G}}$.

Now, when $\mathrm{g}_p = e$, the multiplicity relation (2.18) gives:

$$a_\alpha \;=\; \frac{1}{N_{\mathrm{G}}}\,\chi(e)\,\chi^{*\,(\alpha)}(e) \;=\; \chi^{*\,(\alpha)}(e) \;=\; n_\alpha$$

where, as before, n^α is the dimension of the α^{th} irreducible representation and we have used the fact that only $\chi(e)$ can contribute to the sum over all group elements since the characters of all other group elements in the regular representation vanish.

Therefore, *the multiplicity of an irreducible representation in the decomposition of the regular representation is the dimension of that irreducible representation*. Next, taking the trace of eq. (2.8) for the identity group element of the regular representation yields a relation between its compound and simple characters: $N_{\mathrm{G}} \;=\; a_\alpha\, n^\alpha$. Combining those results, there comes an important **sum rule**:

$$N_{\mathrm{G}} \;=\; \sum_\alpha n_\alpha^2 \tag{2.19}$$

This powerful sum rule tells us that the dimension of any irreducible representation must be smaller than the square root of the order of the group. Thus, when $N_{\mathrm{G}} = 2$ or 3, all inequivalent irreducible representations are one-dimensional. When $N_{\mathrm{G}} = 4$, we can have only four inequivalent 1-d irreducible representations; $n_\alpha = 2$ is ruled out because there would be no room left for the identity 1-d representation. In the case of a group of order 5, eq. (2.19) does allow the identity representation together with one 2-d irreducible representation; but we know that this group, being of order prime, is Abelian, and so admits only five inequivalent 1-d irreducible representations. For a group of order 6, the sum rule allows either six 1-d irreducible representations, or two 1-d and one 2-d irreducible representations. EXERCISE: Can you see why any representation of dimension larger than the square root of the order of a finite group must be reducible?

2.5.5 Another orthogonality relation

Let us go back to our first orthogonality relation, eq. (2.13). What it says is that the set $\{\sqrt{n_\alpha/n_{\mathrm{G}}}\big(D^{(\alpha)}\big)^i{}_j(\mathrm{g_p})\}$, with i and$j$ fixed but $1 \le p \le N_{\mathrm{G}}$, of an irreducible representation α, can be viewed as the N_{G} components of a vector. This vector is orthogonal to any other such vector corresponding to other matrix elements, whether or not they correspond to the same representation as that of the first vector. There are N_{G} such vectors and they form a complete set with completeness relation expressed as:

$$\sum_\alpha^{N_{\mathrm{r}}} \sum_{i,\,j}^{n_\alpha} \frac{n_\alpha}{N_{\mathrm{G}}} \big(D_p^{(\alpha)}\big)^i{}_j \,\big(D_{p'}^{(\alpha)\,*}\big)^i{}_j \;=\; \delta_{p'p} \tag{2.20}$$

where N_r is the number of irreducible representations. Note once again that the left-hand side is not matrix multiplication.

Well, what can we do with this result? We can take the equation for each p, where p is an element of some class k, and sum over all elements of the class; we can also do this with p' over the elements of another class k'. When $k \neq k'$, the right-hand side of the double summation must vanish because classes are distinct; when $k = k'$, the double sum collapses into one which adds up to n_k. On the left-hand side, we may recognise $\sum_p^{n_k} \big(D_p^{(\alpha)}\big)^i{}_j$ as an element of the matrix M introduced in eq. (2.10). The completeness relation now reads:

$$\sum_\alpha^{N_{\mathrm{r}}} \sum_{i,\,j}^{n_\alpha} \frac{n_\alpha}{N_{\mathrm{G}}} \big(M_k^{(\alpha)}\big)^i{}_j \,\big(M_{k'}^{(\alpha)\,*}\big)^i{}_j \;=\; n_k \delta_{k'k}$$

But there we had also found that $\mathbf{M}^{(\alpha)} = (n_k/n_\alpha)\chi^{(\alpha)}\,\mathbf{I}$. Inserting and carrying out the sums over i and j gives another orthogonality relation:

$$\sum_{\alpha=1}^{N_\mathrm{r}} \frac{n_k}{N_\mathrm{G}}\,\chi_k^{(\alpha)}\,\chi_{k'}^{(\alpha)*} \;=\; \delta_{k'k} \tag{2.21}$$

Thus, the characters in the k^{th} class in a given irreducible representation can be considered as components of vectors forming a basis of a space whose dimension is N_r, the number of irreducible representations. There are N_c such orthogonal vectors, one for each class. But, in a N_r-dimensional space, there cannot be more than N_r orthogonal vectors, and $N_\mathrm{c} \le N_\mathrm{r}$.

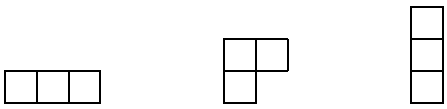
In section 2.5.3, however, we had argued that $N_\mathrm{r} \le N_\mathrm{c}$. These results together lead to the following important statement:

The number of irreducible representations of a group is equal to the number of its classes: $N_\mathrm{r} = N_\mathrm{c}$.

2.5.6 Exploring representations with Young diagrams

We have already discussed how Young diagrams could be used to find and label classes of S_N. But, much more often, it is representations that they help to label.We will be looking at S_N whose classes we have associated with partitions of N and, noting that the number of irreducible representations is also the number of classes, we will construct their Young diagrams with the same partitions λ_i of N, where the λ_i sum up to N and $\lambda_1 \ge \ldots \ge \lambda_N$. So the Young diagrams for the irreducible respresentations of S_N look exactly like those for its classes.

What *will* be different is the labelling of the Young diagrams: instead of taking the partitions as the number of boxes in columns from left to right, we will take them as the number of boxes in rows from top to bottom. For S_3, this gives:



(3) $(2\,1)$ 1^3

The sequence of representation labels is just the reverse of that for classes! But, if they are not cycles, what are they?

To discover the meaning of these Young diagrams we consider how the corresponding permutations act on functions in the carrier space of the $N!$-dimensional regular representation. We start by giving ourselves a set of functions $\{\psi_i\}$ $(1 \le i \le N)$, each of one variable, where the choice of the same symbol as for particle wave-functions in quantum mechanics is intentional (some authors use the Dirac notation for them). Then with products of these we construct functions of N variables x_j. For instance, the product $\psi_{(1\ldots N)} = \psi_1(x_1)\cdots\psi_1(x_N)$ spans a one-dimensional subspace which contains functions which are obviously completely symmetric, ie. invariant, under *any* of the $N!$ possible permutations of the variables. Thus, our subspace qualifies as an invariant subspace of the regular representation, and it makes sense to associate it with the 1-d irreducible identity representation which has the same character, 1, for all elements of S_N. We shall follow the usual convention by associating it with the single Young diagram with one row of N boxes. Its label will therefore always be (N).

With the same set $\{\psi_i\}$, we can also construct the completely antisymmetric function:

$$\psi_{[1\ldots N]} \;=\; \begin{vmatrix} \psi_1(x_1) & \cdots & \psi_1(x_N) \\ \vdots & \cdots & \vdots \\ \psi_N(x_1) & \cdots & \psi_N(x_N) \end{vmatrix}$$

This function changes sign under any transposition in its set of variables, and the 1-d subspace it spans is also invariant, because the function resulting from multiplying $\psi_{[1\ldots N]}$ by ± 1 is obviously in the same subspace. We associate this subspace with the 1-d irreducible representation which sends each element of S_N to its parity, $+1$ or -1. Again by convention, this in turn corresponds to the single one-column Young diagram with N rows.

Other irreducible representations, and thus Young diagrams, have a mixed symmetry which can be used to find their dimension. This is even stronger than eq. (2.19) which is only a constraint on the *possible* dimensions. Here is one way to do this.

- Take the Young diagram for each irrep, and fill each of its N boxes with numbers from 1 to N in all possible permutations to generate $N!$ **Young tableaux**. Then assign a function with N subscripts, living in the carrier space of the regular representation of S_N, to each tableau.The order of the subscripts follows the order of numbers in the first row, then the second row, until the last row. These functions represent products of functions, each of one coordinate, but we no longer treat them explicitly as such. They form a basis for the carrier space of the regular representation.
- Symmetrise each function with respect to the numbers in each row of the tableau, and antisymmettrise the result with respect to the numbers in each column. This yields, for each diagram, a new, mixed-symmetry function, $\psi^{(i)}$ $(1 \leq i \leq N)$, that is a linear combination of the previous $N!$ basis functions for the carrier space of the regular representation.

Example 2.8. For the $(2\,1)$ irreducible representation of S_3, the Young tableaux and corresponding mixed-symmetry functions would be:

$$\begin{array}{|c|c|}\hline 1&2\\\hline 3\\\cline{1-1}\end{array} \quad \Psi^{(1)} = \psi_{123} + \psi_{213} - \psi_{321} - \psi_{231} \qquad \begin{array}{|c|c|}\hline 1&3\\\hline 2\\\cline{1-1}\end{array} \quad \Psi^{(2)} = \psi_{132} + \psi_{312} - \psi_{231} - \psi_{321}$$

$$\begin{array}{|c|c|}\hline 2&1\\\hline 3\\\cline{1-1}\end{array} \quad \Psi^{(3)} = \psi_{213} + \psi_{123} - \psi_{312} - \psi_{132} \qquad \begin{array}{|c|c|}\hline 2&3\\\hline 1\\\cline{1-1}\end{array} \quad \Psi^{(4)} = \psi_{231} + \psi_{321} - \psi_{132} - \psi_{312}$$

$$\begin{array}{|c|c|}\hline 3&1\\\hline 2\\\cline{1-1}\end{array} \quad \Psi^{(5)} = \psi_{312} + \psi_{132} - \psi_{213} - \psi_{123} \qquad \begin{array}{|c|c|}\hline 3&2\\\hline 1\\\cline{1-1}\end{array} \quad \Psi^{(6)} = \psi_{321} + \psi_{231} - \psi_{123} - \psi_{213}$$

The question now is, are these mixed functions independent? Since we expect the regular representation to be reducible (fully reducible, in fact), there should exist a lower-dimensional invariant subspace, the carrier space of our irreducible representation of interest, and we should be able to show that there are only $n_\alpha < 6$ (for S_3) independent combinations, where n_α will be the number of basis functions for the invariant subspace, and therefore the dimension of the irreducible representation of S_3 carried by that space.

We note immediately that linear combinations that differ by a transposition of numbers in a column of their tableaux cannot be independent: they are the negative of one another. So we have at most three linearly independent combinations. But we also see that $\Psi^{(1)} - \Psi^{(2)} - \Psi^{(3)} = 0$, leaving only two independent combinations, which we take to be $\Psi^{(1)}$ and $\Psi^{(2)}$, and which are the basis functions for the carrier space of a 2-dim irreducible representation.

This rather tedious procedure can be made much faster by filling the tableaux in all the possible ways subject to the following rules: the number 1 fills the uppermost, leftmost box; and the numbers must increase down columns and along rows. The number of ways this can be done is the dimension of the representation. For instance, the $(2\,1)$ Young diagram of S_3, when filled, generates the two tableaux with so-called **standard numbering**:

$$\begin{array}{|c|c|}\hline 1&2\\\hline 3\\\cline{1-1}\end{array} \qquad \begin{array}{|c|c|}\hline 1&3\\\hline 2\\\cline{1-1}\end{array}$$

$$\Psi^{(1)} \qquad \Psi^{(2)}$$

each corresponding to one basis function in the 2-dimensional invariant subspace carrying the $(2\ 1)$ irrep of S_3.

There is also a very convenient method for calculating the dimension of the representation associated with a Young diagram if one does not wish to construct bases for the subspaces:

Definition 2.36. For any box in the Young diagram associated with an irreducible representation, draw a straight line *down* to the last box in its column and all the way to the right end of the box's row. The result is called a **hook** and the number of boxes traversed by the hook is the **hook length** of this box.

Then the dimension of the associated irreducible representation is the order of S_N, $N!$, divided by the product of the N hook lengths for this diagram..

Definition 2.37. Irreducible representations for which the Young diagrams are the transpose of each other, ie. for which the length of each row in one is equal to the length of the corresponding column in the other, are said to be **conjugate**. Their dimensions are the same.

A **self-conjugate** irreducible representation is one for which the Young diagram and its transpose are identical.

Now it can be shown that the product of an irreducible representation with a 1-d representation is itself an irreducible representation, which may be the same (when the 1-d representation is the trivial one). This goes for their characters also. When the completely antisymmetric (1^N) 1-d representation exists, as is the case for S_N, the characters of a conjugate irreducible representation can always be written, class by class, as the product of the characters of its conjugate representation and the characters in the (1^N) representation. Therefore, characters for the same class in conjugate representations are either identical or differ only by their sign. Characters of a self-conjugate representation in a class that has negative parity must vanish.

2.5.7 Character tables

A character table contains all the distinct characters for all the irreducible representations of a group. Each row contains the characters of all the group elements in a given representation, and each column contains the characters of a class of the group in all its representations.

The first row corresponds to the identity 1-dim irreducible representation, (N); all entries in that row must be 1. Also, the first column corresponds to the identity class; each entry in that column must be the dimension of the representation (ie. the trace of the identity matrix for each representation) for the corresponding row.

If we are dealing with S_N, we always have another 1-dim irreducible **antisymmetric** representation (called the **sign representation** by mathematicians), (1^N), whose 1×1 matrices, and characters, are the parities ± 1 of its classes. We choose to place the characters of this completely antisymmetric representation in the bottom row of the table .

What about the entries which are neither in the first/last row nor in the first column? Well, we can assign some algebraic symbol to the unknown characters and then spot conjugate representations. If there are any, the character in each column of one of representation in a conjugate pair must be the character of its conjugate multiplied by the character (± 1) in the antisymmetric 1-d row. If there are self-conjugate irreducible representations, any character sitting in the same column as a -1 in the last 1-dim row must be zero.

Some information can also be extracted from the defining representation of the group, whose dimension we denote by N. This representation, which is always reducible, is in fact fully reducible to a 1-dim representation and a (N - 1)-dimensional irreducible representation. Indeed, let $(x_1, \ldots, x_N)$ be a set of coordinates in the carrier space of the defining representation. It is easy to construct a fully symmetric combination of all those coordinates:

$$X = \frac{x_1 + \ldots + x_N}{N}$$

This function spans the 1-dim subspace of $\mathbb{R}^N$ invariant under any permutation of the coordinates and is the carrier space of the irreducible representation of S_N labelled by (N). Since the defining representation is unitary, the

complementary subspace is itself invariant, and is the carrier space of another irreducible representation. Indeed, let this (N-1)-dim subspace be spanned by N - 1 functions of the mixed-symmetry form:

$$Y_{j-1} \;=\; \frac{x_1 \,+\, \ldots \,+\, x_{j-1} \,-\, (j-1)\, x_j}{\sqrt{j(j-1)}} \qquad\qquad 2 \le j \le N$$

These N-1 **Jacobi coordinates** can be shown to be linearly independent, so that there is no proper invariant subspace, and the representation is irreducible. The functions are symmetrised with respect to j - 1 coordinates and then antisymmetrised with respect to the $N^{\rm th}$ one. This allows us to identify the representation with the (N-1 1) irreducible representation of S_N.

The characters of this (N-1 1) representation can be calculated as follows. First, we note that for a class labelled $(\ldots\, 2^\beta\, 1^\alpha)$, the characters of the defining representation are equal to the number of objects that the permutations in the class leave invariant, ie. α (see for instance the defining representation of S_3 in section 2.4.3). Since this character must be the sum of the characters of the (N-1 1) and of the (N) irreducible representations, we find that the characters of each class labelled by α in the (N-1 1) irreducible representation are just $\alpha - 1$.

In the case of S_N, other low-hanging fruit helping to find characters involves the class labelled by the cycle structure (N) in an N-dimensional irreducible representation (if there is one). The corresponding permutations shuffle *all* N objects, and the N-dimensional matrices representing them must have 0 as diagonal entries, resulting in a character that is 0.

Next, we let eq. (2.16) and (2.21) provide constraints on the remaining unknowns:

- The first says that *complete* rows in the table (each belonging to different representations) are orthogonal to each other, with the understanding that each term in the sum is weighted by the number of elements in the class (column).

- The second says that complete columns (each belonging to different classes) are othogonal.

Moreover, if one lets β refer to the identity representation, then, for any irrep α *other* than the identity, eq. (2.16) becomes:

$$\sum_{k=1}^{N_{\rm c}} n_k\, \chi_k^{(\alpha)} \;=\; 0 \qquad (2.22)$$

This is usually the best way to use eq. (2.16), at least at the beginning, because the resulting constraints are linear. Unfortunately, many of these relations will be automatically satisfied and will not yield new information, because of the strong constraints on the characters imposed by conjugation and self-conjugation of the irreducible representations. When all possible information has been extracted from eq. (2.22) and (2.21), and there still remain unknowns, one can try to spot reasonably simple quadratic relations from eq. (2.16) as well as using the normalisation of rows and columns.

Two last but important remarks: the characters of the 1-dim representations of any group (eg. those of an Abelian group) must preserve the group product. Also, although the characters of S_N are real, characters of other groups (eg. Z_n) can be complex.

There exist even more sophisticated methods for determining the characters of a group (eg. by generating them from the characters of a subgroup, or of a factor group), but lack of time and space prevents us from discussing them here. In fact, character tables for well-known groups can be found in specialised books and on the web.

Let us use these rules to find the characters of S_3 as a 3×3 table, with classes corresponding to columns and irreducible representations to rows. The first and last row can be immediately written down from our knowledge of the parity of each class (-1 for the transpositions and $+1$ for the cyclic permutations). Note also that the $(2\,1)$ representation is self-conjugate, so we can put 0 for the character in the $(2\,1)$ class, because the parity of that class (last character in the column) is -1. The $(2\,1)$ representation is the (N-1 1) representation discussed above, and its remaining character is determined by its belonging to a class with $\alpha = 0$; thus, the character must be -1. The linear constraint (2.22), as well as the other orthogonality rules, are automatically satisfied and did not have to be used. Collecting everything, we arrive at:

	(1^3)	$(2\,1)$	(3)
n_k	1	3	2
(3)	1	1	1
$(2\,1)$	2	0	-1
(1^3)	1	-1	1

EXERCISE: work out the character table and irreducible representations of Z_4, the cyclic group of order 4. You may make the task easier by remembering that products of characters belonging to a 1-d irreducible representation, which are the actual representation matrices, must mimic the group product of the corresponding elements.

Example 2.9. Lifting of a degeneracy by a weak interaction.

Consider a physical system in a rotationally-invariant potential that depends only on distance to the origin. This often occurs in quantum mechanics, and the result is that the eigenstates labelled by the integers that characterise eigenvalues of L^2 and L_z, l and m, with $-l \le m \le l$, exhibit a $2l+1$-fold degeneracy, in the sense that they all have the same energy. This is also manifested by the way spherical harmonics, which are eigenfunctions of L^2 and L_z for a given value of l, transform under a rotation by some angle α. Using eq. (2.3), we have:

$$\big[\mathcal{R}_\alpha Y_{lm}\big](\theta,\phi) \;=\; \sum_{m'=-l}^{l} Y_{lm'}(\theta,\phi)\,\big(D^{(l)}\big)^{m'}{}_{m}(\alpha)$$

where the $\mathbf{D}^{(l)}$ matrix is an irreducible representation of the rotation group $SO(3)$ which acts on the invariant space spanned by the $2l+1$ Y_{lm} for that l.

Now, we can simplify things by noting that rotations by an angle α about any axis are all equivalent to (in the same class as) a rotation by that angle around some axis. Since we are only interested in characters, take this axis to be the z-axis. Then $\mathcal{R}_\alpha Y_{lm}(\theta,\phi) = \mathrm{e}^{\mathrm{i}m\alpha} Y_{lm}(\theta,\phi) = Y_{lm}(\theta,\phi+\alpha)$ because the dependence of the spherical harmonics on ϕ is $\mathrm{e}^{\mathrm{i}m\phi}$. The $\mathbf{D}^{(l)}(\alpha)$ matrix is diagonal, with entries running from $\mathrm{e}^{-\mathrm{i}l\alpha}$ to $\mathrm{e}^{\mathrm{i}l\alpha}$, and its character is not hard to compute:

$$\chi^{(l)}(\alpha\neq 0) \;=\; \sum_{m=-l}^{l} (\mathrm{e}^{\mathrm{i}\alpha})^m \;=\; \mathrm{e}^{-\mathrm{i}l\alpha}\sum_{n=0}^{2l}(\mathrm{e}^{\mathrm{i}\alpha})^n \;=\; \mathrm{e}^{-\mathrm{i}l\alpha}\left(\frac{1-\mathrm{e}^{\mathrm{i}(2l+1)\alpha}}{1-\mathrm{e}^{\mathrm{i}\alpha}}\right) \;=\; \frac{\sin(l+1/2)\alpha}{\sin(\alpha/2)} \tag{2.23}$$

where we have recast the sum as a geometric series by redefining the index as $m = n - l$.

Now let us turn on a weak interaction whose corresponding potential is no longer rotationally-invariant, but still retains that invariance for rotations by a restricted, finite set of angles, which we collectively denote by β. Suppose this set actually is a group, or more precisely, a subgroup of $SO(3)$. Then the matrix $\mathbf{D}^{(l)}(\beta)$ should be a representation of that subgroup, *but that representation may no longer be irreducible*. This will certainly happen for any $\mathbf{D}^{(l)}$ whose dimension is too large to satisfy the sum rule (2.19) that applies to the finite subgroup.

The set of Y_{lm}, which we abbreviate to $\{f_i\}$, transform as: $\mathcal{R}_\beta f_i = f_j D^j{}_i$, with summation over repeated indices implied, and where we drop the label (l) and the reference to β to keep the formalism as lean as possible. But now, there may exist a matrix $\mathbf{S}$ independent of β which transforms $\mathbf{D}$ into a block-matrix $\mathbf{D'}$, something which was impossible when there was no restriction on the angles. That is, $\mathbf{D'} = \mathbf{S}^{-1}\mathbf{D}\,\mathbf{S}$. Next , noting that the f_i form a *row* vector for the left action of the subgroup, apply $\mathcal{R}_\beta$ to $S^i{}_l f_i$, which is just a linear combintation of the f_i because the $S^i{}_l$ are independent of β:

$$\mathcal{R}_\beta f_i\, S^i{}_l \;=\; f_j\, D^j{}_i S^i{}_l \;=\; f_j\, S^j{}_i D'^i{}_l = \; D'^i{}_l S^j{}_i f_j$$

where we have made liberal use of the commutativity of expressions written in index notation. This can be rewritten as $\mathcal{R}_\beta h_l \;=\; D'^i{}_l\, h_i$. In other words, there exist linear combinations of the spherical harmonics which transform amongst themselves under the restricted rotations.

But we do not have to find $\mathbf{S}$ to extract useful information. Indeed, we can calculate the characters of $\mathbf{D}^{(l)}(\beta)$ for all elements of the restricted-angle subset in $SO(3)$. Then we find the character table of the isomorphic group, which is finite. If there is a row in the table that exactly matches the $SO(3)$ characters of $\mathbf{D}^{(l)}(\beta)$, then $\mathbf{D}^{(l)}(\beta)$ is not only an irreducible representation of $SO(3)$, it is also an irrep of its subgroup defined by the angles allowed by the restricted symmetry. The corresponding invariant subspaces are identical, and the original $2l+1$-fold degeneracy for that value of l is still present after the perturbation has been turned on. As l increases, however, the dimension $2l+1$ of $\mathbf{D}^{(l)}(0)$, which always appears as the first character corresponding to the identity class of $SO(3)$, will eventually exceed the fixed dimensions of all the identity matrices for the identity class of the subgroup. All the corresponding $\mathbf{D}^{(l)}(\beta)$ will be reducible to a direct sum of the irreducible representations of the subgroup, given by eq. (2.8), with the multiplicity of each irrep calculable from eq. (2.18).

For instance, suppose that the perturbation has cubic symmetry. A cube is invariant under[1]:

- 6 rotations by $\pm\pi/2$ around the three axes through its centre that intersect faces through their centre;
- 3 rotations by π around these same axes;
- 8 rotations by $\pm 2\pi/3$ around the four axes through diagonally opposed corners.
- 6 rotations by π around the six axes intersecting the centre of two diagonally opposed edges;

With the identity rotation, these add up to 24 elements forming a subgroup of $SO(3)$ isomorphic to S_4. The five classes of S_4 are $\mathcal{C}_1$ (the identity), $\mathcal{C}_2$ (rotations by $\pm\pi/2$), $\mathcal{C}_3$ (rotations by $\pm 2\pi/3$), $\mathcal{C}_4$ (3 rotations by π), and $\mathcal{C}_5$ (6 rotations by π). With this ordering, the character table of S_4 is:

	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_4$	$\mathcal{C}_5$
n_k	1	6	8	3	6
$D_{(1)}$	1	1	1	1	1
$D_{(2)}$	1	−1	1	1	−1
$D_{(3)}$	2	0	−1	2	0
$D_{(4)}$	3	1	0	−1	−1
$D_{(5)}$	3	−1	0	−1	1

[1] See, eg: `http://demonstrations.wolfram.com/RotatingCubesAboutAxesOfSymmetry3DRotationIsNonAbelian/`.

These irreps of S_4 (or of the group of symmetries of the cube) are given non-descriptive labels and are ordered by increasing dimension instead of their mixed-symmetry structure. Only the rows of characters matter for our purposes. With eq. (2.23), we calculate the characters of the representations of S_4 induced by $\mathbf{D}^{(l=1)}(\beta)$ and $\mathbf{D}^{(l=2)}(\beta)$, with angles β running through the values corresponding to the five classes of S_4:

	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_4$	$\mathcal{C}_5$
$\mathbf{D}^{(l=1)}$	3	1	0	-1	-1

	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_4$	$\mathcal{C}_5$
$\mathbf{D}^{(l=2)}$	5	-1	-1	1	1

The $l = 1$ irrep of $SO(3)$ restricted to the angles allowed by the cubic-symmetry subgroup has the same dimension and the same characters as the fourth representation of S_4 in the above character table for S_4. The invariant spaces are the same and there is no lifting of the unperturbed 3-fold degeneracy. The $l = 2$ irrrep of $SO(3)$, however, has no identical row in the S_4 character table, and must correspond to a*reducible* representation of S_4. With eq. (2.18), we calculate the following multiplicy for each irrep of S_4 that can appear in the decomposition of $\mathbf{D}^{(l=2)}(\beta)$: $a_1 = a_2 = a_4 = 0$, and $a_3 = a_5 = 1$. Then we have the decomposition:

$$\mathbf{D}^{((=2)}(\beta) \;=\; D_{(3)}(\beta) \oplus D_{(5)}(\beta)$$

The unperturbed 5-fold degeneracy of the $l = 2$ states is partially lifted to become two "levels", one 3-fold and one 2-fold degenerate.

2.6 Other Examples of Symmetry Helping to Solve Problems

In what follows, I wish to illustrate with a couple of other simple examples how powerful the use of symmetry can be in the solution of many problems.

Example 2.10. First, consider something ultra-simple: two identical bodies of mass m separately undergoing simple (undamped) harmonic motion due to the same external force. Let $\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ be their displacement vector away from their respective equilibrium position. (We have assumed that both objects lie some distance away from one another on the x axis.) This vector is a solution of Newton's 2$^{\text{nd}}$ Law for the system, $\ddot{\mathbf{X}} = -\mathbf{M}^{-1}\mathbf{K}\mathbf{X}$, where:

$$\mathbf{M} \;=\; \begin{pmatrix} m & 0 \\ 0 & m \end{pmatrix}, \qquad\qquad \mathbf{K} \;=\; \begin{pmatrix} k & 0 \\ 0 & k \end{pmatrix}$$

with k the restoring constant associated with the motion. We call $\mathbf{M}^{-1}\mathbf{K}$ the **dynamical matrix** of the system. Of course, as we all know, both bodies oscillate at the same frequency $\omega_0 = \sqrt{k/m}$. The space of solutions is spanned by the two eigenvectors $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$, both with the *same* eigenvalue ω_0. This is a two-fold degeneracy.

Now suppose that we introduce an *internal* interaction which is linear in the bodies' displacement (ie., usually, it is weak enough). We then say that the two bodies are **weakly coupled**.

The dynamical matrix becomes:

$$\mathbf{M}^{-1}\mathbf{K} \;=\; \begin{pmatrix} \omega_0^2 + \omega_1^2 & -\omega_1^2 \\ -\omega_1^2 & \omega_0^2 + \omega_1^2 \end{pmatrix}$$

where ω_1 parametrises the coupling.

This coupled system retains a "mirror" symmetry, in the sense that, under reflection about a perpendicular plane halfway between the rest positions of the bodies, $\mathbf{X} \to \mathbf{X}' = -\begin{pmatrix} x_2 \\ x_1 \end{pmatrix}$. Indeed, we write in matrix form:

$$\mathbf{X}' \;=\; \mathbf{S}\,\mathbf{X} \qquad\qquad \mathbf{S} \;=\; \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}$$

Apply S to Newton's 2nd Law. In matrix form:

$$\ddot{\mathbf{X}}' \;=\; \mathbf{S}\,\ddot{\mathbf{X}} \;=\; -\,\mathbf{S}\,\mathbf{M}^{-1}\,\mathbf{K}\mathbf{X} \;=\; -\,(\mathbf{S}\,\mathbf{M}^{-1}\,\mathbf{K}\,\mathbf{S}^{-1})(\mathbf{S}\mathbf{X})$$

that is: $\ddot{\mathbf{X}}' \;=\; -\,(\mathbf{S}\,\mathbf{M}^{-1}\,\mathbf{K}\,\mathbf{S}^{-1})\,\mathbf{X}'$. But, if Newton's 2nd Law is to be invariant under reflection, $\ddot{\mathbf{X}}' = -\mathbf{M}^{-1}\mathbf{K}\mathbf{X}'$. This forces $\mathbf{S}\mathbf{M}^{-1}\mathbf{K} = \mathbf{M}^{-1}\mathbf{K}\mathbf{S}$. In operator language, because $\mathbf{S}$ and $\mathbf{M}^{-1}\mathbf{K}$ commute, they must have a common set of eigenvectors.

Call these eigenvectors $\mathbf{A}$. Finding the eigenvalues and eigenvectors of $\mathbf{M}^{-1}\mathbf{K}$ may be hard; finding the eigenvalues and eigenvectors of $\mathbf{S}$ is not. We know that $\mathbf{S}^2 = \mathbf{I}$, so that if $\mathbf{S}\mathbf{A} = \lambda\mathbf{A}$, $\lambda^2 = 1$. Then $\mathbf{A}_+ = \binom{1}{-1}$, with eigenvalue $\lambda_+ = 1$, and $\mathbf{A}_- = \binom{1}{1}$, with eigenvalue $\lambda_- = -1$. Now it is a simple matter, knowing the eigenvectors of $\mathbf{M}^{-1}\mathbf{K}$, to find the normal frequencies of the system, ie. the eigenvalues $\omega_\pm$ of the $\mathbf{M}^{-1}\mathbf{K}$ operator, which satisfy $\mathbf{M}^{-1}\mathbf{K}\,\mathbf{A}_\pm \;=\; \omega_\pm^2\,\mathbf{A}_\pm$. These eigenvectors are called **normal modes**; the entries of a normal-mode vector are the *relative* amplitudes of the motion for each body when the system oscillates in that particular mode. Their importance lies in the fact that any displacement vector $\mathbf{X}$ of the system must be a linear combination of its normal modes.

Using group-theoretic language, $\mathbf{S}$ forms the regular 2-dim representation of the group of reflections of two objects about the plane, which has the same group-composition table as S_2. This can be reduced to 2 1-dim (irreducible) representations: 1 and -1. Notice how the coupling has broken the initial degeneracy of the uncoupled system: now $\omega_- = \omega_0$ and $\omega_+^2 = \omega_0^2 + 2\omega_1^2$, with one eigenvector for each.

Example 2.11. This time, we take six identical bodies arranged on a circle 60° apart. Each is subject to an identical external linear restoring force giving rise to *small* oscillations about their equilibrium position and tangent to the circle. Each are coupled by an identical weak interaction to their nearest neighbours $\pm 60^\circ$ away; similarly, they are coupled to their second next neighbours, $\pm 120^\circ$ away, by another (even weaker) interaction that is identical for both these neighbours; finally, a third (weakest) interaction couples each one to its opposite counterpart, 180° away. We wish to study the effect of the coupling on the motion of the bodies tangent to the circle.

Because of the symmetry of the interactions and of the system, the dynamical $\mathbf{M}^{-1}\mathbf{K}$ matrix must have the form:

$$\mathbf{M}^{-1}\mathbf{K} \;=\; \begin{pmatrix} \omega_0^2 & -\omega_1^2 & -\omega_2^2 & -\omega_3^2 & -\omega_2^2 & -\omega_1^2 \\ -\omega_1^2 & \omega_0^2 & -\omega_1^2 & -\omega_2^2 & -\omega_3^2 & -\omega_2^2 \\ -\omega_2^2 & -\omega_1^2 & \omega_0^2 & -\omega_1^2 & -\omega_2^2 & -\omega_3^2 \\ -\omega_3^2 & -\omega_2^2 & -\omega_1^2 & \omega_0^2 & -\omega_1^2 & -\omega_2^2 \\ -\omega_2^2 & -\omega_3^2 & -\omega_2^2 & -\omega_1^2 & \omega_0^2 & -\omega_1^2 \\ -\omega_1^2 & -\omega_2^2 & -\omega_3^2 & -\omega_2^2 & -\omega_1^2 & \omega_0^2 \end{pmatrix}$$

How can we use the symmetry to find the normal modes of the system? By recognising that the system must be invariant under 60° rotations. This operation is isomorphic to a cyclic permutation: $(1\,2\,3\,4\,5\,6) \in Z_6$. The regular representation matrix for this element of Z_6 looks like:

$$\mathbf{S} \;=\; \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Invariance under $\mathbf{S}$ means that $\mathbf{M}^{-1}\mathbf{K}$ and $\mathbf{S}$ commute. In fact, this last statement can be used to obtain the form of the $\mathbf{M}^{-1}\mathbf{K}$ matrix given above.

Now, as in the last example, the eigenvectors of $\mathbf{S}$ satisfy $\mathbf{S}\mathbf{A} = \lambda\mathbf{A}$. But since $\mathbf{S}^6 = \mathbf{I}$, we immediately find that the eigenvalues are the sixth roots of 1, as expected for the cyclic group. Therefore. $\lambda_{(m)} = \mathrm{e}^{\mathrm{i}m\pi/3}$, $(0 \le m \le 5)$. To each value of m corresponds an eigenvector $\mathbf{A}_{(m)}$ with components $A^j_{(m)} = \lambda^{j-1}_{(m)} = \mathrm{e}^{\mathrm{i}m(j-1)\pi/3}$.

As in Example 2.10, these eigenvectors are also the normal modes of the system. Inserting into the eigenvalue equation $\mathsf{M}^{-1}\mathsf{K}\,\mathbf{A}_{(m)} = \omega^2_{(m)}\,\mathbf{A}_{(m)}$ with the coupling parameters $\omega_{(5)} = \omega_{(1)}$ and $\omega_{(4)} = \omega_{(2)}$ yields the **dispersion relation**:

$$\omega^2_{(m)} = \sum_{j=1}^{6} \omega^2_{j-1}\, \mathrm{e}^{\mathrm{i}m(j-1)\pi/3} = \omega_0^2 - 2\,\omega_1^2 \cos m\pi/3 - 2\,\omega_2^2 \cos 2m\pi/3 - (-1)^m\,\omega_3^2$$

We note that $\mathbf{A}^*_{(1)} = \mathbf{A}_{(5)}$, and $\mathbf{A}^*_{(2)} = \mathbf{A}_{(4)}$. These modes are complex, which is a problem if they are supposed to correspond to real relative amplitudes. But we also note that $\omega_{(1)} = \omega_{(5)}$, and $\omega_{(2)} = \omega_{(4)}$; therefore, the corresponding eigenvectors span two invariant 2-dim subspaces, which allows us to take appropriate linear combinations of the eigenvectors to turn them into real modes of the same frequency.

Again, the coupling has lifted the original 6-fold degeneracy of the uncoupled system, but there is still some degeneracy left because of the two 2-dim subspaces.

This is as far as we can go without knowing the interaction parameters themselves. But we have succeeded in nailing down the relative amplitudes of motion of the bodies in each normal mode *without that explicit knowledge*!

3 MODULE III — GROUP THEORY 2: Lie Groups

3.1 Definitions

In this module we focus on a class of groups with an infinite number of elements. As groups, they of course satisfy the *algebraic* properties of a group as set out in definition 2.1. But we shall put in an extra requirement: that each group element g_p be in correspondence with a point P in some manifold. This means that the index "p" can in fact be taken as a set of continuous, real variables, and we write a group element as $\mathrm{g}(P)$ and say that the manifold parametrises the group. More precisely:

Definition 3.1. Let P be any point in a n-dim manifold M^n which is obtained from two other points, P_1 and P_2 from a mapping $P = \phi(P_1, P_2)$. Let $\mathrm{g}(P_1) \star \mathrm{g}(P_2) = \mathrm{g}(P)$ be the group product of an infinite group G. If the map ϕ and its inverse are differentiable, then G is a **Lie group**.

The important point to remember here is that since they correspond to points in a manifold, elements of a Lie group can be parametrised in terms of *smooth* coordinates on this manifold.

A Lie group is real if its manifold is real and complex if its manifold is complex.

The dimension of a Lie group is the dimension of its manifold.

Definition 3.2. A Lie group is **connected** if any pair of points on its manifold is connected by a continuous path.

The subset of all elements in a Lie group that are connected by a continuous path to the identity must be a subgroup of the Lie group. Thus, even if it is itself not connected, a Lie group must contain a connected subgroup.

Example 3.1. An infinite line with a coordinate patch $-\infty < x < \infty$ ($x \in \mathbb{R}$) is a 1-dim manifold. In section 2.1.1 we stated that $\mathbb{C}$ was a continuous group under addition. So is $\mathbb{R}$ itself, and if we write a group element as $\mathrm{g}(x) = \mathrm{e}^x$, we can easily deduce the function corresponding to the group product. Indeed, $\mathrm{g}(z) = \mathrm{g}(x) \star \mathrm{g}(y) = \mathrm{g}(x + y)$, and we are not surprised to find that: $\phi(x, y) = x + y$.

Example 3.2. What if we restrict $\theta = x \in \mathbb{R}$ so that $0 \leq \theta < 2\pi$? We can define group elements $\mathrm{g}(\theta) = \mathrm{e}^{\mathrm{i}\theta}$ with the group product:

$$\mathrm{g}(\theta_1) \star \mathrm{g}(\theta_2) \;=\; \mathrm{g}(\theta_1 + \theta_2 \mod 2\pi)$$

The group manifold here is the unit circle, S^1, with each point on the circle parametrised by its polar angle θ, and $\phi(\theta_1, \theta_2) = \theta_1 + \theta_2$. Note that this manifold is real, so that the group is real even though its elements are complex! It is Abelian, and connected.

Example 3.3. Real invertible 2×2 matrices form a group whose elements can be written as $\mathbf{g}(x) = \begin{pmatrix} x_1 & x_2 \\ x_3 & x_4 \end{pmatrix}$. If we impose the condition that the matrices have determinant 1, we see that the number of parameters is lowered by 1. We compute the set of three functions $z_i = \phi_i(x_1, x_2, x_3, y_1, y_2, y_3)$ consistent with the group product:

$$\begin{pmatrix} x_1 & x_2 \\ x_3 & \frac{1+x_2\,x_3}{x_1} \end{pmatrix} \begin{pmatrix} y_1 & y_2 \\ y_3 & \frac{1+y_2\,y_3}{y_1} \end{pmatrix} \;=\; \begin{pmatrix} z_1 & z_2 \\ z_3 & \frac{1+z_2\,z_3}{z_1} \end{pmatrix}$$

We obtain:

$$z_1 \;=\; x_1\,y_1 \,+\, x_2\,y_3 \qquad z_2 \;=\; x_1\,y_2 \,+\, x_2\,\frac{1+y_2\,y_3}{y_1} \qquad z_3 \;=\; x_3\,y_1 \,+\, y_3\,\frac{1+x_2\,x_3}{x_1}$$

In this particular parametrisation, the mappings ϕ_i are all differentiable only off the $x_1 = 0$ and $y_1 = 0$ planes. Whatever the manifold for the group is—see later—it cannot be covered with just this coordinate patch.

The inverse mapping corresponding to $\mathrm{g}(x)^{-1}$ can be read off the inverse matrix $\mathbf{g}^{-1}$.

Example 3.4. If we demand instead that the general complex 2×2 matrices be not only unimodular, but unitary as well, the treatment is simpler. Introduce the parametrisation:

$$\begin{pmatrix} a & b \\ -b^* & a^* \end{pmatrix} = \begin{pmatrix} x + \mathrm{i}\, y & z + \mathrm{i}\, w \\ -(z - \mathrm{i}\, w) & x - \mathrm{i}\, y \end{pmatrix}$$

with the condition $|a|^2 + |b|^2 = x^2 + y^2 + z^2 + w^2 = 1$ which guarantees that the matrix is unitary with determinant equal to 1. The group manifold is thus the unit 3-sphere S^3 embedded in $\mathbb{R}^4$ with coordinates (x, y, z, w); this is a *real* three-dimensional Lie group as was the last example.

Definition 3.3. A Lie group is **compact** when the volume of its manifold is finite.

3.2 Some Matrix Lie Groups

Amazingly enough, it turns out that almost all Lie groups of interest in physics, the so-called **classical** Lie groups, are either matrix groups or groups of transformations isomorphic to matrix groups. The only group product we ever have to consider is matrix multiplication, and inverse elements are just inverse matrices.

One useful and satisfying way of classifying Lie groups is to begin with $n \times n$ invertible matrices over some field $\mathbb{F}$ of numbers, the **general linear** group $GL(n, \mathbb{F})$, and then identify interesting subgroups by imposing constraints on the group elements of $GL(n, \mathbb{F})$. Some of these have already been mentioned in section 2.2.2. For a start, $GL(n, \mathbb{R}) \subset GL(n, \mathbb{C})$.

Following standard usage, we introduce the diagonal matrix I_p^q with p elements equal to $+1$ and q entries equal to -1, where $p + q = n$. In this notation, I_n is the n-dim identity matrix, and also the orthonormal metric in Euclidean n-dim space.

3.2.1 Groups obtained from linear constraints

A host of matrix subgroups are obtained by imposing linear constraints on their elements so that they have some block structure (eg. upper or lower triangular block form), and perhaps further demanding that some of the diagonal elements be 1. We shall not consider these here.

3.2.2 Bilinear or quadratic constraints: the metric (or distance)-preserving groups

Definition 3.4. The set of unitary transformations T of a complex matrix $\mathsf{M} \in GL(n, \mathbb{C})$ is defined by:

$$\mathsf{M} \;\mapsto\; \mathsf{T}\,\mathsf{M}\,\mathsf{T}^\dagger$$

where the subgroup of matrices T leaves $\mathsf{M} = \mathsf{I}_n$ (or the Euclidean n-dim metric) invariant:

$$\mathsf{T}\,\mathsf{I}_n\,\mathsf{T}^\dagger \;=\; \mathsf{T}\,\mathsf{T}^\dagger \;=\; \mathsf{I}_n$$

that is, those matrices for which $\mathsf{T}^{-1} = \mathsf{T}^\dagger$, and we call that subgroup $U(n) \subset GL(n, \mathbb{C})$: Both $U(n)$ and its matrices are **unitary**. Example 3.2 referred to $U(1)$.

Definition 3.5. The set of orthogonal transformations T of a real matrix $\mathsf{M} \in GL(n, \mathbb{R})$ is defined by:

$$\mathsf{M} \;\mapsto\; \mathsf{T}\,\mathsf{M}\,\mathsf{T}^{\mathrm{T}}$$

(T^{T} is the transpose of T), such that T leave I_n invariant:

$$\mathsf{T}\,\mathsf{I}_n\,\mathsf{T}^{\mathrm{T}} \;=\; \mathsf{T}\,\mathsf{T}^{\mathrm{T}} \;=\; \mathsf{I}_n$$

that is, those matrices for which $\mathsf{T}^{-1} = \mathsf{T}^{\mathrm{T}}$. The group of such matrices is called $O(n)$ and is **orthogonal**.

Be aware that n in $O(n)$ or $U(n)$ refers to the dimension of the *matrices*, not that of the group, which is the number of coordinates on its manifold! The $O(n)$ matrices have determinant ± 1, whereas the absolute value of the complex determinant of $U(n)$ matrices is equal to 1. Thus, (can you see why?) $O(n)$ is not connected. Neither, for that matter, is $U(n)$.

Finally, the **symplectic group** $Sp(2n, \mathbb{R}) \subset GL(2n, \mathbb{R})$ leaves invariant the antisymmetric "metric" matrix $\begin{pmatrix} 0 & \mathbf{I}_n \\ -\mathbf{I}_n & 0 \end{pmatrix}$. The classical Hamiltonian equations of motion are form-invariant (covariant) under symplectic transformations.

The group manifolds (and thus these groups themselves) are compact because their matrices define closed, bounded subsets of the manifolds that parametrise $GL(n, \mathbb{C})$ and $GL(n, \mathbb{R})$. $O(n)$ and $U(n)$ preserve the length (or norm) of n-vectors in Euclidean $\mathbb{R}^n$, and therefore also angles between those vectors (eg., the angles of any triangle are determined by the lengths of its sides).

We also have the non-compact groups which preserve the indefinite metric $\mathbf{I}_p^q$, defined by the transformations:

$$\mathbf{T}\,\mathbf{I}_p^q\,\mathbf{T}^{\mathrm{T}} = \mathbf{I}_p^q \qquad O(p,q) \tag{3.1}$$

$$\mathbf{T}\,\mathbf{I}_p^q\,\mathbf{T}^{\dagger} = \mathbf{I}_p^q \qquad U(p,q) \tag{3.2}$$

A famous example is $O(3,1)$, aka the **full Lorentz group**, that leaves the pseudo-Euclidean metric on $\mathbb{R}^4$ (or space-time distance) invariant; equivalently, the norm of a 4-vector $\mathbf{x}$ is left invariant by 3-dim rotations, Lorentz transformations (boosts), and space or time reflections. In principle, from the condition: $\mathbf{T}\,\mathbf{I}_3^1\,\mathbf{T}^{\mathrm{T}} = \mathbf{I}_3^1$, one could work out detailed constraints on the elements of the $O(3,1)$ matrices to find that there are six independent parameters, but this would be needlessly messy. There are far better ways of parametrising the group to extract all this information, and much more, as we shall see.

3.2.3 Multilinear constraints: the special linear groups

The **special linear** subgroups $SL(n, \mathbb{C}) \subset GL(n, \mathbb{C})$ and $SL(n, \mathbb{R}) \subset GL(n, \mathbb{R})$ contain matrices with determinant 1. This last constraint is n-linear.

Example 3.3 actually referred to $SL(2, \mathbb{R})$ with the constraint $x_1 x_4 - x_2 x_3 = 1$, a bilinear constraint.

$SL(n, \mathbb{R})$ is often referred to as the volume-preserving group in $\mathbb{R}^n$. But it does not preserve all lengths!

Important subgroups of the special linear groups and the metric-preserving groups may be obtained as their intersections. Thus, $SO(n) = O(n) \cap SL(n, \mathbb{R})$ and $SU(n) = U(n) \cap SL(n, \mathbb{C})$. These groups are compact. The group of example 3.4 was $SU(2)$.

A little earlier, we stated that $O(n)$ and $U(n)$ are not connected, but we know that they must have connected subgroups, ie. groups with elements connected to the identity by a continuous path. These are $SO(n)$ and $SU(n)$.

3.2.4 Groups of transformations

Continuous transformations in physics act on vectors, or on functions of vectors. These transformations belong to groups which are usually isomorphic to Lie groups.

We have already met such transformations in examples 2.4 (translations) and 2.5 (rotations).

Here, instead of passive transformations as in Module II, we will use the more usual active transformations, which map a vector into a *new* vector in the same basis.

1. **Translations** Let $T_{\mathbf{a}}$ act on $\mathbf{x} \in \mathbb{R}^3$ so that $T_{\mathbf{a}}\mathbf{x} = \mathbf{x} + \mathbf{a}$, where $\mathbf{a} \in \mathbb{R}^3$. This can be generalised; indeed, let f be an analytic function acting on $\mathbb{R} \times \mathbb{R}^n$. The left action on f of the operator $\mathcal{T}_{\mathbf{a}}$ associated with $T_{\mathbf{a}}$ is:

$$[\mathcal{T}_{\lambda\mathbf{a}}\, f](\mathbf{x}) = f(T_{\lambda\mathbf{a}}^{-1}\mathbf{x}) = f(\mathbf{x} - \lambda\mathbf{a}) \qquad \mathbf{a} \in \mathbb{R}^n,\ \lambda \in \mathbb{R}$$

   These are translations in the direction of $\mathbf{a}$; they form a group which is isomorphic to $G = \{\mathbb{R}, +\}$, the addition being on the parameter λ.

2. **Rotations**

   Parametrise 2-dim rotations in the $z = 0$ plane by R_α, with $R_\alpha \phi = \phi + \alpha$ (active rotation of a vector), with $[\mathcal{R}_\alpha\, f](\phi) = f(\phi - \alpha)$, and $-\pi < \phi \leq \pi$. In terms of the left action on $\mathbf{x} \in \mathbb{R}^3$: $\mathbf{x}' = R_\alpha^{(z)}\, \mathbf{x}$ (ie. $\mathbf{x}'$ obtained by rotating $\mathbf{x}$ by $+\alpha$ around the z axis), the matrix associated with $R_\alpha^{(z)}$ now is:

$$\begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad \text{Then} : [\mathcal{R}_\alpha\, f](\mathbf{x}) = f(R_\alpha^{-1}\mathbf{x}) \;=\; f(x\cos\alpha + y\sin\alpha,\; -x\sin\alpha + y\cos\alpha,\; z)\,.$$

   At least formally, this generalises immediately to arbitrary rotations in 3-dim space: $[\mathcal{R}\, f](\mathbf{x}) = f(R^{-1}\mathbf{x})$, where $\mathcal{R}$ can be factorised as $\mathcal{R}_{\alpha,\beta,\gamma} = \mathcal{R}_\alpha\, \mathcal{R}_\beta\, \mathcal{R}_\gamma$, α, β and γ being the famous Euler angles. We are not even going to try to write down the corresponding matrix; it is not so illuminating anyway. But it is not hard to show (exercise) that it must be orthogonal and have determinant 1, which means it is an element of $SO(3)$.

3. We also have **scale transformations** $\mathbf{x}' = a\mathbf{x}$, with $a \in \mathbb{R}$ a non-zero positive constant, and $\mathbf{x} \in \mathbb{R}^n$ in *Cartesian* coordinates. The group of these transformations is isomorphic to $GL_+(1, \mathbb{R})$. The restriction to Cartesian coordinates is important: in spherical coordinates over $\mathbb{R}^3$, only the radial coordinate would scale.

4. **Lorentz and Poincaré transformations**

   Lorentz boosts are given in Jackson's *Classical Electrodynamics*, eq. (11.19), for the $\mathbb{R}^4$ coordinates ct and 3-dim $\mathbf{x}$:

$$ct' \;=\; \gamma(ct - \boldsymbol{\beta}\cdot\mathbf{x}) \qquad\qquad \mathbf{x}' \;=\; \mathbf{x} \,+\, \frac{\gamma - 1}{\beta^2}\,(\boldsymbol{\beta}\cdot\mathbf{x})\,\boldsymbol{\beta} \,-\, \gamma\boldsymbol{\beta}(ct)$$

   where $\boldsymbol{\beta}$ is the velocity of the primed frame in the unprimed frame, and $\gamma = 1/\sqrt{1-\beta^2}$. Jackson's eq. (11.98) expresses this transformation in matrix form. It is not pretty. To include 3-dim rotations, just replace $\mathbf{x}$ by $R^{-1}_{\alpha,\beta,\gamma}\mathbf{x}$ in the second equation. It is not worth writing the resulting 4×4 matrix which will be an element of $SO(3,1)$ if we exclude time reversal and space reflection; otherwise the relevant group will be $O(3,1)$, the full Lorentz group. The transformation is a homogeneous one, which in the 4-vector formalism is written: $\mathbf{x}' = \Lambda\mathbf{x}$, where $\mathbf{x}$ is a *general* 4-vector (not necessarily position).

   We can extend the full Lorentz transformations to include space-time translations $\mathbf{t}$; this makes them inhomogeneous:

$$\mathbf{x}' \;=\; \Lambda\mathbf{x} \,+\, \mathbf{t}$$

   Whereas the homogeneous transformations left the norm of of a 4-vector invariant, these leave invariant the norm of the *difference* between two 4-vectors.

   If we call $\boldsymbol{\Lambda}$ the full Lorentz transformation matrix, we can construct the matrix for these transformations by adding to $\boldsymbol{\Lambda}$ a fifth row and column whose last element is a 1 that does not do anything, that is:

$$\begin{pmatrix} \mathbf{x}' \\ 1 \end{pmatrix} \;=\; \begin{pmatrix} \boldsymbol{\Lambda} & \mathbf{t} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix}$$

   These 5×5 matrices are elements of the 10-parameter Poincaré group $P(4,1)$.

These examples illustrate the isomorphism between physical transformations and matrix Lie groups. We can then identify, say, a rotation with a $SO(3)$ matrix. In fact, we go so far as calling $SO(3)$ the rotation group. The same applies to the other cases.

3.2.5 Operator realisation of groups of transformations: infinitesimal generators

Now we explore more deeply this isomorphism between groups of transformations of functions and Lie groups. We shall express the left action of a couple of transformations from the last section as *differential operators*. This is far from being a gratuitous exercise as we shall see very soon after.

1. **Translations**

   We can first look just at functions $f(x)$ ($x \in \mathbb{R}$). Then the result of a translation $T_a x = x + a$ ($a \in \mathbb{R}$) on f can be Taylor-expanded about x:

   $$[\mathcal{T}_a f](x) \;=\; f(T_a^{-1}x) \;=\; f(x-a) \;=\; (1 \,-\, a\,\mathrm{d}_x \,+\, \ldots)\, f(x) \;=\; \exp(-a\,\mathrm{d}_x)\, f(x)$$

   In $\mathbb{R}^3$ this generalises to:

   $$[\mathcal{T}_{\mathbf{a}} f](\mathbf{x}) \;=\; f(T_{\mathbf{a}}^{-1}\mathbf{x}) \;=\; f(\mathbf{x}-\mathbf{a}) \;=\; \sum_{n=0}^{\infty} \frac{1}{n!}\,(-a^i\,\partial_i)^n\, f(\mathbf{x}) \;=\; \exp(-a^i\,\partial_i)\, f(\mathbf{x}) \tag{3.3}$$

   (We must differentiate the function before evaluating it, so we can be less fussy about notation in the last equality.)

   The operators $-\partial_i$ are called the **infinitesimal generators** of translations, although in quantum mechanics $\mathbf{a}$ would be redefined so that it is the Hermitian momentum operator $\mathbf{p} = -\mathrm{i}\hbar\boldsymbol{\partial}$ that is called the infinitesimal generator of translations. Then we would write the translation operator as: $T_{\mathbf{a}} = \mathrm{e}^{-\mathrm{i}\mathbf{a}\cdot\mathbf{p}/\hbar}$.

   We note that the infinitesimal generators $-\partial_i$ commute amongst themselves.

2. **Rotations**

   For rotations $R_\alpha \phi = \phi + \alpha$ in the (xy) plane by a small angle α:

   $$[\mathcal{R}_\alpha f](\phi) \;=\; f(R_\alpha^{-1}\phi) \;=\; f(\phi-\alpha) \;=\; (1 \,-\, \alpha\,\mathrm{d}_\phi \,+\, \ldots)\, f(\phi) \;=\; \exp(-\alpha\,\mathrm{d}_\phi)\, f(\phi)$$

   As we have seen in the last section, generalising this rotation to $\mathbb{R}^3$ in Cartesian coordinates gives, for a rotation $R_\alpha^{(z)}\mathbf{x} = (x\cos\alpha - y\sin\alpha,\; x\sin\alpha + y\cos\alpha,\; z)$:

   $$[R_\alpha^{(z)} f](\mathbf{x}) \;=\; f(R_\alpha^{-1}\mathbf{x}) \;=\; f(x\cos\alpha + y\sin\alpha,\; -x\sin\alpha + y\cos\alpha,\; z)$$

   If we Taylor-expand the right-hand side of the last equality, we obtain:

   $$[\mathcal{R}_\alpha^{(z)} f](\mathbf{x}) \;=\; \left[1 \,+\, \alpha\,(y\,\partial_x \,-\, x\,\partial_y) \,+\, \ldots\right] f(\mathbf{x}) \;=\; \exp(\alpha\, M_z)\, f(\mathbf{x}) \tag{3.4}$$

   where $M_z = y\,\partial_x - x\,\partial_y$. We proceed in exactly the same way for rotations about the x and y axes to arrive at the general rotation operator:

   $$\mathcal{R}_{\alpha,\beta,\gamma} \;=\; \exp(\alpha M_x)\, \exp(\beta M_y)\, \exp(\gamma M_z) \tag{3.5}$$

   where $M_x = z\,\partial_y - y\,\partial_z, \quad M_y = x\,\partial_z - z\,\partial_x, \quad M_z = y\,\partial_x - x\,\partial_y$. Again, in quantum mechanics, these infinitesimal generators of rotations are redefined as $\mathbf{L} = \mathrm{i}\hbar\mathbf{M}$ and interpreted as the (Hermitian) angular momentum operators in Cartesian coordinates.

   The infinitesimal generators do not commute. Indeed: $[M_i,\, M_j] = \epsilon_{ij}{}^k\, M_k$, or $[L_i,\, L_j] = \mathrm{i}\hbar\,\epsilon_{ij}{}^k\, L_k$.

   (Note: we could have written — some do! — the infinitesimal generators of translations to define $\mathbf{M}$ as the negative of the above. The cost, however, would be an extra minus sign in the commutation relations.)

We can now find the infinitesimal generators of an arbitrary group of transformations with m parameters a^i in the neighbourhood of the identity, such that $a^i = 0 \,\forall\, i$ for the identity group element. These transformations map a point in a manifold M^n (not the group manifold!) to another not far from the first that can certainly be described by the same coordinate patch.

Let the transformations act (left action!) on a space of differentiable functions f defined on M^n:

$$[T_{\mathbf{a}}\, f](\mathbf{x}) \;=\; f(T_{\mathbf{a}}^{-1}\mathbf{x})$$

Focus on $T_{\mathbf{a}}\, f$, and take f as a function of the *parameters* a^i. As before, Taylor-expand the right-hand side to first order, this time around the identity parametrised by $\mathbf{a} = 0$:

$$[T_{\mathbf{a}}\, f](\mathbf{x}) \;=\; \left(1 \,+\, a^i\, \partial_{a^i}(T_{\mathbf{a}}^{-1}x)^j\Big|_{\mathbf{a}=0}\partial_j \,+\, \ldots\right) f(\mathbf{x})$$

where i runs over the number (m) of parameters, ie. the dimension of the group, and j from 1 to the dimension n of the space on which the functions f act.

Definition 3.6. The operators:

$$X_i \;=\; \partial_{a^i}(T_{\mathbf{a}}^{-1}x)^j\Big|_{\mathbf{a}=0}\partial_j \tag{3.6}$$

are called the **infinitesimal generators of the group of transformations**. In some references, the right-hand side is multiplied by $-\mathrm{i}$ (with appropriate adjustment to the expansion) so as to ensure hermiticity.

For example, considering rotations around the z axis in $\mathbb{R}^3$, there is just one parameter (angle) $a^1 = \alpha$, and only x and y derivatives can occur since z does not depend on α, and the second term in the square bracket of eq. (3.4) is recovered.

3.2.6 Infinitesimal generators of matrix Lie groups

Now we show how to linearise matrix groups and find their infinitesimal generators. This is not hard at all if we know the matrices. In general, the matrix elements will be analytic functions of some (non-unique!) set of group parameters a^i, and all we have to do is Taylor-expand the matrix to first order in the group parameters around the identity element $\mathbf{I}_n$, for which the a^i all vanish:

$$\mathbf{M}_{\mathbf{a}} \;=\; \mathbf{I}_n \,+\, a^i\,\mathbf{X}_i \qquad\qquad \mathbf{X}_i \;=\; \partial_{a^i}\mathbf{M}_{\mathbf{a}}\Big|_{\mathbf{a}=0} \tag{3.7}$$

where we understand that differentiating a matrix means differentiating each of its elements. The matrices $\mathbf{X}_i$ are the infinitesimal generators of the group. Again, some prefer the definition $\mathbf{X}_i \;=\; -\mathrm{i}\partial_{a^i}\mathbf{M}_{\mathbf{a}}\big|_{\mathbf{a}=0}$.

Example 3.5. $GL(n,\mathbb{R})$

Let the matrix elements $M^i{}_j$ be written as $M^i{}_j = \delta^i{}_j + a^i{}_j$. In matrix form, this is

$$\mathbf{M}_{\mathbf{a}} \;=\; \mathbf{I}_n \,+\, a^i{}_j\,\mathbf{E}_i^j \qquad\qquad (\mathbf{E}_i^j)^k{}_l \;=\; \delta^j{}_l\,\delta^k{}_i$$

The $\mathbf{E}_i^j$ matrices are the infinitesimal generators of $GL(n,\mathbb{R})$; they have zero elements except for the $(ij)^{\text{th}}$ element which is 1.

Example 3.6. A slightly more exciting example is $\mathbf{M}_\theta \in SO(2)$: $\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}$, for $0 \le \theta < 2\pi$, that effects rotations in a plane. Taylor-expand to first order:

$$\mathbf{M}_\theta \;\approx\; \begin{pmatrix}1 & -\theta\\ \theta & 1\end{pmatrix} \;=\; \mathbf{I}_2 \,+\, \theta\,\mathbf{X}$$

Then the infinitesimal generator of $SO(2)$ is a matrix fully consistent with the constraints on $SO(n)$ generators we shall discover in section 3.3.4:

$$\mathbf{X} \;=\; \partial_\theta \mathbf{M}_\theta\big|_{\theta=0} \;=\; \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

We shall write the space it spans as:

$$\mathfrak{so}(2) \;=\; \begin{pmatrix} 0 & -\theta \\ \theta & 0 \end{pmatrix}$$

Another example (EXERCISE) that is quite easy to work out is $SL(2,\mathbb{R})$; it will have three infinitesimal generators.

How the generators may be found when the group matrices are not known will be discussed after we formalise this linearisation procedure.

A couple of remarks are now in order.

- The term "infinitesimal generator" means exactly that: an operator that effects an infinitesimal transformation away from the identity. We want to reconstruct a *finite* transformation out of a succession of infinitesimal transformations involving only the *first-order* contribution in the expansion of a transformation or a matrix. That is:

  $$\mathbf{M_a} \;=\; \lim_{n\to\infty} (M_{\mathbf{a}/n})^n \;=\; \lim_{n\to\infty} \left(1 \,+\, (a^i/n)\mathbf{X}_i\right)^n$$

  Now the following relations on any (well-behaved) linear operator A hold:

  $$\exp(A) \;\equiv\; \lim_{n\to\infty} \left(1 \,+\, \frac{1}{n}\,A\right)^n \qquad\qquad \mathrm{e}^A \;=\; \sum_{n=0}^{\infty} \frac{1}{n!}\,A^n\,\mathrm{d}_A^n \mathrm{e}^A\Big|_{A=0} \;=\; \sum_{n=0}^{\infty} \frac{A^n}{n!} \tag{3.8}$$

  since the right-hand side of the first relation is equal to its derivative when $n \to \infty$. Therefore, $\mathbf{M_a} = \exp(a^i \mathbf{X}_i)$. This **exponential map**, useful though it is, must still be handled with some care as we shall discover later.

- To understand the importance of infinitesimal generators, consider that taking linear combinations of group elements, such as, for instance, the matrices of $SO(3)$ does not give an element of $SO(3)$. In other words, linear combinations of matrix representations of the rotation group are not rotations. In general, group products are non-linear in the group parameters, so linear combinations cannot be expected to preserve them.

  Linear combinations of infinitesimal generators of rotations, however, *are* generators of rotations! Indeed, there is a set $\{X_i\}$ of infinitesimal generators of a Lie group that forms a basis of a linear vector space. An arbitrary vector in the space can always be written as $b^i X_i$, with b_i the group parameters.

3.3 Lie Algebras

3.3.1 Linearisation of a Lie group product

The study of Lie groups can be considerably simplified by linearising their group product around the identity. To linearise the group product, we first write two group elements in the neighbourhood of the identity as $\mathrm{g} = e + a\epsilon X$ and $\mathrm{g}' = e + b\epsilon Y$, where ϵ is a sufficiently small real number and a and b are real, but arbitrary. X and Y are infinitesimal generators of g and g′, respectively. Now expand $\mathrm{g}\,\mathrm{g}'$ to first order in ϵ:

$$\mathrm{g}\,\mathrm{g}' \;\approx\; e \,+\, \epsilon\,(aX \,+\, bY) \,+\, \ldots$$

Since $aX \,+\, bY$ is a generator for the group product $\mathrm{g}\,\mathrm{g}'$, we see that the generators indeed form a linear vector space, as asserted at the end of the last section.

Now consider the group product $\mathrm{g}\,\mathrm{g}'\,\mathrm{g}^{-1} = h\,\mathrm{g}'$. $h = e$ if the group is Abelian. Also, h itself is a group element since it is equal to $\mathrm{g}\,\mathrm{g}'\,\mathrm{g}^{-1}\mathrm{g}'^{-1}$. Mathematicians often define $h = \mathrm{g}\,\mathrm{g}'\,(\mathrm{g}'\,\mathrm{g})^{-1}$ as the "commutator" for the *group* product, but we shall reserve the term for $[X,\,Y]$. Let us expand it to first non-vanishing order, this time writing $\mathrm{g} = e + \epsilon X + \epsilon^2 X^2/2$, $\mathrm{g}' = e + \epsilon Y + \epsilon^2 Y^2/2$, where ϵ is arbitrarily small:

$$\begin{aligned}\mathrm{g}\,\mathrm{g}'\,(\mathrm{g}'\,\mathrm{g})^{-1} &\approx (e + \epsilon X + \frac{1}{2}\epsilon^2 X^2)(e + \epsilon Y + \frac{1}{2}\epsilon^2 Y^2)(e - \epsilon X + \frac{1}{2}\epsilon^2 X^2)(e - \epsilon Y + \frac{1}{2}\epsilon^2 Y^2) + \dots \\ &\approx e + \epsilon^2 (XY - YX) + \dots\end{aligned}$$

All other order-ϵ^2 contributions cancel out. We write $[X,\,Y] = XY - YX$, the commutator of the infinitesimal generators X and Y. $[X,\,Y]$ is the generator for $\mathrm{g}\,\mathrm{g}'\,(\mathrm{g}'\,\mathrm{g})^{-1}$ and must thus be an element of the same vector space as X and Y. When the group product commutes, the corresponding commutator of the generators vanishes.

When the generators are in matrix form, it is straightforward if a bit tedious to show, just by expanding the expression, that the **Jacobi identity** holds:

$$\big[X,\,[Y,\,Z]\big] + \big[Y,\,[Z,\,X]\big] + \big[Z,\,[X,\,Y]\big] = 0 \tag{3.9}$$

The identity is more difficult to prove when the generators are differential operators, but it still holds.

Because $\big[X,\,[Y,\,Z]\big] - \big[[X,\,Y],\,Z\big] \neq 0$, Lie algebras are *non-associative*.

3.3.2 Definition of a Lie algebra

Now we are ready for an important definition that collects and generalises our findings:

> **Definition 3.7.** A **Lie algebra** $\mathfrak{g}$ (no relation to the metric of Module I) is a finite-dimensional vector space equipped with a bilinear product, the **Lie bracket** $[\ ,\ ]$, which:
>
> - is **linear**: $[aX + bY,\,Z] = a\,[X,\,Z] + b\,[Y,\,Z] \quad \forall\, a,\,b \in \mathbb{R}\,\text{or}\,\mathbb{C}$;
> - is **antisymmetric**: $[X,\,Y] = -\,[Y,\,X]$;
> - satisfies the Jacobi identity: $\big[X,\,[Y,\,Z]\big] + \big[Y,\,[Z,\,X]\big] + \big[Z,\,[X,\,Y]\big] = 0$.
>
> In physics, we usually take the Lie bracket to be the commutator $XY - YX$. Also, many physicists, because they always deal with the algebra, not the group, use G to denote $\mathfrak{g}$; this is confusing.

It is important to keep in mind that in a Lie algebra the action of one element on another is not given by the straightforward composition of the two, as might be naively assumed from the idea of linearisation, but by their *commutator*. The closure property of a Lie group translates into the closure of its algebra, in the sense that the commutator of any two of its elements is also an element of the algebra.

The algebra $\pm\mathrm{i}\,\mathfrak{g}$ is said to be an **essentially real** Lie Algebra. We have already seen that the linear and orbital angular momentum operators of quantum mechanics were related to real infinitesimal generators in that way.

Sometimes, however, it proves very convenient to construct a **complex extension** of a real or essentially real algebra, by allowing the coefficients of the expansion of its elements over a basis to be complex. For instance, we might wish to construct $J_\pm = J_x \pm \mathrm{i}J_y$. This gives us more flexibility in constructing useful bases.

The dimension n of a Lie algebra is that of the manifold of the group with which it is associated.

Representations of a Lie algebra are simply derived from those of the underlying group via the technique introduced in section 3.2.6, and they inherit any equivalence between the group's representations as well as their reducibility or irreducibility. More on this when we discuss representations of Lie algebras a bit later.

3.3.3 Structure constants of a Lie algebra

The commutators of its n infinitesimal generators which form a basis of a Lie algebra are themselves elements of the algebra, so they must be written as linear combinations of those basis generators:

$$[X_i,\, X_j] \;=\; C_{ij}{}^{k}\, X_k \tag{3.10}$$

The coefficients $C_{ij}{}^{k}$ are called the **structure constants** of the Lie algebra. They are said to specify the structure of the algebra. In fact, with some caveats which are rarely relevant, they pretty much tell us everything about the group itself.

Two Lie algebras are said to be isomorphic when they have the same dimension and structure constants.

The structure constants inherit the antisymmetry of the commutators: $C_{ji}{}^{k} = -C_{ij}{}^{k}$. When the structure constants all vanish, we say that the algebra is **Abelian**; then $[X,\, Y] = 0\ \forall\, X,\, Y \in \mathfrak{g}$.

The Jacobi identity on the elements of the algebra imediately gives (EXERCISE) a relation between the structure constants:

$$C_{ij}{}^{l}\, C_{kl}{}^{m} \;+\; C_{jk}{}^{l}\, C_{il}{}^{m} \;+\; C_{ki}{}^{l}\, C_{jl}{}^{m} \;=\; 0 \tag{3.11}$$

Defining a matrix $(\mathbf{D}_i)_j{}^{k} = -C_{ij}{}^{k}$, we find (EXERCISE) that $\mathbf{D}$ satisfies the commutation relation (3.10).

The structure constants for the essentially real algebra $\pm\mathrm{i}\,\mathfrak{g}$ are just (exercise) $\pm\mathrm{i}C_{ij}{}^{k}$. It is straightforward to show (exercise) that if the group's (and therefore algebra's) representations are unitary, as is the case for compact groups such as $SU(n)$ and $SO(n)$, the $C_{ij}{}^{k}$ are real. Very often, in the case of essentially real algebras, people will call the $C_{ij}{}^{k}$ themselves the structure constants instead of $\pm\mathrm{i}C_{ij}{}^{k}$.

3.3.4 A direct way of finding Lie algebras

Suppose we do not have an explicit form for the matrix representation of a Lie group in terms of the group parameters. All we know is the constraint(s) on the group elements. This is sufficient to find the Lie algebra; often the group matrix can then be reconstructed from the algebra.

The first thing to do is to linearise the constraints. We listed some important ones at the beginning of section 3.2. For metric-preserving compact groups, $\mathbf{M}\,\mathbf{I}_n\,\mathbf{M}^\dagger \;=\; \mathbf{I}_n$; for non-compact metric-preserving groups (when the metric is indefinite), we have $\mathbf{M}\,\mathbf{I}^q_p\,\mathbf{M}^\dagger \;=\; \mathbf{I}^q_p$. In both cases, the real version just entails replacing the complex conjugate transpose by the transpose.

Linearising for the compact groups, we get:

$$(\mathbf{I}_n \,+\, \epsilon\,\mathbf{A})^\dagger(\mathbf{I}_n \,+\, \epsilon\,\mathbf{A}) \;\approx\; \mathbf{I}_n \,+\, \epsilon\,(\mathbf{A}^\dagger \,+\, \mathbf{A}) \;=\; \mathbf{I}_n$$

Therefore the matrices representing the algebra are antihermitian: $\mathbf{A}^\dagger = -\mathbf{A}$. Their diagonal matrix elements are pure imaginary for unitary group algebras $\mathfrak{u}(n)$; for orthogonal group algebras $\mathfrak{o}(n)$, $\mathbf{A}$ is real skew-symmetric, and there are thus $n(n-1)/2$ independent parameters. Thus, we can say that $\mathfrak{o}(n)$ is the set of all real skew-symmetric matrices of rank n.

If we choose to use essentially real algebras instead (eg. L as generators of $\mathfrak{so}(3)$ instead of M in section 3.2.5), then $\mathbf{M} = I_n + \mathrm{i}\epsilon\mathbf{A}$, and we find that the representations must be Hermitian: $\mathbf{A}^\dagger = \mathbf{A}$.

If the group is non-compact, the same process yields: $\mathbf{A}^\dagger\mathbf{I}^q_p = -\mathbf{I}^q_p\mathbf{A}$. This is a bit messier, but we can simplify it by breaking $\mathbf{A}$ into block matrices. If $\mathbf{S}$ is a $q\times q$ matrix, $\mathbf{T}$ a $q\times p$ matrix, $\mathbf{U}$ a $p\times q$ matrix, and $\mathbf{V}$ a $p\times p$ matrix, then:

$$\begin{pmatrix} \mathbf{S}^\dagger & \mathbf{U}^\dagger \\ \mathbf{T}^\dagger & \mathbf{V}^\dagger \end{pmatrix}\begin{pmatrix} -\mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_p \end{pmatrix} + \begin{pmatrix} -\mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_p \end{pmatrix}\begin{pmatrix} \mathbf{S} & \mathbf{T} \\ \mathbf{U} & \mathbf{V} \end{pmatrix} \;=\; 0$$

Expanding, we arrive (exercise) at three conditions on the block matrices:

$$\mathbf{S}^\dagger \;=\; -\,\mathbf{S}, \qquad\qquad \mathbf{V}^\dagger \;=\; -\,\mathbf{V}, \qquad\qquad \mathbf{T}^\dagger \;=\; \mathbf{U}$$

Both the $\mathbf{S}$ and $\mathbf{V}$ diagonal blocks are antihermitian. The off-diagonal blocks are each other's adjoint. Over $\mathbb{R}$, this means that $\mathbf{A}$ has two antisymmetric diagonal block matrices, one $q\times q$ and one $p\times p$; the off-diagonal blocks are the transpose of one another.

If $q = 1$ (pseudo-Euclidean case), S is a number which is pure imaginary over $\mathbb{C}$ and 0 over $\mathbb{R}$.

There only remains to notice that the non-zero elements of the infinitesimal generators can only be ± 1 (over $\mathbb{R}$) and also $\pm\,\mathrm{i}$ (over $\mathbb{C}$) because of the linearisation.

As we saw, another important constraint can be imposed on a group matrix $\mathbf{M}$: $\det \mathbf{M} = 1$, which defines $SL(n, \mathbb{R})$. Anticipating a little, its algebra can be written as a sum of matrices $\mathbf{C}_i$, and we have $\mathbf{M} = \exp \mathbf{C}_1 \exp \mathbf{C}_2 \exp \mathbf{C}_3 \cdots$.

Now, since the determinant of a product of matrices is equal to the product of the determinants of the matrix factors, and because—when a matrix $\mathbf{A}$ is diagonalisable— there exists a similarity transformation $\mathbf{SAS}^{-1}$ which takes $\mathbf{A}$ to $\mathbf{A}'$, the diagonal matrix with the eigenvalues A_i of $\mathbf{A}$ as elements, we conclude that $\det \mathbf{A}$ is equal to the product of the eigenvalues of $\mathbf{A}$.

Also:

$$\mathbf{S}\,\mathrm{e}^{\mathbf{A}}\,\mathbf{S}^{-1} \;=\; \mathbf{S}\,\mathbf{I}\mathbf{S}^{-1} + \mathbf{S}\,\mathbf{A}\,\mathbf{S}^{-1} + \frac{1}{2!}\,\mathbf{S}\,\mathbf{A}\,\mathbf{S}^{-1}\,\mathbf{S}\,\mathbf{A}\,\mathbf{S}^{-1} + \ldots \;=\; \mathbf{I} + \mathbf{A}' + \frac{1}{2!}\,(\mathbf{A}')^2 + \ldots \;=\; \mathrm{e}^{\mathbf{A}'}$$

where $\mathrm{e}^{\mathbf{A}'}$ is a diagonal matrix with e^{A_i} as entries. In other words, the eigenvalues of $\mathrm{e}^{\mathbf{A}}$ are just e^{A_i}. Then:

$$\det \mathrm{e}^{\mathbf{A}} \;=\; \prod_i \mathrm{e}^{A_i} \;=\; \exp \sum_i A_i \;=\; \mathrm{e}^{\mathrm{Tr}\,\mathbf{A}'}$$

But $\mathrm{Tr}\,\mathbf{A}' = \mathrm{Tr}(\mathbf{SAS}^{-1}) = \mathrm{Tr}\,\mathbf{A}$. We obtain via this elegant (but limited to diagonalisable matrices!) derivation an important *basis-independent* relation:

$$\det\left(\mathrm{e}^{\mathbf{A}}\right) \;=\; \mathrm{e}^{\mathrm{Tr}\,\mathbf{A}} \tag{3.12}$$

valid for *any* square matrix. This is immediately extended to $\det\left(\mathrm{e}^{\mathbf{A}}\mathrm{e}^{\mathbf{B}}\cdots\right) = \mathrm{e}^{\mathrm{Tr}\,(\mathbf{A}+\mathbf{B}+\ldots)}$, and we immediately deduce that all matrices in the algebra $\mathfrak{sl}(n,\,\mathbb{R})$ must have vanishing trace. This includes the generators that live in $\mathfrak{su}(n)$ and $\mathfrak{so}(n)$. This is why it can be said that the $\mathfrak{sl}(n,\,\mathbb{R})$ algebra consists of the set of all traceless matrices of rank n.

Note the interesting fact that since antisymmetric real matrices must be traceless, $\mathfrak{o}(n)$ and $\mathfrak{so}(n)$ are identical! This is very much related to the absence of a continuous path from the $O(n)$ identity (which is unimodular) to orthogonal matrices with determinant -1: $O(n)$ is not connected. For instance, spatial inversions cannot be linearised; one cannot invert axes by a "small" amount! It does not make sense to speak of the infinitesimal generators of spatial inversions if we cannot take a system *gradually*, in steps as small as we wish, from its initial state to its final transformed state. In three dimensions parametrised with spherical coordinates, spatial inversion is performed by $f(r, \theta, \phi) \to f(r,\, \pi - \theta,\, \pi + \phi)$, but without intermediate steps. So the infinitesimal generators of $O(3)$ are those corresponding to its connected $SO(3)$ subgroup, the group of rotations.

While we are at it, and since the exponential representation of group matrices in terms of their algebra matrices can allow us to reconstruct the group elements from the algebra, we note an important but difficult to prove expression which says that the familiar rule $\mathrm{e}^a\mathrm{e}^b = \mathrm{e}^{a+b}$ does not hold for matrices *unless they commute*! This is the so-called **Baker-Campbell-Hausdorff** (BCH) formula:

$$\mathrm{e}^{\mathbf{A}}\mathrm{e}^{\mathbf{B}} = \mathrm{e}^{\mathbf{C}} \qquad \mathbf{C} \;=\; \mathbf{A} + \mathbf{B} + \frac{1}{2}[\mathbf{A}, \mathbf{B}] + \frac{1}{12}\Big(\big[\mathbf{A}, [\mathbf{A}, \mathbf{B}]\big] + \big[[\mathbf{A}, \mathbf{B}],\, \mathbf{B}\big]\Big) + \ldots \tag{3.13}$$

By comparison, it requires little effort to show (EXERCISE) that:

$$\mathrm{d}_x \mathbf{A}^{-1} \;=\; -\,\mathbf{A}^{-1}\,(\mathrm{d}_x \mathbf{A})\,\mathbf{A}^{-1}$$

Also, if $\mathbf{A}$ and $\mathbf{B}$ are square matrices, and if the explicit dependence of another matrix $\mathbf{C}$ on a real parameter λ is given by $\mathbf{C}(\lambda) = \mathrm{e}^{\lambda\mathbf{B}}\mathbf{A}\mathrm{e}^{-\lambda\mathbf{B}}$, expanding plus repeated differentiation with respect to λ produces (EXERCISE) another useful expression:

$$\mathrm{e}^{\mathbf{B}}\,\mathbf{A}\,\mathrm{e}^{-\mathbf{B}} \;=\; \mathbf{A} \,+\, [\mathbf{B},\, \mathbf{A}] \,+\, \frac{1}{2}\,\big[\mathbf{B},\, [\mathbf{B},\, \mathbf{A}]\big] \,+\, \ldots \,+\, \frac{1}{n!}\,\big[\mathbf{B}, [\mathbf{B}, \ldots\, [\mathbf{B},\, \mathbf{A}]\big] \,+\, \ldots$$

This relation also holds for linear operators.

Example 3.7. The generators of $SO(3)$ live in a three-parameter algebra:

$$\mathfrak{so}(3) \;=\; \begin{pmatrix} 0 & -\theta_z & \theta_y \\ \theta_z & 0 & -\theta_x \\ -\theta_y & \theta_x & 0 \end{pmatrix} \;=\; \theta^i\, \mathbf{M}_i^{(3)} \tag{3.14}$$

But how did we pick the signs above the diagonal? Consider *counterclockwise* rotations by a small angle θ around an axis whose direction is specified by the vector $\hat{\mathbf{n}}$. An active transformation rotates a vector $\mathbf{x}$ by adding a small vector that is perpendicular to both the axis and to $\mathbf{x}$. By geometry, we find that, to first-order, the transformed vector is $\mathbf{x}' = \mathbf{x} + \theta\hat{\mathbf{n}} \times \mathbf{x}$. Expanding gives:

$$x' \;\approx\; x + \theta(n_y\, z - n_z\, y) \qquad y' \;\approx\; y + \theta(n_z\, x - n_x\, z) \qquad z' \;\approx\; z + \theta(n_x\, y - n_y\, x)$$

$$\Longleftrightarrow \qquad \mathbf{x}' = \mathbf{x} \;+\; \begin{pmatrix} 0 & -\theta_z & \theta_y \\ \theta_z & 0 & -\theta_x \\ -\theta_y & \theta_x & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

where $\boldsymbol{\theta} = \theta\hat{\mathbf{n}}$. The matrix is indeed the $\mathfrak{so}(3)$ algebra of eq. (3.14). How does this compare to the operator algebra as laid out in eq. (3.4) and the couple of equations that follow it? By identifying $\alpha = \theta n_z$, etc., we can write the first order in the expansion of the general rotation operator as:

$$\begin{pmatrix} x & y & z \end{pmatrix} \begin{pmatrix} 0 & -\theta_z & \theta_y \\ \theta_z & 0 & -\theta_x \\ -\theta_y & \theta_x & 0 \end{pmatrix} \begin{pmatrix} \partial_x \\ \partial_y \\ \partial_z \end{pmatrix}$$

The matrix is the $\mathfrak{so}(3)$ algebra matrix. The operator algebra has the same commutator structure as $\mathfrak{so}(3)$, which establishes the isomorphism of the two algebras.

A rotation by a *finite* angle θ around an axis $\hat{\mathbf{n}}$ can now be written as:

$$R(\boldsymbol{\theta}) \;=\; \mathrm{e}^{-\mathrm{i}\theta^k \mathbf{L}_k} \;=\; \mathrm{e}^{-\mathrm{i}\theta\hat{\mathbf{n}}\cdot\mathbf{L}}$$

where, as we saw we could do with the real operators, we have multiplied the real matrices by i to obtain the Hermitian angular momentum matrices familiar from quantum mechanics:

$$\mathbf{L}_x \;=\; \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\mathrm{i} \\ 0 & \mathrm{i} & 0 \end{pmatrix}, \qquad \mathbf{L}_y \;=\; \begin{pmatrix} 0 & 0 & \mathrm{i} \\ 0 & 0 & 0 \\ -\mathrm{i} & 0 & 0 \end{pmatrix}, \qquad \mathbf{L}_z \;=\; \begin{pmatrix} 0 & -\mathrm{i} & 0 \\ \mathrm{i} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

or, rather more succinctly: $(\mathbf{L}_i)^j{}_k = -\mathrm{i}\epsilon_{ijk}$. Sometimes, the $SO(3)$ generators are written as $J_{ij} = \epsilon_{ijk}J^k$, with matrix elements $(J_{ij})^l{}_m = -\mathrm{i}\epsilon_{ijk}\epsilon^{klm} = -\mathrm{i}(\delta_i{}^l\delta_j{}^m - \delta_i{}^m\delta_j{}^l)$. The labels $(ij) \quad i < j$ refer to the *plane* in which the rotation is generated. This allows a straightforward generalisation to rotations in $n > 3$ dimensions, where an axis does not uniquely define a plane of rotation.

Example 3.8. The $\mathfrak{so}(4)$ Lie algebra is the set of all antisymmetric 4×4 real matrices. It is now appropriate to use the $4(4-1)/2 = 6$ generators: $J_{ij} = \epsilon_{ijk}J^k \quad (i < j)$ introduced in example 3.7, that generate rotations in the (ij)-plane. Their matrix elements are: $(J_{ij})^l{}_m = -\epsilon_{ijk}\epsilon^{klm} = -(\delta_i{}^l\delta_j{}^m - \delta_i{}^m\delta_j{}^l)$. A general expression for their commutators can now be obtained by computing (EXERCISE): $\left[J_{mn},\, J_{pq}\right]^i{}_j = (J_{mn})^i{}_k(J_{pq})^k{}_j - (J_{pq})^i{}_l(J_{mn})^l{}_j$ and rearranging the eight resulting terms, yielding the important and *general* N-dim result:

$$[J_{mn}, J_{pq}] \;=\; \delta^p{}_m\, J_{nq} \;+\; \delta^q{}_n\, J_{mp} \;-\; \delta^q{}_m\, J_{np} \;-\; \delta^p{}_n\, J_{mq} \qquad 1 \le m < n \le N,\; 1 \le p < q \le N \tag{3.15}$$

One sees (EXERCISE) that only one term on the right can contribute, and that the commutator vanishes unless one (and only one) number in the pair (ij) is equal to one (and only one) number in the pair (pq). It is now almost trivial to compute the nine non-trivial $\mathfrak{so}(4)$ commutators. With $N_i = J_{i4}$ and $M_i = \epsilon_{ijk}J^{jk}$ $(1 \leq i,\, j < k \leq 3)$, they can be written as:

$$[M_i,\, M_j] \;=\; \epsilon_{ij}{}^{k}\, M_k, \qquad\qquad [M_i,\, N_j] \;=\; \epsilon_{ij}{}^{k}\, N_k, \qquad\qquad [N_i,\, N_j] \;=\; \epsilon_{ij}{}^{k}\, M_k \tag{3.16}$$

The commutators can be decoupled by transforming to the basis:

$$Y_i \;= \frac{1}{2}\,(M_i \,+\, N_i) \qquad\qquad Z_i \;=\; \frac{1}{2}\,(M_i \,-\, N_i)$$

from which we immediately obtain the *decoupled* relations:

$$[Y_i,\, Y_j] \;=\; \epsilon_{ij}{}^{k}\, Y_k, \qquad\qquad [Y_i,\, Z_j] \;=\; 0, \qquad\qquad [Z_i,\, Z_j] \;=\; \epsilon_{ij}{}^{k}\, Z_k \tag{3.17}$$

By inspection, the Y_i and Z_i separate into two $\mathfrak{so}(3)$ (or $\mathfrak{su}(2)$) algebras, and $\mathfrak{so}(4) = \mathfrak{so}(3) \oplus \mathfrak{so}(3)$ (or $\mathfrak{su}(2) \oplus \mathfrak{su}(2)$). In terms of dimensions, $\mathbf{6} = \mathbf{3} \oplus \mathbf{3}$, and at group level, $SO(4)$ is isomorphic to the direct product $SO(3) \times SO(3)$.

The algebra now reads: $\mathfrak{so}(4) \;=\; a^i\, Y_i \,+\, b^i\, Z_i$, and since $[Y_i, Z_j] = 0$, an element of $SO(4)$ takes the form; $\mathrm{e}^{a^i Y_i}\, \mathrm{e}^{b^i Z_i}$.

Example 3.9. With the three angles θ_i rotating around Cartesian axis $1 \leq i \leq 3$, and the rapidity parameters $\zeta_i = \hat{\beta}_i\, \tanh^{-1}\beta$ for pure Lorentz boosts along the x, y and z axes, the $\mathfrak{so}(3,\, 1)$ algebra of the Lorentz group $SO(3,1)$ derived from the metric-preserving constraint is:

$$\mathfrak{so}(3,1) \;=\; \mathsf{L} \;=\; \begin{pmatrix} 0 & \zeta_x & \zeta_y & \zeta_z \\ \zeta_x & 0 & -\theta_z & \theta_y \\ \zeta_y & \theta_z & 0 & -\theta_x \\ \zeta_z & -\theta_y & \theta_x & 0 \end{pmatrix} \;=\; \theta^\mu\, \mathsf{M}_\mu \,+\, \zeta^\nu\, \mathsf{K}_\nu \tag{3.18}$$

where the infinitesimal generators can be read off:

$$\mathsf{M}_x \;=\; \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \qquad \mathsf{M}_y \;=\; \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix} \qquad \mathsf{M}_z \;=\; \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

and

$$\mathsf{K}_x \;=\; \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \qquad \mathsf{K}_y \;=\; \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \qquad \mathsf{K}_z \;=\; \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

One shows (EXERCISE) that the commutators of the infinitesimal generators are:

$$[M_i,\, M_j] \;=\; \epsilon_{ij}{}^{k}\, M_k \qquad\qquad [M_i,\, K_j] \;=\; \epsilon_{ij}{}^{k}\, K_k \qquad\qquad [K_i,\, K_j] \;=\; -\,\epsilon_{ij}{}^{k}\, M_k \tag{3.19}$$

There is an important difference between these and the superficially similar relations for $\mathfrak{so}(4)$ that we have derived in eq. (3.16): because of the minus sign in the third relation, the *complex* basis in which the commutators decouple is: $L_i^{\pm} = (M_i \pm \mathrm{i}K_i)/2$, yielding (EXERCISE) : $[L_i^{\pm},\, L_j^{\pm}] = \epsilon_{ij}{}^{k} L_k^{\pm}$ and $[L_i^{\pm},\, L_j^{\mp}] = 0$. This means that: $\mathfrak{so}(3,1) \cong \mathfrak{su}(2) \oplus \mathfrak{su}(2)$.

A matrix in $SO(3,1)$ can then be written: $\mathrm{e}^{\mathsf{L}} = \mathrm{e}^{\theta^\mu\, \mathsf{M}_\mu \,+\, \zeta^\nu\, \mathsf{K}_\nu} = \mathrm{e}^{(\theta^\mu - \mathrm{i}\,\zeta^\mu)\mathsf{L}_\mu^{+}}\, \mathrm{e}^{(\theta^\nu + \mathrm{i}\,\zeta^\nu)\mathsf{L}_\nu^{-}}$.

3.3.5 Hard-nosed questions about the exponential map — the fine print

Three theorems by Lie, which we have implicitly made use of, show that for any Lie group a Lie algebra can be found which is characterised by the structure constants. But can we go back from the algebra to the group? Up to now, we have relied on the exponential map to do this, but it is not always possible, at least with just one map.

We already know that we cannot recover orthogonal matrices of determinant -1 in $O(3)$ from the $\mathfrak{o}(3)$ algebra. Do we fare better with special linear transformations whose matrices $SL(n,\mathbb{R})$ all have determinant 1? No, according to the following counter-example (provided, I believe, by Cartan). One form of the $\mathfrak{sl}(2,\mathbb{R})$ algebra, which is the set of all traceless matrix operators, is: $\mathsf{X} = \begin{pmatrix} x_1 & x_2 - x_3 \\ x_2 + x_3 & -x_1 \end{pmatrix}$. Exponentiating gives the basis-independent result (EXERCISE):

$$\mathrm{e}^{\mathsf{X}} = \sum_n \frac{1}{n!}\mathsf{X}^n = \begin{cases} \mathsf{I}_2 \cosh r + \mathsf{X}\frac{\sinh r}{r} & r^2 > 0 \\ \mathsf{I}_2 + \mathsf{X} & r^2 = 0 \\ \mathsf{I}_2 \cos r + \mathsf{X}\frac{\sin r}{r} & r^2 < 0 \end{cases}$$

where $r^2 = x_1^2 + x_2^2 - x_3^2$. The results are basis-independent because $r^2 = -\det \mathsf{X}$. The structure is rather reminiscent of the light-cone structure that we get by endowing the parameter space $\mathbb{R}^3$ with an indefinite metric. Inside the light-cone, for any value of x_3, the values of the other two parameters are confined inside a circle of radius smaller than x_3. The corresponding generators map to *compact* group elements. Outside the light-cone, however, the generators can grow without restriction; they map to non-compact elements of $SL(2,\mathbb{R})$.

So far, so good. But a glance at the above expressions shows that $\mathrm{Tr}\,\mathrm{e}^{\mathsf{X}} \geq -2$ always. Yet $SL(2,\mathbb{R})$ has a large subset of elements with trace smaller than -2: matrices of the type $\begin{pmatrix} -\lambda & 0 \\ 0 & -1/\lambda \end{pmatrix}$ ($\lambda > 1$), for instance. These cannot be reached with the above exponential map.

Cartan argued that all the group elements could nevertheless be reached by a product of two exponentials, one of the non-compact generators, and one of the compact generators. In our case, one would write:

$$\mathsf{X} = \mathsf{X}_a + \mathsf{X}_b = \begin{pmatrix} x_1 & x_2 \\ x_2 & -x_1 \end{pmatrix} + \begin{pmatrix} 0 & -x_3 \\ x_3 & 0 \end{pmatrix}$$

Then (EXERCISE):

$$\mathrm{e}^{\mathsf{X}_a}\,\mathrm{e}^{\mathsf{X}_b} = \begin{pmatrix} z + y & x \\ x & z - y \end{pmatrix} \begin{pmatrix} \cos x_3 & -\sin x_3 \\ \sin x_3 & \cos x_3 \end{pmatrix}$$

where $z \equiv \cosh r' \geq 1$, $x \equiv \frac{x_2}{r'}\sinh r'$, and $y \equiv \frac{x_1}{r'}\sinh r'$, with $r'^2 = x_1^2 + x_2^2$. Each matrix is unimodular, and the trace of the product is now $2z\cos x_3 = 2\cosh r' \cos x_3$ which is unrestricted. One exponential, $\mathrm{e}^{\mathsf{X}_b}$, corresponds to a compact generator, and the other to non-compact generators.

In example 3.3 we noted that we needed more tools to tell us what the manifold of $SL(2,\mathbb{R})$ was. Now we know! The parameters of the non-compact matrix satisfy $z^2 - (x^2 + y^2) = 1$ which is the positive-z hyperboloid. *Topologically*, it is equivalent to $\mathbb{R}^2$. The parameter values $-\pi \leq x_3 \leq \pi$ map the X_b subalgebra to $SO(2) \subset SL(2,\mathbb{R})$, whose manifold is S^1. We conclude that $SL(2,\mathbb{R})$ is non-compact, and that its manifold is $\mathbb{R}^2 \times S^1$.

These considerations do not detract from the usefulness of the exponential map as a link between a Lie algebra and a Lie group; they just mean that in the case of non-compact groups a little more care should be exercised.

3.4 Representations of Lie Groups and Algebras

3.4.1 Representations of Lie Groups

Definition 3.8. As for finite groups, a **representation** T_{g} of a Lie group G ($\mathrm{g} \in G$) is a homomorphism of G to the group of general linear matrix transformations $GL(\mathcal{V})$ acting on a vector space $\mathcal{V}$ called its **carrier space**.

For compact Lie groups, $\mathcal{V}$ is a finite-dimensional Hilbert space $\mathcal{H}$, ie. a vector space over $\mathbb{C}$ equipped with an inner product. For non-compact groups, however, it may well happen that the Hilbert space is infinite-dimensional.

Of particular interest are the irreducible representations of G. They satisfy Schur's lemma in its version for continuous groups:

A unitary representation T_{g} of a group G is irreducible if, and only if, the only operator A on $\mathcal{H}$ that satisfies $A\,\mathsf{T}_{\mathrm{g}} = \mathsf{T}_{\mathrm{g}}\,A\;\forall \mathrm{g} \in G$ is a multiple of the identity.

The following statements, which we quote without proof, apply to compact Lie groups:

- An irreducible representation of a compact Lie group must be equivalent to an unitary representation. All unitary representations of a compact Lie group are finite-dimensional. Thus, all irreducible representations are finite-dimensional.
- Every representation of a compact Lie group that is not already irreducible is fully reducible, in the sense that it can be written as the direct sum of irreducible unitary representations.
- An important theorem by Peter and Weyl states that:

  Let $\{\mathsf{T}^{(\alpha)}\}$ be the set of all irreducible non-equivalent unitary representations of a *compact* Lie group. Then the functions $\sqrt{n_\alpha}\,T^{(\alpha)}_{ij}$, where n_α is the dimension of the representation and $1 \le i, j \le n_\alpha$, form an orthonormal basis for $(\mathcal{L}_2)_G$, the Hilbert space of square-integrable functions on the manifold of G; that is, if $f_{\mathrm{g}} \in (\mathcal{L}_2)_G$, we have that:

$$f_{\mathrm{g}} \;=\; \sum_{\alpha,i,j} b^{(\alpha)}_{ij}\,\big(T^{(\alpha)}_{\mathrm{g}}\big)_{ij}, \qquad\qquad b^{(\alpha)}_{ij} \;=\; n_\alpha \int_G f_{\mathrm{g}}\,\big(T^{(\alpha)}_{\mathrm{g}}\big)^*_{ij}\,\mathrm{d}\mu_{\mathrm{g}}$$

  where $\mathrm{d}\mu_{\mathrm{g}}$ is the volume element of the group manifold.

  Example 3.10. Consider the Abelian group $U(1)$ with S^1 as its manifold. Its unitary irreducible representations are of course 1-dim: $T^{(m)}(\theta) = \mathrm{e}^{\mathrm{i}m\theta}$, ($m \in \mathbb{Z}$ because $T^{(m)}(\theta + 2\pi) = T^{(m)}(\theta)$). Then the Peter-Weyl theorem asserts that $\{T^{(m)}\}$ is a basis for $\mathcal{L}_2(S^1)$, the space of differentiable square-integrable functions on S^1, which must be *periodic*:

$$f(\theta) \;=\; \sum_{m=-\infty}^{\infty} c_m\,\mathrm{e}^{\mathrm{i}m\theta} \qquad\qquad c_m \;=\; \int_0^{2\pi} f(\theta)\,\mathrm{e}^{-\mathrm{i}m\theta}\,\mathrm{d}\theta$$

  which we recognise as a Fourier expansion.

3.4.2 Representations of Lie algebras

Lie algebras, as we have seen, can be realised as (differential) operators, or also as $\mathrm{gl}(\mathcal{V})$, the set of all linear transformations on some vector space $\mathcal{V}$. $\mathcal{V}$ will be a Hilbert space over $\mathbb{R}$ or $\mathbb{C}$, and we will have $\mathfrak{gl}(n,\mathbb{R})$ or $\mathfrak{gl}(n,\mathbb{C})$ realised as $n \times n$ real or complex matrices. In fact, a finite-dimensional algebra will always be isomorphic to some matrix algebra.

Definition 3.9. Let $\mathfrak{g}$ be a Lie algebra. A **representation** T of $\mathfrak{g}$ maps elements of the algebra to elements of the general linear invertible matrix transformations on its **carrier space** (or **module**) $\mathcal{V}$. The mapping is a homomorphism. The **dimension of a representation** is that of its carrier space.

$\mathfrak{g}$ has a Lie bracket, the commutator, and its representations must satisfy this product. Thus, if T is a representation of $\mathfrak{g}$, we must have $\forall\, X,\ Y \in \mathfrak{g}$:

$$\mathsf{T}_{[X,Y]} \;=\; \big[\mathsf{T}_X,\,\mathsf{T}_Y\big]$$

The **defining (fundamental) representation** of $\mathfrak{g}$ is the matrix obtained directly from the constraints on the algebra.

3.4.3 The regular (adjoint) representation and the classification of Lie algebras

We have already noted how eq. (3.11) for the structure constants could be written as the commutator of matrices whcih we now recognise as providing a new representation of the algebra:

Definition 3.10. The **regular (adjoint) representation** of a Lie algebra associates with each element Z of the algebra a matrix $\mathbf{R}_Z$ (or ad_Z) such that $[Z,\, X_i] = \left(R_Z\right)_i{}^j\, X_j$, where the X_i are the infinitesimal generators that form a basis for the algebra. (Warning! Some authors use the definition $[Z,\, X_i] = \left(R_Z\right)^j{}_i\, X_j$.)

Clearly, then, the regular representation of a basis generator is just the structure constants: $[X_i,\, X_j] = \left(R_{X_i}\right)_j{}^k\, X_k = C_{ij}{}^k\, X_k$. And its dimension is that of the algebra, ie. the number of generators (or parameters).

One can think of the regular representation as a matrix operator that acts on elements of the Lie algebra: $\mathrm{ad}_X(Y) = [X,\, Y]$. If X_i is a generator, ie. a basis element of the algebra, then $\mathrm{ad}_{X_i}(X_j) = [X_i,\, X_j] = C_{ij}{}^k X_k$, and the regular representation of X_i is the matrix ad_{X_i} whose elements are $(\mathrm{ad}_{X_i})_j{}^k = C_{ij}{}^k$, as before.

We confirm that $\mathbf{R}$ is a representation (EXERCISE, using the Jacobi identity): $[\mathbf{R}_{X_i},\, \mathbf{R}_{X_j}]X_k = \mathbf{R}_{[X_i,\, X_j]}X_k$.

Example 3.11. Take the defining, two-dimensional representation of the essentially real version of the $\mathfrak{su}(2)$ algebra with as basis set the three *Hermitian* Pauli matrices, $\boldsymbol{\sigma}_i$, with commutators: $[\boldsymbol{\sigma}_i,\, \boldsymbol{\sigma}_j] = 2\mathrm{i}\,\epsilon_{ij}{}^k\boldsymbol{\sigma}_k$. Then $\mathrm{ad}_{\boldsymbol{\sigma}_i}(\boldsymbol{\sigma}_j) = 2\mathrm{i}\,\epsilon_{ij}{}^k\boldsymbol{\sigma}_k$, and $(\mathrm{ad}_{\boldsymbol{\sigma}_i})_j{}^k = 2\mathrm{i}\,\epsilon_{ij}{}^k$, and we have:

$$\mathrm{ad}_{\boldsymbol{\sigma}_1} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 2\mathrm{i} \\ 0 & -2\mathrm{i} & 0 \end{pmatrix} \qquad \mathrm{ad}_{\boldsymbol{\sigma}_2} = \begin{pmatrix} 0 & 0 & -2\mathrm{i} \\ 0 & 0 & 0 \\ 2\mathrm{i} & 0 & 0 \end{pmatrix} \qquad \mathrm{ad}_{\boldsymbol{\sigma}_3} = \begin{pmatrix} 0 & 2\mathrm{i} & 0 \\ -2\mathrm{i} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

If we write the regular representation of a generic element of this algebra as $\mathbf{R} = a^i\mathrm{ad}_{\boldsymbol{\sigma}_i}$, we obtain the Hermitian matrix:

$$\mathbf{R} = \begin{pmatrix} 0 & 2\mathrm{i}\,a_3 & -2\mathrm{i}\,a_2 \\ -2\mathrm{i}\,a_3 & 0 & 2\mathrm{i}\,a_1 \\ 2\mathrm{i}\,a_2 & -2\mathrm{i}\,a_1 & 0 \end{pmatrix}$$

Note that the commutation relations $[\boldsymbol{\sigma}_i,\, \boldsymbol{\sigma}_j] = 2\mathrm{i}\,\epsilon_{ij}{}^k\boldsymbol{\sigma}_k$ can be written equivalently as: $[S_i,\, S_j] = \mathrm{i}\,\epsilon_{ij}{}^k S_k$, simply by the redefinition $\mathbf{S} = \boldsymbol{\sigma}/2$. This particular form is often preferred in quantum mechanics, particle physics and related fields.

Definition 3.11. A **subalgebra** of an algebra is just a subspace of the algebra $\mathfrak{g}$ which closes under commutation. A subalgebra $\mathfrak{g}_{\mathrm{sub}}$ is **invariant** if, for each $X \in \mathfrak{g}_{\mathrm{sub}}$ and $\forall\, Y \in \mathfrak{g}$, $[X,\, Y] \in \mathfrak{g}_{\mathrm{sub}}$. An invariant subalgebra is sometimes called an **ideal**, but we shall not be using this term.

The **centre** $\mathfrak{z}$ of an algebra is the largest subalgebra whose elements commute with *all* elements of the algebra. This means that, for a commutative (Abelian) algebra, the centre is the algebra itself. By definition, $\mathfrak{z}$ is an Abelian invariant subalgebra, but not not necessarily the only one.

An invariant subalgebra generates an invariant subgroup via the exponential map. To prove this, we must show that if $\mathrm{g} = \mathrm{e}^Y$ and $\mathrm{g}' = \mathrm{e}^{Y'}$, where Y is a generator in an invariant subalgebra and $Y' \in \mathfrak{g}$, then $\mathrm{g}'^{-1}\mathrm{g}\,\mathrm{g}' = \mathrm{e}^{Y'^{-1}}\mathrm{e}^Y\mathrm{e}^{Y'} = \mathrm{e}^Z$ is also in the subgroup generated by the subalgebra. But this is the case by virtue of the Baker-Campbell-Hausdorff formula, eq. (3.13), since Z is calculated from commutators of Y with other generators and so must be in the *invariant* subalgebra.

Definition 3.12. If a sequence of transformations exists that puts the regular representation of a *non-Abelian* Lie algebra into block-diagonal form, with the blocks irreducible *non-zero* subrepresentations, the representation is said to be **fully reducible**. In this case, the regular representation can be written as a direct sum of irreducible representations. Of course, these irreducible representations cannot all be one-dimensional. In this basis, the block submatrices commute with one another.

Like the structure constants, the regular representation summarises the structure of the Lie algebra. This algebra is a vector space spanned by a basis of generators. But we can decide to transform to another basis via a similarity transformation. The question is: can we transform the regular representation to a basis where it takes a form that might help classify the algebra?

Definition 3.13. If an algebra has no non-trivial invariant subalgebra, the *regular* representation of a generic element of the algebra is irreducible and we say that the algebra is **simple**. In that case, the representation leaves no *proper* subspace of its carrier space invariant.

Definition 3.14. A Lie algebra that contains no Abelian invariant subalgebra is said to be **semisimple**, ie. it has zero centre (it contains no non-zero element that commutes with all the other elements). A semisimple Lie algebra is either simple or the sum of simple Lie algebras (that may occur more than once in the sum), and its representations are the direct sums of irreducible representations. A semisimple algebra always has at least two complementary invariant subalgebras, with all the generators of one commuting with all the generators of the other(s), but not amongst themselves (inside a block in the reduced adjoint representation).

From these two definitions it follows that all simple algebras are semisimple since they are already in (single) block form. Those semisimple algebras which are *not* simple must contain a proper, *non-Abelian*, invariant subalgebra.

Abelian Lie algebras (eg. $\mathfrak{u}(1)$, $\mathfrak{so}(2)$) are not semisimple, and therefore not simple. Apart from $\mathfrak{so}(4)$ (see below), the non-Abelian $\mathfrak{so}(n)$ algebras are all simple, and so are the $\mathfrak{su}(n)$ and $\mathfrak{sl}(n,\mathbb{R})$ algebras.

Example 3.12. From eq. (3.16), no basis generator of $\mathfrak{so}(4)$ commutes with all others: The algebra has no non-zero centre! It is therefore[†] semisimple. Its structure constants determine the regular representation of a generic element of $\mathfrak{so}(4)$ in block-diagonal form:

$$\mathbf{R} \;=\; \begin{pmatrix} 0 & a_3 & -a_2 & 0 & 0 & 0 \\ -a_3 & 0 & a_1 & 0 & 0 & 0 \\ a_2 & -a_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & b_3 & -b_2 \\ 0 & 0 & 0 & -b_3 & 0 & b_1 \\ 0 & 0 & 0 & b_2 & -b_1 & 0 \end{pmatrix}$$

The blocks cannot be further reduced, $\mathfrak{so}(3)$ being simple; $\mathfrak{so}(4)$ is a semisimple (but not simple) algebra.

3.4.4 The Cartan-Killing form

Again, we recall that a Lie algebra is a vector space. As such, not only does it have a basis which can be chosen at our convenience, it can also be equipped with a (non-unique!) inner product. One such inner product[‡] is:

$$(Y,\, Z) \;=\; \mathrm{Tr}\, \mathbf{Y}\, \mathbf{Z}$$

Definition 3.15. The **Cartan-Killing form** (CK-form) of a n-paraneter Lie algebra is a symmetric bilinear form defined as the inner product for elements of the adjoint (regular) representation:

$$(Y,\, Z) \;=\; \mathrm{Tr}\left(\mathbf{R}_Y\, \mathbf{R}_Z\right) \;=\; (\mathbf{R}_Y)_l^{\,k}\, (\mathbf{R}_Z)_k^{\,l} \tag{3.20}$$

The CK-form for basis generators X_i is easily calculated: $(X_i,\, X_j) = C_{il}{}^k\, C_{jk}{}^l$. It has $n(n+1)/2$ components.

[†]The fact that $\{Y_i,\, Z_j\}$ in the decoupled basis of eq. (3.17) are Abelian subalgebras does not invalidate our argument, because these subalgebras are not invariant.

[‡]Another inner product which is sometimes introduced is the Hilbert-Schmidt product: $(\mathbf{A},\, \mathbf{B}) = \mathrm{Tr}(\mathbf{A}^\dagger \mathbf{B})$.

An important property of the CK-form is its invariance under the action of any element g in the Lie group associated with a Lie algebra. Let X and Y be elements of a Lie Algebra. Then:

$$\left(\mathrm{g}\,X\,\mathrm{g}^{-1},\,\mathrm{g}\,Y\,\mathrm{g}^{-1}\right) \;=\; \mathrm{Tr}\,\left(\mathbf{R}_{\mathrm{g}}\,\mathbf{R}_X\,\mathbf{R}_{\mathrm{g}^{-1}}\,\mathbf{R}_{\mathrm{g}}\,\mathbf{R}_Y\,\mathbf{R}_{\mathrm{g}^{-1}}\right) \;=\; \mathrm{Tr}\,(\mathbf{R}_X\,\mathbf{R}_Y) \;=\; (X,\,Y)$$

where we have used the property $\mathrm{Tr}\,\mathbf{A}\,\mathbf{B} = \mathrm{Tr}\,\mathbf{B}\,\mathbf{A}$. Linearising this after writing $\mathrm{g} = \mathrm{e}^{\epsilon Z}$, we obtain (EXERCISE):

$$\big([Z,\,X],\,Y\big) \;+\; \big(X,\,[Z,\,Y]\big) \;=\; 0 \tag{3.21}$$

This could be shown directly from $(Y,\,Z) = \mathrm{Tr}\,\mathbf{Y}\,\mathbf{Z}$, but it would not be so easy to interpret the result. Because it can be derived from the invariance of the CK-form under the group of which the algebra is the linearisation, we can say that eq. (3.21) is a manifestation of the invariance of the CK-form under the algebra as well as the group.

Definition 3.16. A CK-form is **degenerate** (or **singular**) if there exists at least one element Z in the algebra for which $(Z,\,Y) = 0\;\forall\,Y \in \mathfrak{g}$. In this case, the matrix with elements $(X_i,\,X_j)$ has a row and column entirely populated with zeros, which forces its determinant to vanish. Otherwise, the CK-form is **non-degenerate**

Alternatively, a CK-form is non-degenerate if there exists a basis in which all its diagonal elements are non-zero (so that its determinant does not vanish). Then we say that it induces a **Cartan metric g** on a Lie algebra, with components $g_{\mu\nu} = (X_\mu,\,X_\nu)$, where $\{X_\mu\}$ is a basis for the algebra. If the algebra is compact, we can transform to an orthonormal Cartan metric $\mathbf{g} = k\mathbf{I}_n$; if the algebra is non-compact, we can transform to an indefinite metric $k\mathbf{I}^q_p$, with $p+q=n$, the dimension of the algebra (and rank of the adjoint rep). In these two cases, it is habitual to call $\mathbf{I}_n$ and $\mathbf{I}^q_p$ themselves the metric, which is then manifestly orthonormal.

Like all metrics, an orthonormal Cartan metric can be used to raise and lower indices. In particular, introduce $f_{\mu\nu\lambda} \;=\; C_{\mu\nu}{}^{\rho}g_{\rho\lambda}$. Inserting $g_{\rho\lambda} = (X_\rho,\,X_\lambda)$, and using eq. (3.21), it is straightforward to show (EXERCISE) that $f_{\mu\nu\lambda}$ is antisymmetric in all its indices.

Now, if a Lie algebra has a non-zero centre (ie. a subalgebra whose elements commute with all the elements of the algebra), its CK-form must be degenerate because the adjoint representation of any element of the centre vanishes trivially. We have also said that the centre of an algebra is an Abelian invariant subalgebra. **Cartan's criterion** asserts that a CK-form is degenerate if, and only if, an algebra has an Abelian invariant subalgebra — that is, if there exists at least one generator which has vanishing commutator with all other generators.

This leads to the useful alternate definition:

Definition 3.17. A Lie algebra is **semisimple** if, and only if, its CK-form is non-degenerate,

Example 3.13. $x^i\partial_j$ is a basis of the operator realisation of $\mathfrak{gl}(3,\mathbb{R})$. Then $x^i\partial_i$ commutes with every other element of the algebra, and $\mathfrak{gl}(3,\mathbb{R})$ has a non-zero centre; its CK-form is degenerate, therefore, it is not semisimple.

Example 3.14. In example 3.11, we have already obtained the adjoint representation for the generators of $\mathfrak{su}(2)$. If we use the more standard representation corresponding to $\mathbf{S} = \boldsymbol{\sigma}/2$, then the adjoint representation matrix is merely divided by 2, and is also the one for $\mathfrak{so}(3)$ because the structure constants for the two algebras are now identical. With $\mathbf{S}$ in the adjoint rep, eq. (3.20) then gives:

$$(S_i,\,S_i) \;=\; \mathrm{Tr}\,(\mathbf{S}_1\mathbf{S}_1) \;=\; \mathrm{Tr}\,(\mathbf{S}_2\mathbf{S}_2) \;=\; \mathrm{Tr}\,(\mathbf{S}_3\mathbf{S}_3) \;=\; 2$$

with all other traces 0. The CK-form is then $2\mathbf{I}$. This confirms that the CK-form for $\mathfrak{su}(2)$ induces an invertible definite (Euclidean) orthonormal metric, $\mathbf{g} = \mathbf{I}$. Therefore, the group is compact as well as semisimple, and we can write the structure constants as the skew-symmetric $f_{ijk} = \mathrm{i}\epsilon_{ijk}$ (or in some sources just ϵ_{ijk}).

The CK-form as defined above in terms of the regular representation for each generator can be tedious to calculate if we need all its components. That representation generates huge matrices for any but the smallest groups. If it is invertible (non-degenerate), however, we can extract useful information from it with much less work by working in terms of the parameters a^μ such that $\mathbf{R} = a^\mu \mathbf{X}_\mu$:

$$(R,\, R) \;=\; a^\mu\, a^\nu\, \mathrm{Tr}\,(\mathbf{X}_\mu\, \mathbf{X}_\nu) \;=\; a^\mu\, a^n u\, g_{\mu\nu} \;=\; a^\mu\, a_\mu \tag{3.22}$$

where we have used the Cartan metric to lower the parameter index. Inspection of $a^\mu\, a_\mu$ will reveal whether the group is compact or non-compact.

Example 3.15. Going back to the defining representation that we wrote for $Z \in \mathfrak{sl}(2,\mathbb{R})$ in section (3.3.5):

$$\mathbf{Z} \;=\; \begin{pmatrix} x_1 & x_2+x_3 \\ x_2-x_3 & -x_1 \end{pmatrix} \;=\; x_1 \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} + x_2 \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + x_3 \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

The corresponding independent non-zero stucture constants are: $C_{12}{}^3 = 2$, $C_{31}{}^2 = -2$, and $C_{23}{}^1 = -2$. From these we build the regular-representation matrix:

$$\mathbf{R} \;=\; \begin{pmatrix} 0 & -2x_3 & -2x_2 \\ 2x_3 & 0 & 2x_1 \\ -2x_2 & 2x_1 & 0 \end{pmatrix}$$

Now, we only need to calculate the diagonal elements of $\mathbf{R}^2$ and sum them to get: $(R,\, R) = 8(x_1^2 + x_2^2 - x_3^2)$. We deduce that the algebra is non-compact: X_1 and X_2 are non-compact, while X_3 is compact. This is consistent with what we found earlier about the $SL(2,\mathbb{R})$ manifold.

Interestingly enough, using the defining representation directly, we would find (EXERCISE) $2(x_1^2 + x_2^2 - x_3^2)$. Far from being a coincidence, this opens up the possibility of calculating $a^\mu\, a_\mu$ in eq. (3.22) directly from the defining representation instead of the more unwieldy regular representation. This is because for semisimple algebras the defining and regular representations are both faithful, and thus contain the same amount of information.

3.4.5 Cartan subalgebra

Now we would very much like to find whether some elements H_i of a semisimple algebra have a *diagonalisable* adjoint-representation matrix, and satisfy the eigenvalue equation:

$$R_{H_i}(Y) \;=\; [H_i,\, Y] \;=\; \lambda_Y\, Y \tag{3.23}$$

for some $Y \in \mathfrak{g}$, which makes Y an **eigengenerator** of H_i. In fact, we would like to know the maximal subset of r *simultaneously* diagonalisable operators in a given Lie algebra. These operators, H_i, must all commute between themselves, ie. they form an *Abelian* (but non-invariant!) subalgebra $\mathfrak{H}$.

Definition 3.18. A maximal Abelian subalgebra of a semisimple Lie algebra is called a **Cartan subalgebra** $\mathfrak{h}$. Its dimension $r < n$ defines the **rank** of the algebra. It is not unique, although its dimension is. The elements of a Cartan subalgebra are called the **Cartan generators** of the algebra.

Example 3.16. Let $\{J_+,\; J_-,\, J_0\} = \{J_1 + \mathrm{i}J_2,\, J_1 - \mathrm{i}J_2, J_3\}$ be a basis of a Lie algebra, with $[J_i,\, J_j] = \mathrm{i}\epsilon_{ij}{}^k J_k$. These commutation relations translate into: $[J_0,\, J_\pm] = \pm J_\pm$, $[J_+,\, J_-] = 2J_0$. The corresponding adjoint-representation matrices of J_0 and J_+ are immediately found:

$$\mathrm{ad}_{J_0} \;=\; \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad\qquad \mathrm{ad}_{J_+} \;=\; \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 2 \\ -1 & 0 & 0 \end{pmatrix}$$

Because J_0 is diagonal, it is a Cartan generator; comparing with eq. (3.23), ad_{J_0} has a *complete* set $\{J_+,\ J_-,\ J_0\}$ of eigengenerators for the corresponding eigenvalues $\{1,\ -1,\ 0\}$. But ad_{J_+} has only one eigenvalue, 0, and therefore one eigengenerator which cannot form a basis of the algebra. It is not diagonalisable and not a Cartan generator; the same holds for J_-. Thus, the algebra contains only one Cartan generator and is of rank 1.

We arrive at the same results by working out the secular equation for eq. (3.23), $||\mathbf{R} - \lambda \mathbf{I}_3|| = 0$, for the *generic* regular matrix:

$$\mathbf{R} \;=\; a^k \,\mathrm{ad}_{J_k} \;=\; \begin{pmatrix} 0 & \mathrm{i}\,a_3 & -\mathrm{i}\,a_2 \\ -\mathrm{i}\,a_3 & 0 & \mathrm{i}\,a_1 \\ \mathrm{i}\,a_2 & -\mathrm{i}\,a_1 & 0 \end{pmatrix}$$

Then $||\mathbf{R} - \lambda \mathbf{I}_3|| = -\lambda(\lambda^2 + a^2) = 0$, where $a^2 = a_1^2 + a_2^2 + a_3^2$. The fact that the eigenvalues $\{0,\ \pm \mathrm{i}a\}$ are complex tells us that we should use the complex extension of the algebra (which indeed we did when using the basis $\{J_+,\ J_-,\ J_0\}$!), and the number of independent non-zero eigenvalues gives the rank of the algebra, here 1.

The secular equation itself also yields the rank of the algebra: the number of independent functions of the parameters (just one, a, in our example) in the secular equation is also the rank of the algebra. It can be shown that the rank of a $\mathfrak{su}(n)$ algebra is $n-1$; also, $\mathfrak{so}(2n)$ and $\mathfrak{so}(2n+1)$ algebras have rank n.

An important point is that the same analysis could have been carried through using the three 2×2 matrices of the defining representation of the algebra ($\mathfrak{su}(2)$), with the same conclusions. Indeed, in the usual Pauli matrix basis:

$$||\mathfrak{su}(2) - \lambda \mathbf{I}_2|| \;=\; \begin{vmatrix} a_3 - \lambda & a_1 - \mathrm{i}a_2 \\ a_1 + \mathrm{i}a_2 & -a_3 - \lambda \end{vmatrix} \;=\; \lambda^2 + \frac{1}{4}a^2 \;=\; 0$$

gives $\lambda = \pm \mathrm{i}a/2$. Again, the secular equation has one independent function of the parameters.

As noted previously, the advantage of the defining representation is that the dimension of its matrices is most often much smaller than those of the adjoint representation.

3.5 Weights and Roots of a Representation of a Compact Semisimple Lie Algebra

Definition 3.19. Let $|\mu\rangle$ be an eigenvector common to all Cartan basis generators H_i, living in the *carrier space* of some representation $\mathbf{D}$ of the generator. Then $H_i|\mu\rangle = \mu_i|\mu\rangle$. The set $\{\mu_i\}$ $(1 \leq i \leq r)$ corresponding to each eigenvector can be viewed as the components of a r-dimensional vector called a **weight $\boldsymbol{\mu}$** of the representation. The number of these weights, ie. the number of eigenvectors, is equal to the dimension of $\mathbf{D}$.

Definition 3.20. Following standard notation, let $\{E_{\boldsymbol{\alpha}} \in \mathfrak{g}\}$ denote the set of *eigengenerators* of an element H_i of the a Cartan subalgebra of the *complex extension* of a semisimple Lie algebra, such that in some representation $\mathbf{D}$, $[H_i,\ E_{\boldsymbol{\alpha}}] = \alpha_i E_{\boldsymbol{\alpha}}$, $1 \leq i \leq r$. Then the set of eigenvalues, $\{\alpha_i\}$, of the $[H_i,\ \ ]$ operator associated with each eigengenerator $E_{\boldsymbol{\alpha}}$ can be viewed as the components of a r-dimensional vector called the **root $\boldsymbol{\alpha}$**. We can also write $[\mathbf{H},\ E_{\boldsymbol{\alpha}}] = \boldsymbol{\alpha} E_{\boldsymbol{\alpha}}$.

Do keep in mind the crucial distinction between the eigengenerators, whose associated eigenvalues are the root components, and the eigenvectors that live in the carrier space and whose eigenvalues are the weights.

Clearly, any Cartan generator H_i has the root $\boldsymbol{\alpha} = 0$, because $[H_i,\ H_j] = 0\ \forall\, j$. Thus, we can write an algebra $\mathfrak{g}$ as the sum of its Cartan subalgebra, with roots zero, and the non-Cartan generators with non-zero roots. The set

of all non-zero roots is called the **root system** of the algebra. The Cartan and non-Cartan generators form a basis of the algebra.

As we are soon to discover, all the information about a semisimple algebra is encoded in its system of r-dimensional roots. Like the weight vectors, these roots can be seen as living in a r-dimensional space. A Euclidean metric is induced on this root space by the Euclidean metric of the Cartan subalgebra, so that we can represent it as having r Cartesian axes, each associated with a Cartan generator H_i. The root vectors can then be represented in a **root diagram**. The i^{th} component of each root is the projection of the root along the H_i axis. This root space is almost always much easier to work with than the original algebra, simply because it is of smaller dimension.

Also, each n-dimensional representation (often called a **multiplet**) of a semisimple Lie algebra is associated with n weights which are r-dim "vectors" with as i^{th} component the corresponding eigenvalue of the simultaneously diagonalisable Cartan generators, H_i, of the algebra, acting on their carrier function space. Just like roots, these weights can be plotted on a **weight diagram**, or **lattice**, with each eigenvalue component of a weight projected along the associated H_i axis.

Those eigengenerators of H_i, $E_{\boldsymbol{\alpha}} \in \mathfrak{g}$ (the latter is now the complex extension!), which are *not* Cartan generators are quite interesting. An important fact, which we shall not prove, is that they are uniquely labelled by their roots. To each non-zero root corresponds one and only one such generator, which spans a 1-dim subalgebra.

Now let $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ be two non-zero roots. Then, from the Jacobi identity and definition 3.20 of the roots, there comes:

$$\big[H_i,\,[E_{\boldsymbol{\alpha}},\,E_{\boldsymbol{\beta}}]\big] \;=\; \big[[H_i,\,E_{\boldsymbol{\alpha}}],\,E_{\boldsymbol{\beta}}\big] \;+\; \big[E_{\boldsymbol{\alpha}},\,[H_i,\,E_{\boldsymbol{\beta}}]\big] \;=\; (\alpha_i \,+\, \beta_i)\,[E_{\boldsymbol{\alpha}},\,E_{\boldsymbol{\beta}}]$$

Therefore, either $[E_{\boldsymbol{\alpha}},\,E_{\boldsymbol{\beta}}] = 0$, which happens when $\boldsymbol{\alpha}+\boldsymbol{\beta}$ is not a root, or it is proportional to the eigengenerator $E_{\boldsymbol{\alpha}+\boldsymbol{\beta}}$ of H_i, with eigenvalue $\alpha_i \,+\, \beta_i$. and we can write:

$$[E_{\boldsymbol{\alpha}},\,E_{\boldsymbol{\beta}}] \;=\; C_{\boldsymbol{\alpha}\boldsymbol{\beta}} E_{\boldsymbol{\alpha}+\boldsymbol{\beta}} \tag{3.24}$$

When either $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$ vanishes, $C_{\boldsymbol{\alpha}\boldsymbol{\beta}}$ is equal to the other root, and we recover the defining equation for the roots in def. 3.20.

Going to definition 3.10 of the adjoint represepresentation, one should now be able to see that the adjoint representation of a Cartan generator is a diagonal matrix, with r 0s and α_i as the rest of the $n-r$ diagonal entries. And the adjoint representation of any other generator $E_{\boldsymbol{\alpha}}$ must have all diagonal entries zero. From this, the following statements about the CK-form of an algebra can be derived (EXERCISE):

$$h_{ij} \;=\; (H_i,\,H_j) \;=\; \sum_{\boldsymbol{\alpha}} \alpha_i \alpha_j, \qquad\qquad (H_i,\,E_{\boldsymbol{\alpha}}) \;=\; 0, \qquad\qquad (E_{\boldsymbol{\alpha}},\,E_{\boldsymbol{\beta}}) \;=\; 0 \quad (\boldsymbol{\alpha}+\boldsymbol{\beta} \;\neq\; 0 \text{ and not a root})$$

where h_{ij} are the components of the metric of the Cartan subalgebra.

To go further, work with Hermitian Cartan generators: $H_i^\dagger \;=\; H_i$ of the essentially real algebra. Then, if $[H_i,\,E_{\boldsymbol{\alpha}}] \;=\; \alpha_i E_{\boldsymbol{\alpha}}$, we immediately find by taking the adjoint that $[H_i,\,E_{\boldsymbol{\alpha}}^\dagger] \;=\; -\alpha_i E_{\boldsymbol{\alpha}}^\dagger$, which we can write as $[H_i,\,E_{-\boldsymbol{\alpha}}] = -\alpha_i E_{-\boldsymbol{\alpha}}$. Thus, non-Cartan generators and non-zero roots always come in pairs, $\{E_{\boldsymbol{\alpha}},\,E_{-\boldsymbol{\alpha}} = E_{\boldsymbol{\alpha}}^\dagger\}$ and $\pm\boldsymbol{\alpha}$. In fact, $-\boldsymbol{\alpha}$ is the *only* possible multiple of $\boldsymbol{\alpha}$ that is a root. The generators $E_{\pm\boldsymbol{\alpha}}$ of the complex extension are computed from the pairs X_k and X_l of non-Cartan generators of the algebra as: $(X_k \pm \mathrm{i}X_l)/\sqrt{2}$.

When $\boldsymbol{\beta} = -\boldsymbol{\alpha}$, eq. (3.24) would map $[E_{\boldsymbol{\alpha}},\,E_{\boldsymbol{\beta}}]$ to a generator with zero root, ie. one that lives in the Cartan subalgebra. Therefore, $[E_{\boldsymbol{\alpha}},\,E_{\boldsymbol{\beta}}]$ must belong to $\mathfrak{H}$ as a linear combination of the r Cartan basis generators:

$$[E_{\boldsymbol{\alpha}},\,E_{-\boldsymbol{\alpha}}] \;=\; \lambda^i H_i \qquad 1 \leq i \leq r$$

Now take the inner product of this equation with H_j. For semisimple algebras, we know that $(X_i,\,X_j) \;=\; g_{ij}$. Then the right-hand side gives $\lambda^j h_{ij}$. For the left-hand side, we have:

$$\big(H_i,\,[E_{\boldsymbol{\alpha}},\,E_{-\boldsymbol{\alpha}}]\big) \;=\; \big(E_{-\boldsymbol{\alpha}},\,[H_i,\,E_{\boldsymbol{\alpha}}]\big) \;=\; \alpha_i\,(E_{-\boldsymbol{\alpha}},\,E_{\boldsymbol{\alpha}})$$

where we have used eq. (3.21) (or, alternatively, the cyclic property of the trace of a product). Therefore, $\lambda^i \;=\; h^{ij}\alpha_j(E_{-\boldsymbol{\alpha}},\,E_{\boldsymbol{\alpha}})$. Now the E generators live in the complex extension, so that $(E_{-\boldsymbol{\alpha}},\,E_{\boldsymbol{\alpha}}) = (X_k,\,X_k) = g_{kk}$.

We also recall that for a compact semisimple algebra, $\mathbf{g} = k\mathbf{I}$: the metric on $\mathfrak{g}$, including its restriction on $\mathfrak{H}$, is diagonal with identical entries. Then $h^{ii} = 1/g_{kk}$, and $\lambda^i = \alpha_i$. We arrive at yet another useful result for the commutator of two non-Cartan generators with opposite root vectors:

$$[E_{\boldsymbol{\alpha}},\, E_{-\boldsymbol{\alpha}}] \;=\; \alpha_i H_i \;\equiv\; \boldsymbol{\alpha}\cdot\mathbf{H} \tag{3.25}$$

Here, $\boldsymbol{\alpha}\cdot\mathbf{H}$ should be seen as just another way to write $\alpha_i H_i$, and there is a summation over i implied.

Now is a good time to discover what those non-Cartan generators of the complex extension do for a living. We have:

$$H_i\, E_{\pm\boldsymbol{\alpha}} \;=\; [H_i,\, E_{\pm\boldsymbol{\alpha}}] \;+\; E_{\pm\boldsymbol{\alpha}} H_i \;=\; \pm\alpha_i E_{\pm\boldsymbol{\alpha}} \;+\; E_{\pm\boldsymbol{\alpha}} H_i$$

When this acts on a common eigenvector of the H_i in the carrier space, there comes:

$$H_i\, E_{\pm\boldsymbol{\alpha}}|\mu\rangle \;=\; \pm\alpha_i E_{\pm\boldsymbol{\alpha}}|\mu\rangle \;+\; E_{\pm\boldsymbol{\alpha}} H_i|\mu\rangle \;=\; (\mu_i \pm \alpha_i) E_{\pm\boldsymbol{\alpha}}|\mu\rangle \tag{3.26}$$

We see that $E_{\pm\boldsymbol{\alpha}}|\mu\rangle$ is an eigenvector of H_i with eigenvalue $\mu_i \pm \alpha_i$, from which we conclude that the $E_{\pm\boldsymbol{\alpha}}$ act as raising and lowering operators on the carrier space of the Cartan generators, changing weights by $\pm\boldsymbol{\alpha}$. This result holds for any irreducible representation. Thus, if we can find the weight vectors, the root vectors must be all the possible differences between them; often, this is the quickest way to obtain the roots.

We now concern ourselves with finding the (real!) structure constants $C_{\boldsymbol{\alpha\beta}}$ in eq. (3.24). They satisfy several symmetry relations. For instance:

$$C_{\boldsymbol{\beta\alpha}} \;=\; -\,C_{\boldsymbol{\alpha\beta}} \qquad\qquad C_{-\boldsymbol{\alpha},-\boldsymbol{\beta}} \;=\; -\,C^*_{\boldsymbol{\alpha\beta}} \;=\; -\,C_{\boldsymbol{\alpha\beta}} \tag{3.27}$$

The first is read off the equation, while taking the adjoint of the equation gives the second relation.

Also, let $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}+\boldsymbol{\beta}$ be non-zero roots; then $\boldsymbol{\gamma} = -(\boldsymbol{\alpha}+\boldsymbol{\beta})$ is also a non-zero root, Using the Jacobi identity on $E_{\boldsymbol{\alpha}}$, $E_{\boldsymbol{\beta}}$, and $E_{\boldsymbol{\gamma}}$, plus eq. (3.25), leads (EXERCISE) to:

$$(\boldsymbol{\alpha}\, C_{\boldsymbol{\beta\gamma}} \;+\; \boldsymbol{\beta}\, C_{\boldsymbol{\gamma\alpha}} \;+\; \boldsymbol{\gamma}\, C_{\boldsymbol{\alpha\beta}})\cdot\mathbf{H} \;=\; 0$$

since the H_i are linearly independent, this equation can only be satisfied if:

$$\boldsymbol{\alpha}\, C_{\boldsymbol{\beta\gamma}} \;+\; \boldsymbol{\beta}\, C_{\boldsymbol{\gamma\alpha}} \;+\; \boldsymbol{\gamma}\, C_{\boldsymbol{\alpha\beta}} \;=\; \boldsymbol{\alpha}(C_{\boldsymbol{\beta\gamma}} \;-\; C_{\boldsymbol{\alpha\beta}}) \;+\; \boldsymbol{\beta}(C_{\boldsymbol{\gamma\alpha}} \;-\; C_{\boldsymbol{\alpha\beta}}) \;=\; 0$$

which yields additional symmetries on the structure constants of a semisimple algebra:

$$C_{\boldsymbol{\beta},-\boldsymbol{\alpha}-\boldsymbol{\beta}} \;=\; C_{-\boldsymbol{\alpha}-\boldsymbol{\beta},\boldsymbol{\alpha}} \;=\; C_{\boldsymbol{\alpha\beta}} \tag{3.28}$$

A very important result can now be derived.

Going back again to eq. (3.24), we can write: $[E_{\boldsymbol{\alpha}},\, E_{\boldsymbol{\beta}+\boldsymbol{\alpha}}] \;=\; C_{\boldsymbol{\alpha},\boldsymbol{\beta}+\boldsymbol{\alpha}} E_{\boldsymbol{\beta}+2\boldsymbol{\alpha}}$, …, $[E_{\boldsymbol{\alpha}},\, E_{\boldsymbol{\beta}+k\boldsymbol{\alpha}}] \;=\; C_{\boldsymbol{\alpha},\boldsymbol{\beta}+k\boldsymbol{\alpha}} E_{\boldsymbol{\beta}+(k+1)\boldsymbol{\alpha}}$. But there must exist a value $k = p \geq 0$ such that $\boldsymbol{\beta} + (p+1)\boldsymbol{\alpha}$ is not a root, so that $C_{\boldsymbol{\alpha},\boldsymbol{\beta}+p\boldsymbol{\alpha}} = 0$. Similarly, if we start from: $[E_{-\boldsymbol{\alpha}},\, E_{\boldsymbol{\beta}}] = C_{-\boldsymbol{\alpha},\boldsymbol{\beta}} E_{\boldsymbol{\beta}-\boldsymbol{\alpha}}$, there must exist a value $k = -q \leq 0$ such that $\boldsymbol{\beta} - (q+1)\boldsymbol{\alpha}$ is not a root, and $C_{-\boldsymbol{\alpha},\boldsymbol{\beta}-q\boldsymbol{\alpha}} = 0$.

Next, start from the always useful Jacobi identity and evaluate the commutators using eq. (3.24) and (3.25):

$$\begin{aligned}
\big[E_{\boldsymbol{\alpha}},\, [E_{\boldsymbol{\beta}+k\boldsymbol{\alpha}},\, E_{-\boldsymbol{\alpha}}]\big] \;+\; \big[E_{\boldsymbol{\beta}+k\boldsymbol{\alpha}},\, [E_{-\boldsymbol{\alpha}},\, E_{\boldsymbol{\alpha}}]\big] \;+\; \big[E_{-\boldsymbol{\alpha}},\, [E_{\boldsymbol{\alpha}},\, E_{\boldsymbol{\beta}+k\boldsymbol{\alpha}}]\big] \;&=\; 0 \\
[E_{\boldsymbol{\alpha}},\, E_{\boldsymbol{\beta}+(k-1)\boldsymbol{\alpha}}]\, C_{\boldsymbol{\beta}+k\boldsymbol{\alpha},-\boldsymbol{\alpha}} \;-\; [E_{\boldsymbol{\beta}+k\boldsymbol{\alpha}},\, \boldsymbol{\alpha}\cdot\mathbf{H}] \;+\; [E_{-\boldsymbol{\alpha}},\, E_{\boldsymbol{\beta}+(k+1)\boldsymbol{\alpha}}]\, C_{\boldsymbol{\alpha},\boldsymbol{\beta}+k\boldsymbol{\alpha}} \;&=\; 0 \\
C_{\boldsymbol{\alpha},\boldsymbol{\beta}+(k-1)\boldsymbol{\alpha}}\, C_{\boldsymbol{\beta}+k\boldsymbol{\alpha},-\boldsymbol{\alpha}} \;+\; \boldsymbol{\alpha}\cdot(\boldsymbol{\beta} \;+\; k\boldsymbol{\alpha}) \;+\; C_{-\boldsymbol{\alpha},\boldsymbol{\beta}+(k+1)\boldsymbol{\alpha}}\, C_{\boldsymbol{\alpha},\boldsymbol{\beta}+k\boldsymbol{\alpha}} \;&=\; 0
\end{aligned}$$

Applying the symmetry relations (3.27) and then (3.28) to the first and last term on the left yields the recursion relation:

$$C^2_{\boldsymbol{\alpha},\boldsymbol{\beta}+(k-1)\boldsymbol{\alpha}} \;=\; C^2_{\boldsymbol{\alpha},\boldsymbol{\beta}+k\boldsymbol{\alpha}} \;+\; \boldsymbol{\alpha}\cdot(\boldsymbol{\beta} \;+\; k\boldsymbol{\alpha})$$

We already know that, by definition of p, $C^2_{\boldsymbol{\alpha},\boldsymbol{\beta}+p\boldsymbol{\alpha}} = 0$. Then, from our recursion relation, $C^2_{\boldsymbol{\alpha},\boldsymbol{\beta}+(p-1)\boldsymbol{\alpha}} = \boldsymbol{\alpha}\cdot\boldsymbol{\beta} + p|\boldsymbol{\alpha}|^2$, $C^2_{\boldsymbol{\alpha},\boldsymbol{\beta}+(p-2)\boldsymbol{\alpha}} = C^2_{\boldsymbol{\alpha},\boldsymbol{\beta}+(p-1)\boldsymbol{\alpha}} + \boldsymbol{\alpha}\cdot\boldsymbol{\beta} + (p-1)|\boldsymbol{\alpha}|^2 = 2\boldsymbol{\alpha}\cdot\boldsymbol{\beta} + (p-2)|\boldsymbol{\alpha}|^2$, etc. Generically:

$$C^2_{\boldsymbol{\alpha},\boldsymbol{\beta}+(k-1)\boldsymbol{\alpha}} \;=\; (p-k+1)\left[\boldsymbol{\alpha}\cdot\boldsymbol{\beta} + \frac{p+k}{2}\,|\boldsymbol{\alpha}|^2\right]$$

The recursion stops when $k = -q$, ie. when:

$$0 \;=\; C^2_{\boldsymbol{\alpha},\boldsymbol{\beta}-(q+1)\boldsymbol{\alpha}} \;=\; (p+q+1)\left[\boldsymbol{\alpha}\cdot\boldsymbol{\beta} + \frac{p-q}{2}\,|\boldsymbol{\alpha}|^2\right]$$

or:

$$2\,\frac{\boldsymbol{\alpha}\cdot\boldsymbol{\beta}}{|\boldsymbol{\alpha}|^2} \;=\; -\,(p-q) \tag{3.29}$$

the so-called " **master formula**". We will use it to prove a crucial fact a little later.

Example 3.17. Start with the defining representation of the 3-dim $\mathfrak{su}(2)$ algebra:

$$J_1 \;=\; \frac{1}{2}\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix} \qquad J_2 \;=\; \frac{1}{2}\begin{pmatrix}0 & -\mathrm{i}\\ \mathrm{i} & 0\end{pmatrix} \qquad J_0 \;=\; \frac{1}{2}\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}$$

$\mathfrak{su}(2)$ is semisimple, and one Cartan and one pair of non-Cartan generators can fit in it, thus one *independent* non-zero root vector. This is an algebra of rank 1, and the roots are one-dimensional. The diagonal generator J_0 is identified with the sole Cartan generator. The weights of J_0 in the defining representation are $1/2$ and $-1/2$ (corresponding to the doublet of eigenvectors $\binom{1}{0}$ and $\binom{0}{1}$). Once the weights have been found, the roots must be among all the possible differences between the weights, ie. ± 1. These roots raise or lower the weights by 1.

Then, without any direct computation, we can state from eq. (3.25) that $[E_1,\, E_{-1}] = J_0$. Also, from the definition of roots, $[J_0,\, E_{\pm 1}] = \pm E_{\pm 1}$. Nailing down the structure of the algebra determines the non-Cartan generators. $[J_0,\, E_{\pm 1}] = \pm E_{\pm 1}$ gives, up to a normalisation constant A:

$$E_1 \;=\; A\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix} \;=\; A(J_1 + \mathrm{i}\,J_2), \qquad E_{-1} \;=\; A\begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix} \;=\; A(J_1 - \mathrm{i}\,J_2)$$

A is determined by the other commutation relation, yielding finally: $E_{\pm 1} = (J_1 \pm \mathrm{i}\,J_2)/\sqrt{2}$, as we should have expected. We recognise the set $\{J_0, E_{\pm 1}\}$ as the complex extension of $\mathfrak{su}(2)$.

Because $\mathfrak{su}(2)$ has rank 1, this is a somewhat trivial application of the machinery we are developing, but it comes into its own with algebras of higher rank.

Each pair $F_{\pm\boldsymbol{\alpha}} \equiv \sqrt{2}E_{\pm\boldsymbol{\alpha}}/|\boldsymbol{\alpha}|$ of normalised non-Cartan generators of a semisimple algebra, together with the combination: $\tilde{H}_{\boldsymbol{\alpha}} = 2\boldsymbol{\alpha}\cdot\mathbf{H}/|\boldsymbol{\alpha}|^2$, forms a $\mathfrak{su}(2)$ subalgebra. *There is a distinct* $\mathfrak{su}(2)$ *subalgebra for each pair of non-zero roots*. Indeed:

$$\left[\frac{2\boldsymbol{\alpha}\cdot\mathbf{H}}{|\boldsymbol{\alpha}|^2},\, \frac{\sqrt{2}E_{\pm\boldsymbol{\alpha}}}{|\boldsymbol{\alpha}|}\right] \;=\; 2\sqrt{2}\,\frac{\boldsymbol{\alpha}}{|\boldsymbol{\alpha}|^3}\cdot[\mathbf{H},\, E_{\pm\boldsymbol{\alpha}}] \;=\; \pm 2\sqrt{2}\,\frac{|\boldsymbol{\alpha}|^2}{|\boldsymbol{\alpha}|^3}\,E_{\pm\boldsymbol{\alpha}} \;=\; \pm 2\left(\frac{\sqrt{2}\,E_{\pm\boldsymbol{\alpha}}}{|\boldsymbol{\alpha}|}\right)$$

Together with eq. (3.25), we recover the $\mathfrak{su}(2)$ structure constants in the basis of example 3.16, with $\tilde{H}_{\boldsymbol{\alpha}} = 2J_0$ and $F_{\pm\boldsymbol{\alpha}} = J_\pm$. The non-immediately zero commutators with elements of the other $\mathfrak{su}(2)$ subalgebras can also be found. For instance, one easily shows (EXERCISE) that $\left[\tilde{H}_{\boldsymbol{\alpha}},\, F_{\pm\boldsymbol{\beta}}\right] = \pm(q-p)F_{\pm\boldsymbol{\beta}}$. And, provided one accepts the following formula: $|\boldsymbol{\alpha}+\boldsymbol{\beta}|/|\boldsymbol{\beta}| = \sqrt{q+1/p}$ (check with all possible angles allowed by eq. (3.30) below), then it is not hard to obtain: $\left[F_{\pm\boldsymbol{\alpha}},\, F_{\boldsymbol{\beta}}\right] = \pm(q+1)F_{\boldsymbol{\alpha}+\boldsymbol{\beta}}$. In this basis, all the structure constants are

real integers. This result, due I think to Chevalley (∼1955), means that any semisimple algebra, not just $\mathfrak{su}(2)$, has a real version.

Thus, if a semisimple algebra has dimension n and rank r, it contains $(n-r)/2$ distinct $\mathfrak{su}(2)$ subalgebras, each having as Cartan generator a different element of the Cartan subalgebra, plus two non-Cartan generators corresponding to the root that labels the $\mathfrak{su}(2)$ subalgebra.

Once we have put all the roots on a root diagram, we notice how symmetrical the diagram looks. This is because, of course, not all the $n-r$ roots can be linearly independent; only r of them can be.

Definition 3.21. A **positive** root is one whose first non-zero component is positive; otherwise, it is **negative**. The r (rank of the algebra) positive roots which cannot be obtained from a linear combination of other positive roots are called **simple**, or **independent**. The other positive roots can be obtained as linear combinations of the simple roots, with *positive* coefficients.

Roots may be obtained from other roots via a **Weyl reflection** with respect to a **Weyl hyperplane** in the root diagram. There is a Weyl hyperplane perpendicular to each root through the origin of the root space. For instance, the reflection of a root with respect to its own Weyl hyperplane gives $\boldsymbol{\alpha} \to -\boldsymbol{\alpha}$.

Roots are in fact extremely constrained. If, when we derived eq. (3.29), we had started instead with $[E_{\boldsymbol{\beta}},\, E_{\boldsymbol{\alpha}}] = C_{\boldsymbol{\beta\alpha}} E_{\boldsymbol{\alpha}+\boldsymbol{\beta}}$, and added/subtracted integer multiples of $\boldsymbol{\beta}$ to $\boldsymbol{\alpha}$, we would have found in the same way that $2\boldsymbol{\beta}\cdot\boldsymbol{\alpha}/|\boldsymbol{\beta}|^2 = -(p'-q')$. Multiplying the two expressions, we arrive at the important expression:

$$\frac{(\boldsymbol{\alpha}\cdot\boldsymbol{\beta})^2}{|\boldsymbol{\alpha}|^2|\boldsymbol{\beta}|^2} \;=\; \cos^2\theta_{\boldsymbol{\alpha\beta}} \;=\; \frac{1}{4}\,(p-q)(p'-q') \tag{3.30}$$

The relative length of the roots is seen to be constrained to $|\boldsymbol{\alpha}|/|\boldsymbol{\beta}| = \sqrt{(p'-q')/(p-q)}$. Also, if $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are simple roots, $\boldsymbol{\alpha}-\boldsymbol{\beta}$ cannot be a root; otherwise, either it or $\boldsymbol{\beta}-\boldsymbol{\alpha}$ must be positive, and a simple root could be constructed out of two different positive roots: eg., $\boldsymbol{\beta} = (\boldsymbol{\beta}-\boldsymbol{\alpha})+\boldsymbol{\alpha}$. This means that $C_{\boldsymbol{\alpha},-\boldsymbol{\beta}} = 0$. Comparing with $C_{\boldsymbol{\alpha},-\boldsymbol{\beta}+q\boldsymbol{\alpha}} = 0$, derived from eq. (3.27) applied to $C_{-\boldsymbol{\alpha},\boldsymbol{\beta}-q\boldsymbol{\alpha}} = 0$, we see that $q = 0$ for simple roots. Therefore, from the master formular (3.29), the angle between two *simple* roots satisfies $\cos\theta_{\boldsymbol{\alpha\beta}} \le 0$, so that $\pi/2 \le \theta_{\boldsymbol{\alpha\beta}} \le \pi$.

Since $(p-q)(p'-q')$ must be an integer, There are only five possible values allowed for $\cos^2\theta_{\boldsymbol{\alpha\beta}}$ in eq. (3.30), and this, for any two roots of any semisimple algebra: $0 \;\Rightarrow \theta_{\boldsymbol{\alpha\beta}} = \pm 90°$; $1/4 \;\Rightarrow\; \theta_{\boldsymbol{\alpha\beta}} = 60°, 120°$; $1/2 \Rightarrow \theta_{\boldsymbol{\alpha\beta}} = 45°, 135°$; $3/4 \Rightarrow \theta_{\boldsymbol{\alpha\beta}} = 30°, 150°$; and $1 \Rightarrow \theta_{\boldsymbol{\alpha\beta}} = 0°, 180°$.

Thanks to all these constraints, a systematic and exhaustive procedure exists to construct the root space for all four families of classical semisimple groups, and for the five so-called exceptional groups. With the subscript denoting the rank of the algebra, the four families are:

- A_{n-1} $(n > 1)$, corresponding to $SU(n)$, $SL(n,\mathbb{R})$, $SU(p,q)$, with $p+q = n$ (no relation to the p and q above!)
- B_n, corresponding to $SO(2n+1)$ and $SO(p,q)$, with $p+q = 2n+1$.
- C_n, corresponding to $Sp(n)$ and $Sp(p,q)$, with $p+q = 2n$.
- D_n, corresponding to $SO(2n)$ and $SO(p,q)$, with $p+q = n$.

$SU(2)$, $SL(2,\mathbb{R})$, both A_1, $SO(2)$ (D_1), and $SO(3)$ (B_1) , all have the same one-dim root space with the two roots ± 1. Only five two-dimensional root spaces (four classical and one exceptional) can satisfy all our constraints; but B_2 and C_2 are rotated from each other by $45°$, so are taken to be the same. And there are only four three-dimensional root spaces. Beyond three dimensions, root spaces can no longer be represented on root diagrams. Other methods, such as Dynkin diagrams, take over.

Finally, a few words about weight diagrams. One of the Cartan generators, say H_1, will always be the Cartan generator of a $\mathfrak{su}(2)$ (and $\mathfrak{so}(3)$ - see section 3.6.1 below) subalgebra. Then weight points are arranged on lines

parallel to the H_1 axis, with each line corresponding to an irreducible representation (multiplet) of $\mathfrak{su}(2)$ labelled with j, an integer multiple of $1/2$, and containing $2j+1$ weights. These weights can be generated by applying the lowering non-Cartan generator of $\mathfrak{su}(2)$ to the weights in each $\mathfrak{su}(2)$ multiplet, starting with the highest one, ie. by repeated addition of the r-dim root, $(-1, 0, \ldots, 0)$, to that highest weight. This root, as well as $(1, 0, \ldots, 0)$ (which moves you up from the lowest to the highest weight) , is always a root of the semisimple algebra. Needless to say, as you move parallel to the H_1 axis, all other eigenvalues in the weights remain the same. since you are moving perpendicular to all other H_i axes. One can use other roots to move from one line of weights to its neighbouring lines.

The number of weights for these different $\mathfrak{su}(2)$ multiplets must add up to the dimension of the multiplet of the semisimple algebra. The $\mathfrak{su}(2)$ multiplets must fit snugly inside this multiplet. For instance, take the 10-dimensional representation (decuplet) of $\mathfrak{su}(3)$. The rank of this algebra is 2, and thus the weights are 2-component "vectors". There are four eigenvalues for the other Cartan generator, H_2, in this representation, and the weights are organised on an inverted-triangle lattice made of one horizontal $\mathfrak{su}(2)$ quadruplet, triplet, doublet and singlet, in the direction of decreasing H_2 eigenvalues.

3.5.1 Casimir invariant operators

Each irreducible representation of a Lie algebra can be labelled with the eigenvalues of some function of the basis generators of the algebra.

> **Definition 3.22.** A **Casimir invariant operator** C for a representation of a Lie algebra is an operator that commutes with *all* the generators of the representation.

When the representation is irreducible, C has to be a multiple of the identity by Schur's lemma. All elements of an invariant subspace of the carrier space of the representation will be eigenvectors of C *with the same eigenvalue*. When the algebra is semisimple, a theorem by Chevalley guarantees the existence of a set of Casimir operators as polynomials in the generators, whose eigenvalues may be used to characterise the irreducible representations of the algebra. More precisely, each invariant subspace of the carrier space has a set of basis vectors, each labeled by an eigenvalue of *each* Casimir operator. The number of Casimir operators is the rank of the algebra.

In other words, if $f(\mathbf{x})$ is in an invariant subspace of the Hilbert space which is the carrier space of the algebra, for each Casimir operator C_i, $C_i f(\mathbf{x}) = g(\mathbf{x})$ is a set of equations invariant under the group transformations, in the sense that under the action of the group, any $f(\mathbf{x})$ in the invariant subspace is sent into another function $g(\mathbf{x})$ in that same invariant subspace.

Because a Cartan metric can always be defined for a semisimple algebra, I claim that $g^{\mu\nu} X_\mu X_\nu$ is a Casimir operator, where the X_μ are basis generators of such an algebra. Indeed:

$$\begin{aligned}
\left[g^{\mu\nu}\, X_\mu\, X_\nu,\, X_\rho\right] &= g^{\mu\nu}\left(X_\mu\,[X_\nu,\, X_\rho] \,+\, [X_\mu,\, X_\rho]\, X_\nu\right) \\
&= g^{\mu\nu}\, C_{\mu\rho}{}^{\lambda}\,(X_\nu\, X_\lambda \,+\, X_\lambda\, X_\nu) \\
&= g^{\mu\nu}\, g^{\alpha\lambda}\, f_{\mu\rho\alpha}\,(X_\nu\, X_\lambda \,+\, X_\lambda\, X_\nu) \\
&= 0
\end{aligned}$$

because $g^{\mu\nu}\, g^{\alpha\lambda} f_{\mu\rho\alpha}$ is antisymmetric in, and the term in round brackets is symmetric in, ν and λ. For instance, we found in example 3.14 that for $\mathfrak{so}(3)$, the Cartan metric is $g_{\mu\nu} = \delta^{\mu}{}_{\nu}$. We immediately get:

$$C \;=\; X^\mu\, X_\mu \;=\; J_x^2 \,+\, J_y^2 \,+\, J_z^2 \;=\; J^2$$

where $\mathbf{J}$ is the angular momentum operator of quantum mechanics. Because $\mathfrak{so}(3)$ has rank 1, this is the only Casimir invariant in this case. Then the eigenvalues of J^2 each label an irreducible representation of $\mathfrak{so}(3)$.

3.6 More on finding irreducible representations

3.6.1 Irreducible representations of so(3)

To find the irreducible representations of $\mathfrak{so}(3)$, our approach is to find the eigenvalues λ of the Casimir operator J^2. $\mathfrak{so}(3)$ has one Cartan generator, J_z, whose eigenvalues m are the weights of the representations. The other generators, J_x and J_y, do not commute with J_z and therefore are non-Cartan. We choose a Hermitian representation for the generators.

First, go to the complex extension of $\mathfrak{so}(3)$ via the change of basis: $J_\pm = (J_x \pm \mathrm{i}J_y)/\sqrt{2}$, $J_0 = J_z$. J_x and J_y are Hermitian, but $J_\pm$ are not: we have $J_+^\dagger = J_-$, and $J_-^\dagger = J_+$. J_0 and $J_\pm$ commute with J^2, by definition of a Casimir operator. The commutation relations between J_x, J_y and J_z give $[J_0\,, J_\pm] = \pm J_\pm$, and we identify from definition 3.20 the roots of $J_\pm$ as ± 1, with $+1$ the simple root. Eq. (3.25) then leads directly to: $[J_+,\ J_-] = J_0$.

Also, eq. (3.26) becomes for $\mathfrak{so}(3)$:

$$J_0(J_\pm f) \;=\; (m \,\pm\, 1)\,(J_\pm f)$$

and, since J^2 must commute with $J_\pm$:

$$J^2(J_\pm f) \;=\; J_\pm(J^2 f) \;=\; \lambda\,(J_\pm f)$$

True to their nature as ladder operators, J_+ raises, and J_- lowers, the weights m by 1, but they cannot transform f into an eigenfunction of J^2 with a different eigenvalue. Consequently, *all* the eigenfunctions of J_0 reachable with the ladder operators in a given invariant subspace are also eigenfunctions of J^2 with the *same* eigenvalue λ.

Another expression will come in handy, relating J^2 to the generators. Use the definition of $J_\pm$ to write:

$$J_\pm\, J_\mp \;=\; \frac{1}{2}\,\left(\,J_x^2 \,+\, J_y^2 \,\mp\, \mathrm{i}\,[J_x, J_y]\right) \;=\; \frac{1}{2}\,\left(J_x^2 \,+\, J_y^2 \,\pm\, J_0\right) \;=\; \frac{1}{2}\,\left(J^2 \,-\, J_0^2 \,\pm\, J_0\right)$$

so that:

$$J^2 \;=\; 2J_\pm\, J_\mp \,+\, J_0^2 \,\mp\, J_0 \tag{3.31}$$

For a given value of λ, we expect that there should exist a highest weight, $m_{\max} \equiv j$, as well as a lowest weight, $m_{\min} \equiv j'$, since the corresponding irreducible representation must be finite-dimensional.

Now act with J_+ on the highest-weight eigenfunction of J_0, f_j. Then $J_+ f_j = 0$ and, from the identity (3.31), we find:

$$J^2 f_j \;=\; j^2 f_j \,+\, j\, f_j \;=\; j(j+1)\, f_j \;=\; \lambda\, f_j$$

Similarly, act with J^2 on the lowest-weight eigenfunction of J_0, $f_{j'}$, keeping in mind that $f_{j'}$ is also an eigenfunction of J^2 with the *same* eigenvalue as f_j:

$$J^2 f_{j'} \;=\; (j')^2 f_{j'} \,-\, j' f_{j'} \;=\; j'(j'-1)\, f_{j'} \;=\; \lambda\, f_{j'}$$

Comparing yields $\lambda \;=\; j(j+1) \;=\; j'(j'-1)$, and thus $j' = -j$. It follows that the weights m go from $-j$ to j in N integer steps, ie, $j = -j + N$, so $j = N/2$.

We conclude that:

- The eigenvalues of the Casimir operator J^2 are $j(j+1)$, where j is a positive integer or a half-integer.
- For a given value of j, the weights m can take $2j+1$ values, from $-j$ to j. Therefore, odd-dimensional irreducible representations correspond to integer j and even-dimensional ones to half-integer j.

With the help of eq. (3.31), we can now exhibit the full action of J_- on an eigenstate f_{jm} of J^2 and J_0. Let $J_- f_{jm} = c_-\, f_{j,m-1}$. Then, if the f_{jm} are normalised:

$$\begin{aligned}\langle f_{jm}|J_+ J_-\, f_{jm}\rangle \;&=\; \int f_{jm}^*\, J_+ J_-\, f_{jm}\,\mathrm{d}^3x \\ &=\; \langle J_-\, f_{jm}|J_-\, f_{jm}\rangle \\ &=\; c_-^* c_- \;=\; |c_-|^2\end{aligned}$$

But since $2J_\pm\, J_\mp = J^2 - J_0^2 \pm J_0$, we also have that:

$$\langle f_{jm}|J_+J_-\, f_{jm}\rangle \;=\; \frac{1}{2}\,\langle f_{jm}|(J^2 \,-\, J_0^2 \,+\, J_0)\, f_{jm}\rangle \;=\; \frac{1}{2}\,\big(j(j+1) \,-\, m^2 \,+\, m\big)$$

Comparing yields c_- up to an unimportant phase factor which we put equal to 1. We find the coefficient in $J_+ f_{jm} = c_+\, f_{j,m+1}$ in a strictly analogous way. The results for both ladder operators are:

$$J_\pm f_{jm} \;=\; \frac{1}{\sqrt{2}}\,\sqrt{j(j+1) \,-\, m(m\,\pm 1)}\, f_{j,m\pm1} \tag{3.32}$$

Each value of j labels an invariant $2j+1$-dimensional subspace of the carrier space of $\mathfrak{so}(3)$ of which the $2j+1$ eigenfunctions f_{jm} form a basis.

The entries of the three representation matrices $\mathcal{D}^j(J_0) = \langle f_{jm'}|J_0\, f_{jm}\rangle$, $\mathcal{D}^j(J_\pm) = \langle f_{jm'}|J_\pm\, f_{jm}\rangle$ are:

$$\mathcal{D}^j_{m'm}(J_0) \;=\; m\,\delta_{m'm} \qquad\qquad \mathcal{D}^j_{m'm}(J_\pm) \;=\; \frac{\delta_{m',m\pm1}}{\sqrt{2}}\,\sqrt{(j\mp m)(j\pm m+1)} \qquad |m| \,\le\, j \tag{3.33}$$

This form for the coefficients is often quoted, but the equivalent form in eq. (3.32) is often easier to use since only the second factor in the root changes. The representation matrices for $J_x = (J_+ + J_-)/\sqrt{2}$, $J_y = (J_+ - J_-)/(\mathrm{i}\sqrt{2})$ and $J_z = J_0$ are easily recovered if needed. Keeping in mind that the rows and columns are labelled by the *values* of m from j to $-j$, we have for, say, $j = 1$:

$$\mathcal{D}^1(J_0) \;=\; \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix} \qquad \mathcal{D}^1(J_+) \;=\; \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \qquad \mathcal{D}^1(J_-) \;=\; \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

3.6.2 Representations of su(2), SU(2), and SO(3)

The $j = 1/2$, defining irreducible representation of $\mathfrak{su}(2)$ can be written in terms of the three matrices $\mathbf{s}_0$ and $\mathbf{s}_\pm$:

$$\mathbf{s}_0 \;=\; \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{pmatrix} \qquad \mathbf{s}_+ \;=\; \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \qquad \mathbf{s}_- \;=\; \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$$

Transform to the basis $\mathbf{s}_x = (\mathbf{s}_+ + \mathbf{s}_-)/2$, $\mathbf{s}_y = (\mathbf{s}_+ - \mathbf{s}_-)/2\mathrm{i}$, $\mathbf{s}_z = \mathbf{s}_0$. Then the structure constants for this representation are identical to those of the $j = 1/2$ irreducible representation of $\mathfrak{so}(3)$. Indeed,

$$[s_i,\, s_j] \;=\; \mathrm{i}\,\epsilon_{ij}{}^k\, s_k$$

The $\mathbf{s}$ matrices are Hermitian; they are also traceless, as expected for $\mathfrak{su}(2)$. As before, a finite $SU(2)$ transformation can be reconstructed with the exponential map, corresponding to a rotation parametrised by $\boldsymbol{\theta} = \theta\hat{\mathbf{n}}$: $S(\boldsymbol{\theta}) = \mathrm{e}^{\mathrm{i}\theta\hat{\mathbf{n}}\cdot\mathbf{s}}$, where the direction of $\hat{\mathbf{n}}$ is the axis of rotation.

But the isomorphism between $\mathfrak{su}(2)$ and $\mathfrak{so}(3)$ does not translate into an isomorphism between $SU(2)$ and $SO(3)$! Indeed, whereas a $SO(3)$ rotation by 2π is identical to the identity, a $SU(2)$ rotation by 2π is equivalent to *minus* the identity, because of the factor $1/2$ lurking in the $\mathbf{s}$ matrices.

There *is* a $2 \to 1$ homomorphism that maps $SU(2)$ to $SO(3)$: $\pm S(\boldsymbol{\theta}) \to R(\boldsymbol{\theta})$, and because of this $SU(2)$ *can* be represented by $SO(3)$ matrices. But the map is not uniquely invertible, and therefore $SU(2)$ matrices are not *stricto sensu* representations of $SO(3)$. Only the ones that correspond to integer $SU(2)$ j are; those with half-integer $SU(2)$ j are called **spinor** representations, and we say that integer and half-integer representations of $SU(2)$ together form **projective** representations R_g of $SO(3)$, in the sense that $R_{\mathrm{g}_1}.R_{\mathrm{g}_2} = \alpha_{\mathrm{g}_1,\mathrm{g}_2} R_{\mathrm{g}_1\mathrm{g}_2}$, with $\alpha \in \mathbb{C}$

Wigner matrices $\mathcal{D}^j_\theta = \mathrm{e}^{\mathrm{i}\theta\hat{\mathbf{n}}^j\cdot\mathbf{s}_j}$ (with $\mathbf{s}_j$ the triplet of $\mathfrak{su}(2)$ basis generators of the irreducible representation labelled by j), is the name given to the irreducible representations of $SU(2)$, and the matrix elements are called Wigner functions. They can be rather complicated, except when $\hat{\mathbf{n}} = \hat{\mathbf{z}}$ and $s_z = s_0$ is diagonal, in which case $(\mathcal{D}^j_\theta)^{m'}{}_m = \mathrm{e}^{\mathrm{i}m\theta}\delta^{m'}{}_m$ ($|m| \le j$). Fortunately, they are tabulated in many places for small values of j and are easily calculated by computer.

3.6.3 Tensor product representations

Definition 3.23. Let $f_{j_1 m_1}$ and $f_{j_2 m_2}$ be two basis functions associated, respectively, with irreducible representations $\mathcal{D}^{j_1}_{\rm g}$ and $\mathcal{D}^{j_2}_{\rm g}$ of ${\rm g} \in SU(2)$ or $SO(3)$, such that:

$$S_{\rm g}\, f_{j_1 m_1} \;=\; f_{j_1 m'_1}\, (\mathcal{D}^{j_1}_{\rm g})^{m'_1}{}_{m_1}, \qquad\qquad S_{\rm g}\, f_{j_2 m_2} \;=\; f_{j_2 m'_2}\, (\mathcal{D}^{j_2}_{\rm g})^{m'_2}{}_{m_2}$$

where $S_{\rm g}$ is the transformation associated with g. Then we form the **tensor product representation** $\mathcal{D}^{j_1}_{\rm g} \otimes \mathcal{D}^{j_2}_{\rm g}$:

$$S_{\rm g}\, f_{j_1 m_1}\, f_{j_2 m_2} \;=\; f_{j_1 m'_1}\, f_{j_2 m'_2}\, (\mathcal{D}^{j_1})^{m'_1}{}_{m_1}\, (\mathcal{D}^{j_2})^{m'_2}{}_{m_2} \tag{3.34}$$

In Dirac notation, the product of the basis functions would read: $|j_1\, m_1,\, j_2\, m_2\rangle \;=\; |j_1\, m_1\rangle\, |\, j_2\, m_2\rangle$.

Such a product is needed when a system responds to transformations in more than one way, either because of the coupling of two separate systems (eg. particles) or because two distinct dynamical variables of one system get coupled. A common transformation on the whole system is to be written as a direct product of transformations on each of its parts *in its own subspace*.

Linearise eq. (3.34) using the generic expansion $\mathcal{D} = \mathsf{I} + a^i \mathbf{X}_i$, where $\mathbf{X}$ stands for the representation $\mathcal{D}(X)$ of generators of $SU(2)$ or $SO(3)$. We find that the generators of the composite representation are the sums of the generators of the distinct terms in the tensor product, so that:

$$\mathbf{X}^{(1\otimes 2)}(f_{j_1 m_1}\, f_{j_2 m_2}) \;=\; (X^{(1)}\, f_{j_1 m_1})\, f_{j_2 m_2} \;+\; f_{j_1 m_1}\, (X^{(2)}\, f_{j_2 m_2}) \tag{3.35}$$

that is: $\mathbf{X}^{(1\otimes 2)} = X^{(1)} \otimes \mathsf{I} + \mathsf{I} \otimes X^{(2)}$ or, more sloppily, $\mathbf{X} = \mathbf{X}^{(1)} + \mathbf{X}^{(2)}$. When the generators have diagonal representations, as happens with J_0 ($SO(3)$) or s_0 ($SU(2)$), we find, eg.:

$$J_0(f_{j_1 m_1}\, f_{j_2 m_2}) \;=\; (m_1 \,+\, m_2)\, f_{j_1 m_1}\, f_{j_2 m_2}$$

Note that $[X^{(1)},\, X^{(2)}] = 0$, because they act on distinct subspaces.

As before, we expect the product representation to be reducible, ie. there should exist linear combinations ϕ_{jm} (or $|j\, m\rangle$) of the product basis functions $f_{j_1 m_1}\, f_{j_1 m_2}$ which transform among themselves. In other words, we are looking for invariant subspaces of the Hilbert product space. Those linear combinations take the form of the invertible transformation:

$$\phi_{jm} \;=\; \sum_{m_1, m_2} \big(j_1 m_1,\, j_2, m_2\, |jm\big)\, f_{j_1 m_1}\, f_{j_2 m_2} \tag{3.36}$$

where $m = m_1 + m_2$, and $|j_1 - j_2| \le j \le j_1 + j_2$. The *real* coefficients $\big(j_1 m_1,\, j_2, m_2 | jm\big)$ are known as **Clebsch-Gordan** or **Wigner** coefficients. They are unique up to a phase convention and can be calculated or looked up in tables.

One easy way to obtain the ϕ_{jm} in terms of the $f_{j_1 m_1}\, f_{j_2 m_2}$ is to start with the highest weight component, $m = j_1 + j_2$, of the highest j irreducible representation: $j = j_1 + j_2$. Of course, $\phi_{j_1+j_2, j_1+j_2} = f_{j_1 j_1}\, f_{j_2 j_2}$. Next, apply J_- on the left and on the right, using eq. (3.35), until the lowest weight component of the j irreducible representation, $\phi_{j,-j}$, is reached. Now obtain the linear combination for the highest weight of the j - 1 representation, $\phi_{j-1,j-1}$, by demanding that it be orthogonal to $\phi_{j,j-1}$, and repeat with J_-. Continue until all values of j allowed by $|j_1 - j_2| \le j \le j_1 + j_2$ have been reached.

3.6.4 Irreducible (Spherical) tensors

Suppose that a set of functions f_{jm} in the carrier space of $SU(2)$ or $SO(3)$ transforms under a group element parametrised by $\boldsymbol{\theta} = \theta \mathbf{n}$ as: $R_{\boldsymbol{\theta}}\, f_{jm} \;=\; f_{jm'} (\mathcal{D}^j_{\boldsymbol{\theta}})^{m'}{}_m$. Then the set $\{f_{jm}\}$ form a basis for an irreducible representation of $SU(2)$ or $SO(3)$ labelled by j.

Definition 3.24. The components T_{jm} of an **irreducible spherical tensor** $\mathbf{T}_j$ of rank j are operators that transform as:

$$R_{\boldsymbol{\theta}}\, T_{jm}\, R_{\boldsymbol{\theta}}^{-1} \;=\; T_{jm'}\, (\mathcal{D}^{j}_{\boldsymbol{\theta}})^{m'}{}_{m} \tag{3.37}$$

If we linearise this equation, we obtain (EXERCISE) an alternative definition of spherical tensors in terms of generators $\mathbf{J}^{(j)}$ of an irreducible representations of the algebra:

$$[J^{(j)},\, T_{jm}] \;=\; T_{jm'}\, (J^{(j)})^{m'}{}_{m} \qquad\qquad (\text{no summation on } j) \tag{3.38}$$

where j is the label of the irreducible representation.

For SU(2) or SO(3):

$$[J_0^{(j)},\, T_{jm}] \;=\; m\, T_{jm}, \qquad\qquad [J_\pm^{(j)},\, T_{jm}] \;=\; \sqrt{(j \mp m)(j \pm m + 1)}\, T_{j,m\pm 1} \tag{3.39}$$

As a direct consequence of these commutation relations, the matrix element of T_{jm}, $\langle j_2 m_2 | T_{jm} | j_1 m_1 \rangle$, vanishes unless $m_2 = m_1 + m$ and $|j_1 - j| \le j_2 \le j_1 + j$. These are the famous vector addition rules as applied to spherical tensors.

3.6.5 The Wigner-Eckart theorem

The Wigner-Eckart theorem says that if $\mathbf{T}_j$ is a spherical tensor under $SU(2)$, then its matrix elements, written in bra-ket notation, $\langle j_2 m_2 | T_{jm} | j_1 m_1 \rangle$, can be factored as:

$$\langle j_2 m_2 |\, T_{jm}\, | j_1 m_1 \rangle \;=\; \frac{(j_1 m_1,\, jm | j_2 m_2)}{\sqrt{2j_2 + 1}}\, \langle j_2 ||\, \mathbf{T}_j\, || j_1 \rangle \tag{3.40}$$

where $\langle j_2 || \mathbf{T}_j || j_1 \rangle$ is called the **reduced** matrix element and *does not depend on* m, m_1 *or* m_2. So the dependence of the matrix element on these numbers is carried entirely by the Clebsch-Gordan coefficient!

The Wigner-Eckart theorem applies to unitary representations of Lie groups, not only to those of $SU(2)$. The Clebsch-Gordan coefficients and the labelling with eigenvalues of Casimir operators will be appropriate to the Lie group.

As a result, ratios of matrix elements for a given j but different values of m are just ratios of Clebsch-Gordan coefficients.

Example 3.18. Suppose that $\mathbf{T}$ transforms as a scalar under some Lie group. Then the relevant representation matrix of the group is just the identity matrix. If the Lie group is $SU(2)$, $j = m = 0$, and the vector-addition rules collapse the Wigner-Eckart theorem to:

$$\langle j_2 m_2 |\, \mathbf{T}\, | j_1 m_1 \rangle \;=\; \frac{\langle j_2 ||\, \mathbf{T}\, || j_1 \rangle}{\sqrt{2j_2 + 1}}\, \delta^{j_1}{}_{j_2}\, \delta^{m_1}{}_{m_2}$$

From this we see that matrix elements of scalar operators between weights associated with different irreducible representations of a group vanish.

Essentially, the importance of the Wigner-Eckart theorem resides in its separating symmetry-related ("geometrical") aspects of matrix elements from other ("dynamical") aspects which may be difficult to calculate and which the theorem shoves into the reduced matrix element.

3.6.6 Decomposing product representations

The problem of decomposing representations of a semisimple group into their irreducible representations can often be treated in a fairly intuitive way. Consider $SO(3)$ again, and its 3-dim carrier space of functions $f(\mathbf{x})$ and $g(\mathbf{y})$ (eg. in quantum mechanics, the wave-functions of two particles), each of which transforms in some known way under 3-dim rotations. We can form tensor products, $f(\mathbf{x}) \otimes g(\mathbf{y})$, of such functions, whose transformation properties are derived from those of the functions.

For instance, if our functions were 3-dim vectors, we would have a 9-dim product representation (ie. one with nine weights, or basis vectors for its carrier space) , with components T^{ij}, which under rotations R would tranform as:

$$T'^{ij} \;=\; R^i{}_k \, R^j{}_l \, T^{kl} \tag{3.41}$$

We know that the T^{ij} can be decomposed into a symmetric and an antisymmetric part, each of which transforms *separately* under rotations, in the sense that the 6-dim symmetric part rotates into a symmetric object, and the 3-dim antisymmetric part into an antisymmetric one. Thus, we have easily found invariant subspaces. But we can go even further. The trace of T^{ij}, $T^i{}_i$, is invariant under rotations, forming a 1-dim invariant subspace which should be separated out from the symmetric part.

Note that the trace is obtained by contracting T^{ij} with the metric of the *carrier space*, with components g_{ij}, which here is just the identity matrix invariant under rotations. Similarly, the antisymmetric part can be obtained with the Levi-Civita symbol which is also invariant under rotation. Thus, we can write:

$$T^{ij} \;=\; \frac{1}{2}\left(T^{ij} + T^{ji}\right) + \frac{1}{2}\,\epsilon^{ijk}\,\epsilon_{klm}\,T^{lm} \;=\; \frac{1}{2}\left(T^{ij} + T^{ji} - \frac{2}{3}\,g^{ij}\,T^k{}_k\right) + \frac{1}{2}\left(T^{ij} - T^{ji}\right) + \frac{1}{3}\,g^{ij}\,T^k{}_k \tag{3.42}$$

The numerical coefficient of the trace term has been chosen so as to make the symmetric term traceless.

But we can also think of eq. (3.41) as a $3\otimes 3$ exterior direct product a rotation with itself, so a 9×9 matrix, with each row labelled by a *pair* $\{ij\}$ and each column labelled by a *pair* $\{kl\}$, acting on a 9×1 matrix with entries T^{kl} labelled by the pairs $\{kl\}$. The direct-product matrix is a representation of $SO(3)$. Indeed, under a rotation R_1 followed by R_2, $T^{ij} \;\longrightarrow\; (R_2\,R_1)^i{}_m (R_2\,R_1)^j{}_n T^{mn}$, where now the 9×9 matrix is formed from the *matrix* product $R_2 R_1$. The representation is reducible, that is, it can be transformed via an angle-independent similarity matrix to a block-diagonal matrix with a symmetric traceless 6×6 block (which acts only on the symmetric traceless part of $\mathbf{T}$) , an amtisymmetric 3×3 block acting only on the antisymmetric part of $\mathbf{T}$, and a 1 acting only on the trace of $\mathbf{T}$.

We obtain the following decomposition into irreducible representations: $\mathbf{9} \;=\; \mathbf{5} \oplus \mathbf{3} \oplus \mathbf{1}$

As expected, the total dimensions on the left and right match. The result is also consistent with what we would find by decomnposing a $j_1 \otimes j_2 = 1 \otimes 1$ $SO(3)$ product representation with the method of section 3.6.3 to obtain a direct sum of three irreducible representations labelled by $j = 2$ (of dimension 5), $j = 1$ (of dimension 3), and $j = 0$ (of dimension 1).

4 MODULE IV — Elementary Theory of Analytic Functions

4.1 Complex Numbers

We construct an extended algebra over the real numbers with the following structure in a representation: $z = x + \epsilon\, y$, where x (the **real** part) and y (the **imaginary** part) are real numbers, ϵ is an independent, non-real quantity endowed with a rule that characterises the algebra. One can think of it as a device that allows us to extend arithmetical operations from the real numbers with essentially no change.

Addition and multiplication are straightforward:

$$\begin{aligned} z_1 + z_2 &= x_1 + \epsilon\, y_1 + x_2 + \epsilon\, y_2 = x_1 + x_2 + \epsilon\,(y_1 + y_2) \\ z_1\, z_2 &= (x_1 + \epsilon\, y_1)(x_2 + \epsilon\, y_2) = x_1 x_2 + \epsilon^2\, y_1 y_2 + \epsilon\,(x_1 y_2 + y_1 x_2) \end{aligned}$$

If the extended structure is to be at least a ring of numbers, it should close under multiplication. This is achieved if ϵ^2 itself is of the form $a + \epsilon b$, with $a,\ b \in \mathbb{R}$. It is usual to choose $b = 0$ and $a = \pm 1$.

To any extended number z one associates its **conjugate** $z^* \equiv \overline{z} = x - \epsilon\, y$. $z = z^*$ if and only if z is real. Since ϵ^2 is real, so is $|z|^2 = z\, z^* = x^2 - \epsilon^2 y^2$. Also, to perform a division, one *must* first make the denominator real: $z_1/z_2 = z_1 z_2^*/|z_2|^2$. This of course requires that $|z_2| \neq 0$. Multiplication is both commutative and associative.

One extended algebra, that of the **hyperbolic-complex** numbers, is obtained by choosing $\epsilon^2 = +1$. If we write a spacetime event at position x and time t as $ct + \epsilon x$, hyperbolic-complex numbers provide[†] a natural language for special relativity. Another algebra, that of the **complex** numbers, is characterised by $\epsilon \equiv \mathrm{i}$, where $\mathrm{i}^2 = -1$. It is the one we will consider here.

When multiplying or dividing complex numbers it is usually more convenient to use the **polar** representation:

$$z = r\, \mathrm{e}^{\mathrm{i}\theta}$$

To find $r \equiv |z|$ (the **modulus**, or **absolute value**) and θ (the **phase**, or **argument**), we invoke the beautiful Euler relation:

$$\mathrm{e}^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\, \sin\theta$$

which can be proved by defining $w = \cos\theta + \mathrm{i}\, \sin\theta$, then taking its derivative $\mathrm{d}_\theta w = \mathrm{i}(\cos\theta + \mathrm{i}\, \sin\theta) = \mathrm{i}w$. The solution of this differential equation is $w = \mathrm{e}^{\mathrm{i}\theta}$. Then we write:

$$z = r\, \cos\theta + \mathrm{i}\, r\, \sin\theta = x + \mathrm{i}y$$

Comparing yields the transformation between the Cartesian and polar representations of complex numbers:

$$\begin{aligned} x &= r\, \cos\theta & \qquad r &= \sqrt{x^2 + y^2} \\ y &= r\, \sin\theta & \qquad \theta &= \tan^{-1}(y/x) \end{aligned}$$

We recognise these expressions as the transformations between Cartesian and polar coordinates for vectors in two-dimensional Euclidean space. Thus, complex numbers can be mapped to points in a **complex plane** with coordinates (x, y) or (r, θ). The last expression containing the tangent cannot distinguish between angles in the first and third quadrant and between the second and fourth quadrant of the complex plane. Which angle to choose must be decided from the signs of x and y.

Two-dim rotations and complex numbers are representations of one another.

[†]For an accessible introduction, see http://dx.doi.org/10.1063/1.530244 (free access from a UofT computer) and references cited therein.

4.2 Holomorphic Functions of a Complex Variable (BF 6.1)

4.2.1 Differentiability of complex functions

Introduce a one-to-one mapping from the z complex plane to another complex plane $w = u + \mathrm{i}v$:

$$f(z) \;=\; u(x,y) \;+\; \mathrm{i}\, v(x,y)$$

where u and v are real functions of x and y. f maps regions in the z plane to regions in the w plane.

We are of course interested in functions that are continuous and differentiable. But $\mathrm{d}_z f(z)$ is not quite like the gradient of a two-dimensional vector which is function of x and y. With the gradient, we ask how the vector changes under $x + \mathrm{d}x$ or $y + \mathrm{d}y$, keeping the other variable constant; in the case of $f(z) = f(r\mathrm{e}^{\mathrm{i}\theta})$, we think that the derivative $\mathrm{d}_z f(z)$ is well-defined only if it is invariant under $\theta + \mathrm{d}\theta$. In other words, it should not matter in which *direction* we vary z: the derivative should be isotropic. More formally:

$$\mathrm{d}_z f \;=\; \lim_{\Delta z \to 0} \frac{f(z+\Delta z) \;-\; f(z)}{\Delta z} \tag{4.1}$$

where the limit exists, and therefore must be independent of the phase of Δz.

It is difficult to overemphasise how strong a constraint this is on functions of a complex variable! It is much more restrictive than for functions of a real variable. All the properties and applications we are going to review follow from it.

4.2.2 Cauchy-Riemann Relations

We need a more workable criterion than eq. (4.1) for complex differentiability. If we write $\mathrm{d}_z f(z) = p(x,y) + \mathrm{i}q(x,y)$, then:

$$\delta f \;=\; \mathrm{d}_z f(z)\, \delta z \;=\; \big[p(x,y) \;+\; \mathrm{i}\, q(x,y)\big]\, (\delta x \;+\; \mathrm{i}\, \delta y) \;=\; \delta u \;+\; \mathrm{i}\, \delta v$$

Equating real and imaginary parts gives:

$$\begin{aligned} \delta u \;&=\; p\delta x \;-\; q\delta y \;=\; \partial_x u\, \delta x \;+\; \partial_y u\, \delta y \\ \delta v \;&=\; q\delta x \;+\; p\delta y \;=\; \partial_x v\, \delta x \;+\; \partial_y v\, \delta y \end{aligned}$$

By omparing these equations, we find that complex differentiability yields a system of coupled differential equations for u and v, the **Cauchy-Riemann equations**:

$$\partial_x u(x,y) \;=\; \partial_y v(x,y), \qquad\qquad \partial_y u(x,y) \;=\; -\,\partial_x v(x,y) \tag{4.2}$$

We see that the real and imaginary parts of a complex-differentiable function are not independent; in fact, if one is known, the other can be calculated up to a constant. For instance, given $u(x,y)$, $\mathrm{d}_y v(x,y)$ is calculable from one Cauchy-Riemann condition, and if it can be integrated with respect to y, v is known up to a function of x. The other Cauchy-Riemann equation provides the derivative of the unknown function which can then be retrieved up to a constant.

> **Definition 4.1.** Let the function $f(z)$ satisfy the Cauchy-Riemann equations (let it be complex-differentiable) at a point z_0; *in addition*, let it be complex-differentiable over a finite, open neighbourhood of z_0, (or, alternatively, continuous over that neighbourhood). Then $f(z)$ is said to be **holomorphic**, or *regular*, at z_0.
>
> A function f is said to be analytic at z_0 if there exists a series $\sum a_n (z - z_0)^n$ ($n \geq 0$), and also a $r > 0$ such that the series converges to $f(z)$ for $|z - z_0| < r$. We see that an analytic function is holomorphic; the reverse will be shown to hold (but not for functions of a real variable!) later.
>
> Points at which a function is not holomorphic are called **singularities** (or singular points); if a function is nowhere singular in the complex plane, we say that it is an **entire function**.

By differentiating one Cauchy-Riemann relation with respect to x and the other with respect to y (and vice-versa), we show that both u and v must satisfy the Laplace equation: $\nabla^2 u = \nabla^2 v = 0$, and so are both harmonic. Because, as we shall see in a later chapter, boundary conditions uniquely determine the solutions of the Laplace equation, we expect functions holomorphic over a region in the complex plane to be determined by their values on the boundary of the region. This, as we will soon discover, provides the justification for the important process of analytic continuation of a function.

In general, a function $f = u(x, y)+\mathrm{i}v(x, y)$ might be expected to depend on both z and z^* since $x = (z+z^*)/2$ and $y = (z - z^*)/2\mathrm{i}$. If so, then:

$$\begin{aligned} \partial_z f &= \partial_x f\, \partial_z x + \partial_y f\, \partial_z y = \frac{1}{2}\big(\partial_x u + \partial_y v\big) + \frac{\mathrm{i}}{2}\big(\partial_x v - \partial_y u\big) \\ \partial_{z^*} f &= \partial_x f\, \partial_{z^*} x + \partial_y f\, \partial_{z^*} y = \frac{1}{2}\big(\partial_x u - \partial_y v\big) + \frac{\mathrm{i}}{2}\big(\partial_x v + \partial_y u\big) \end{aligned} \tag{4.3}$$

The Cauchy-Riemann conditions are seen to be equivalent to imposing $\partial_{z^*} f = 0$, so that f is a function of a complex variable, not just a complex function of two real variables. The message is clear: a holomorphic function of a complex variable cannot depend on the complex conjugate of that variable.

Holomorphic functions can always be constructed out of harmonic functions u of two real variables. When this is done, one should be able to write the functions $f = u + \mathrm{i}v$ in terms of z only, with no z^* dependence.

From eq. (4.3), it should also be clear that $\mathrm{d}_z f$ can be written in terms of derivatives of either u *or* v. For instance, $\partial_z f = \partial_x u - \mathrm{i}\partial_y u$, and the complex derivative of f can be calculated with the sole knowledge of the real part of f!

4.2.3 Single- and Multi-valued Functions (BF 6.2)

It is not hard to come up with holomorphic functions. We have already seen how they could be generated from harmonic functions of two real variables. Useful examples are the exponential function e^z and the trigonometric functions ($\sin z$, $\cos z$, $\tan z$), which share with their real cousins all the usual relations and properties. They are obtained by a process called **analytic continuation** from the real axis into the complex plane. Apart from $\tan z$, which is singular at $z = (n + 1/2)\pi$, they are entire functions.

Simply by replacing $z = x + \mathrm{i}y$, one readily obtains the $u + \mathrm{i}v$ forms:

$$\begin{aligned} \mathrm{e}^z &= \mathrm{e}^x\,(\cos y + \mathrm{i}\,\sin y) \\ \cos z &= \cos x\,\cosh y - \mathrm{i}\,\sin x\,\sinh y \\ \sin z &= \sin x\,\cosh y + \mathrm{i}\,\cos x\,\sinh y \end{aligned} \tag{4.4}$$

Unlike $\cos x$ and $\sin x$, however, neither $|\cos z|$ nor $|\sin z|$ is bounded:

$$|\cos z|^2 = \cos^2 x + \sinh^2 y, \qquad |\sin z|^2 = \sin^2 x + \sinh^2 y$$

Another, and perhaps the most important, example is the power function z^n, with $n \in \mathbb{Z}^+$. Indeed, most common functions can be written as power series, which means that they can be approximated by polynomials. The closed unit disk centered on the origin of the complex plane is defined by $|z| \le 1$. Then it should be clear that the function z^n maps the disk into itself.

Roots and logarithms require a bit more care. Let us look at roots first. Here we must use the polar representation of z, $r\mathrm{e}^{\mathrm{i}\theta}$, and we have:

$$z^{1/n} = r^{1/n}\,\mathrm{e}^{\mathrm{i}(\theta+2\pi k)/n} \qquad k = 0,\, 1, \ldots,\, n-1$$

While $r^{1/n}$ is unique, the whole function is multi-valued: if we start anywhere on the positive real axis ($\theta = 0$), every time we go around by 2π, we arrive at a different value; only after n loops do we recover the value $z(\theta = 0) = r^{1/n}$. This means that, starting from the singular point at the origin (where the root has no finite derivative), the function is not continuous, and therefore not analytic anywhere on the positive real axis!

Whenever there exists a point z_0 around which we must circle by an angle larger than 2π in order to come back to the same value $f(z_1)$ that we had at z_1 when we started, we say that z_0 is a **branch point**. We can recover single-valuedness of a function by using the **Riemann construction**: we imagine that along the line of singularities starting at the branch point, the complex plane defined by $0 \leq \theta \leq 2\pi$ is joined smoothly to *another* complex plane $2\pi \leq \theta \leq 4\pi$, and so on until the last plane $2(n-1)\pi \leq \theta \leq 2n\pi$ is joined smoothly to the first complex plane. The different planes are known as **Riemann sheets** on each of which a different **branch** of the function is defined, with the line on which the sheets are joined a **branch cut**.

More simply, we can think of the root as a bunch of n single-valued functions, each defined on one sheet $2(k-1)\pi \leq \theta \leq 2k\pi$, with $1 \leq k \leq n$.

Another well-known example of a multi-valued function is $\ln z$:

$$\ln z \;=\; \ln r \;+\; \mathrm{i}\,(\theta \;+\; 2\pi n)$$

where any integer value of n corresponds to the same point z. Thus, for a given z there are infinitely many values of the log, each separated from its neighbours by $\pm 2\pi\,\mathrm{i}$. This time, each Riemann sheet $2(k-1)\pi \leq \theta \leq 2k\pi$ may be joined on the real axis to make $\ln z$ holomorphic everywhere except at $z = 0$, where it goes to $-\infty$. Or we can think of each sheet as providing one branch which is non-holomorphic on the real axis. In the range $0 \leq \theta \leq 2\pi$, $\ln z$ is often called the **principal branch**, or **principal value**, and denoted by $\mathrm{Log}\, z$ in tables of integrals or mathematical handbooks.

Useful representations of inverse trigonometric functions in terms of logarithms can be obtained, for instance, to discover the principal branch of those functions:

$$z \;=\; \tan w \;=\; \frac{1}{\mathrm{i}}\,\frac{\mathrm{e}^{\mathrm{i}w} \;-\; \mathrm{e}^{-\mathrm{i}w}}{\mathrm{e}^{\mathrm{i}w} \;+\; \mathrm{e}^{-\mathrm{i}w}}$$

is easily inverted to yield:

$$w \;=\; \tan^{-1} z \;=\; \frac{\mathrm{i}}{2}\,\ln\left(\frac{1-\mathrm{i}z}{1+\mathrm{i}z}\right)$$

with principal branch $\mathrm{i}\big[\mathrm{Log}\,(1-\mathrm{i}z) - \mathrm{Log}\,(1+\mathrm{i}z)\big]/2$. Similar expressions can be found (EXERCISE) for $\cos^{-1} z$ and $\sin^{-1} z$.

4.2.4 Conformal mappings

When $f = u + \mathrm{i}v$ is analytic, the functions $u(x,y)$ and $v(x,y)$ define a transformation whose Jacobian is:

$$\frac{\partial(u,v)}{\partial(x,y)} \;=\; \partial_x u\,\partial_y v \;-\; \partial_y u\,\partial_x v \;=\; (\partial_x u)^2 \;+\; (\partial_y u)^2 \;=\; |\partial_x u \;-\; \mathrm{i}\,\partial_y u|^2 \;=\; |\mathrm{d}_z f|^2$$

because of the Cauchy-Riemann conditions and eq. (4.3). The transformation is one-to-one at points where the Jacobian, and therefore $\mathrm{d}_z f$, does not vanish. Points where this is not true are said to be **critical**.

Example 4.1. Consider the transformation $w = \ln z$, with $z = r\mathrm{e}^{\mathrm{i}\phi}$. Then $u = \ln r$ and $v = \phi$. Focus on the principal branch $0 \leq \phi \leq 2\pi$. Circles of radius r centered on the origin are mapped into vertical lines $u = \ln r$. Annular regions $r_1 \leq r \leq r_2$ are mapped into vertical strips $\ln r_1 \leq u \leq \ln r_2$ and height $0 \leq v \leq 2\pi$. Finally, rays centered on the origin $\phi = \phi_0$ are mapped into horizontal lines $v = \phi_0$.

Definition 4.2. Suppose two curves, C_1 and C_2, intersect at point z_0. A **conformal** transformation preserves the magnitude and sense of the angle between the two curves at the point w_0 where their images intersect.

At any point where it is analytic and where its derivative does not vanish, a function $f(z)$ is a conformal transformation. This is easily seen by showing that tangents to all curves at that point are rotated by the same amount. Indeed, consider the tangent to some curve at z_0 and its image w via f. If the curve is parametrised by t, $\mathrm{d}_t z$ and $\mathrm{d}_t w$ are the tangent vectors at z_0 and at its image w_0, respectively. Then:

$$\mathrm{d}_t w\big|_{w_0} \;=\; \mathrm{d}_z w\, \mathrm{d}_t z\big|_{z_0} \;=\; \mathrm{d}_z f\, \mathrm{d}_t z\big|_{z_0}$$

Writing $\mathrm{d}_z f = R\mathrm{e}^{\mathrm{i}\alpha}$, we conclude that $\mathrm{d}_t z\big|_{z_0}$ is rotated by an angle α to obtain $\mathrm{d}_t w\big|_{w_0}$. Since this applies to the tangent of any curve through z_0, the angle between the tangents of any two curves at z_0 is left unchanged by the transformation, and the latter is conformal.

If $f = u + \mathrm{i}v$ is holomorphic, one can show (exercise) by applying the Cauchy-Riemann conditions to the scalar product of the gradients of u and v that the curves of constant u are perpendicular to the curves of constant v.

One of the most useful conformal transformations is the so-called **fractional** (aka bilinear, homographic, Möbius) transformation:

$$f(z) \;=\; \frac{az \,+\, b}{cz \,+\, d} \qquad\qquad ad \,-\, bc \;\neq\; 0 \tag{4.5}$$

where the condition $ad - bc \neq 0$ arises from the need to keep $\mathrm{d}_z f$ non-zero. The transformation can be rewritten:

$$f(z) \;=\; \lambda \,+\, \frac{\mu}{z \,+\, \nu}$$

where λ, μ and ν are constants. Since $z + \nu$ represents a translation, $1/(z + \nu)$ an inversion, and multiplication by μ a dilation combined with a rotation, the bilinear transformation is a combination of all these.

Example 4.2. If z_0 lies in the upper half-plane, the bilinear transformation:

$$w \;=\; \mathrm{e}^{\mathrm{i}\phi_0}\, \frac{z \,-\, z_0}{z \,-\, \overline{z}_0}$$

maps the upper half-plane into the interior of the unit circle, $|w| \leq 1$. Indeed, $|w| = |z - z_0|/|z - \overline{z}_0|$ which is smaller than 1 except when z lies on the real axis, in which case $|w| = 1$.

Example 4.3. Under transformation (4.5), circles are mapped into circles. To see this, we note first that the general equation of a circle in the xy plane can be written: $A(x^2 + y^2) + Bx + Cy + D = 0$, with $A > 0$ and $D > 0$. Transforming to complex coordinates yields $\alpha z z^* + \beta z + \beta^* z^* + \gamma = 0$, where $\alpha = A$, $\beta = \frac{1}{2}(B - \mathrm{i}C)$, and $\gamma = D$. When $\alpha = A = 0$, the circle collapses to a straight line.

Now, under $w = 1/z$, this becomes $\gamma w w^* + \beta w^* + \beta^* w + \alpha = 0$, a circle. Under $w = az$ ($a \in \mathbb{C}$), it becomes $\alpha w w^* + \beta a^* w + \beta^* a w^* + \gamma a a^* = 0$, another circle. Translations also map circles to circles. Since a fractional transformation is a combination of inversion, dilation plus rotation, and translation, then it indeed maps circles to circles.

Conformal transformations are often used in some fields of physics or engineering to map a complicated set of objects to a simpler one. Finding the right one can be a black art, but extensive catalogues do exist, eg. S. G. Krantz, *Handbook of Complex Variables* (Birkhäuser, 1999), pp. 163–184.

4.3 Complex Integrals (BF 6.3)

4.3.1 The Cauchy-Goursat theorem

The integral of $f(z) = u + \mathrm{i}v$ over some path C between two points A and B in the complex plane is written:

$$\int_C f(z)\,\mathrm{d}z \;=\; \int_C (u\,\mathrm{d}x \,-\, v\,\mathrm{d}y) \,+\, \mathrm{i}\int_C (v\,\mathrm{d}x \,+\, u\,\mathrm{d}y) \;=\; \int_C \mathbf{F}\cdot\mathrm{d}\mathbf{l} \,+\, \mathrm{i}\int_C \mathbf{G}\cdot\mathrm{d}\mathbf{l} \tag{4.6}$$

where $\mathbf{F} = u\hat{\mathbf{x}} - v\hat{\mathbf{y}}$ and $\mathbf{G} = v\hat{\mathbf{x}} + u\hat{\mathbf{y}}$. The integral from point B to point A on the *same* path is the negative of the integral from A to B.

Now consider a closed path, also called a **contour**. If it encloses a simply-connected region S (no "holes" with points not belonging to S) in the complex plane, and if the derivatives are continuous everywhere in S, Stokes theorem requires that:

$$\oint_C f(z)\,\mathrm{d}z \;=\; \int_S (\boldsymbol{\nabla}\times\mathbf{F})\cdot\mathrm{d}\mathbf{S} \,+\, \mathrm{i}\int_S (\boldsymbol{\nabla}\times\mathbf{G})\cdot\mathrm{d}\mathbf{S}$$

where $\mathrm{d}\mathbf{S} = \hat{\mathbf{z}}\mathrm{d}S$. Evaluating the "$z$" component of the two curls gives $(\boldsymbol{\nabla}\times\mathbf{F})_z = -(\partial_y u + \partial_x v)$ and $(\boldsymbol{\nabla}\times\mathbf{G})_z = \partial_x u - \partial_y v$.

Now, if $f(z)$ is holomorphic everywhere in S and on its border C, the curls vanish because of the Cauchy-Riemann conditions, and the Cauchy-Goursat theorem[†] holds:

$$\oint_C f(z)\,\mathrm{d}z \;=\; 0 \qquad\qquad f(z) \text{ holomorphic within and on } C \tag{4.7}$$

The Cauchy-Goursat theorem is readily extended to multiply-connected regions. Indeed, suppose there is a hole in S. Consider a path that goes some distance along the exterior contour, then leaves it on a path that takes it to the boundary of the hole, which is then traversed in the direction *opposite* the exterior contour, coming back to the latter on the reverse interconnecting path, and then finishes the trip along the exterior contour back to the initial point. The net contribution to the integral of the two interconnecting paths vanishes. The complete contour encloses a simply connected region, so $\oint f(z)\,\mathrm{d}z \;=\; 0$. But this is equivalent to an integral over the exterior boundary plus an integral over the interior boundary, along a path such that if you walk on this boundary, S is always to your left.

The converse of the Cauchy-Goursat theorem, **Morera's theorem**, also holds: if $f(z)$ is continuous in a simply connected region and if $\oint f(z)\,\mathrm{d}z = 0$ around every simple closed curve in that region, then $f(z)$ is holomorphic all over the region.

4.3.2 A few consequences of the Cauchy-Goursat Theorem

If $f(z)$ is holomorphic within and on a (closed) contour C around a simply connected region, then we have the following:

- Obviously, the integral of $f(z)$ between any two points in that region is path-independent.
- **Fundamental Theorem of Calculus:**

$$\int_a^b f(z)\,\mathrm{d}z \;=\; F(b) \,-\, F(a) \qquad\qquad \mathrm{d}_z F(z) \;=\; f(z) \tag{4.8}$$

  where F is analytic.

[†]The best proof of the Cauchy-Goursat theorem relies only on $f(z)$ being analytic on S and its boundary C, but it is a lot fussier than the one we present, which strictly speaking applies to the Cauchy theorem (continuity of derivatives *assumed*).

- **Winding number**:

  On a circle C of radius r centered on z_0, we obtain by direct integration:

$$\oint_C \frac{1}{z - z_0}\,\mathrm{d}z \;=\; \int_0^{2\pi} \frac{1}{r\mathrm{e}^{\mathrm{i}\theta}}\,\mathrm{i}r\,\mathrm{e}^{\mathrm{i}\theta}\,\mathrm{d}\theta \;=\; 2\pi\,\mathrm{i}$$

  This result is totally independent of the radius of the circle, or indeed the shape of the contour, as well as of z_0 itself!

  Definition 4.3. We define the **winding number with respect to** z_0 as:

$$W_C(z_0) \;=\; \frac{1}{2\pi\,\mathrm{i}} \oint_C \frac{\mathrm{d}z}{z - z_0} \tag{4.9}$$

  where $W_C(z_0) \in \mathbb{Z}^+$ and C is *any* loop enclosing z_0. Since the circle can be deformed into a loop of any shape that encloses z_0, clearly the winding number counts the number of times any closed loop goes around z_0 in the counterclockwise direction.

- **Cauchy-Gauss Integral Formula:**

  Suppose that $f(z)$ is holomorphic on a closed region of the complex plane with border C. Consider a circle C_0, centered on a point z_0, with no point outside C. Then, from the Cauchy-Goursat theorrem:

$$\oint_C \frac{f(z)}{z - z_0}\,\mathrm{d}z \;=\; \oint_{C_0} \frac{f(z)}{z - z_0}\,\mathrm{d}z$$

  We can also write:

$$\oint_{C_0} \frac{f(z)}{z - z_0}\,\mathrm{d}z \;=\; f(z_0) \oint_{C_0} \frac{\mathrm{d}z}{z - z_0} + \oint_{C_0} \frac{f(z) - f(z_0)}{z - z_0}\,\mathrm{d}z \;=\; 2\pi\,\mathrm{i}\,f(z_0) + \oint_{C_0} \frac{f(z) - f(z_0)}{z - z_0}\,\mathrm{d}z$$

  The last term on the right vanishes. To see this, invoke **Darboux's inequality**:

$$\left|\int_C g(z)\,\mathrm{d}z\right| \;\leq\; \int_C \left|g(z)\,\mathrm{d}z\right| \;=\; \int_C \left|g(z)\right|\,|\mathrm{d}z| \;\leq ML \tag{4.10}$$

  where M is the maximum value of $g(z)$ on C and L is the length of C. The middle equality results from $|z_1 z_2| = |z_1||z_2|$. , working n the complex representation, Now call $\delta = |z - z_0|$ the radius of C_0. Because $f(z)$, being holomorphic, is continuous at z_0, we know that for any $\epsilon > 0$, no matter how small, we can choose δ small enough that $|f(z) - f(z_0)| < \epsilon$. From the Darboux inequality:

$$\left|\oint_{C_0} \frac{f(z) - f(z_0)}{z - z_0}\,\mathrm{d}z\right| \;\leq\; \frac{\epsilon}{\delta}\,(2\pi\,\delta) \;=\; 2\pi\,\epsilon$$

  Since ϵ can be arbitrarily small, we can make the absolute value of the integral as small as we wish.

  Therefore, for any z_0 inside a region, enclosed by a contour C, where f is holomorphic, the **Cauchy-Gauss Integral formula** holds:

$$f(z_0) \;=\; \frac{1}{2\pi\,\mathrm{i}} \oint_C \frac{f(z)}{z - z_0}\,\mathrm{d}z \tag{4.11}$$

  Thus, $f(z)$ on the (closed) boundary of a region in the complex plane determines its values *everywhere inside*, provided only that f is holomorphic over the whole region.

- Write the integration variable in the Cauchy-Gauss formula as ζ, with $z_0 = z$, and differentiate it with respect to z:

$$\mathrm{d}_z f(z) \;=\; \frac{1}{2\pi\,\mathrm{i}} \oint_C \frac{f(\zeta)}{(\zeta - z)^2}\,\mathrm{d}\zeta$$

where the integral is well-defined so long as z is not on the contour (see p. BF333 for a different proof). We can go on differentiating any number n of times. Unlike for functions of a real variable, $f(z)$ being holomorphic implies the existence of not only its first derivative, but of *all* its derivatives, and the following rather cute relation holds:

$$\mathrm{d}_z^n f(z)\Big|_{z_0} \;=\; \frac{n!}{2\pi\,\mathrm{i}} \oint_C \frac{f(\zeta)}{(\zeta - z_0)^{n+1}}\,\mathrm{d}\zeta \tag{4.12}$$

Example 4.4. Evaluate, over the circle $|z| = 3$:

$$\oint \frac{\mathrm{e}^{2z}}{(z+2)^4}\,\mathrm{d}z$$

This might seem quite hard until we spot the resemblance of the integrand with that in eq. (4.12) with $z_0 = -2$, and $n = 3$. Then:

$$\mathrm{d}_z^3 f(z)\Big|_{z=-2} \;=\; \frac{3!}{2\pi\,\mathrm{i}} \oint_C \frac{f(\zeta)}{(\zeta - z)^4}\,\mathrm{d}\zeta$$

Here, $f(\zeta) = \mathrm{e}^{2\zeta}$, and:

$$\oint_{|z|=3} \frac{\mathrm{e}^{2\zeta}}{(\zeta + 2)^4}\,\mathrm{d}\zeta \;=\; \frac{\pi\,\mathrm{i}}{3}\,\mathrm{d}_z^3 \mathrm{e}^{2z}\Big|_{z=-2} \;=\; \frac{8\pi\,\mathrm{i}}{3}\,\mathrm{e}^{-4}$$

Note again that it matters not a jot that the contour is a circle and what its radius is, so long as the point $z = -2$ lies *inside*: the result is exactly the same.

- **Liouville's Theorem:**

If $f(z)$ is an entire function whose modulus is bounded, ie., if $|f(z)| \leq M$, then $f(z)$ is a constant. To prove this, all we have to show is that $\mathrm{d}_z f(z) = 0$. We have just seen (eq. (4.12)) that:

$$\mathrm{d}_z f(z) \;=\; \frac{1}{2\pi\,\mathrm{i}} \oint_C \frac{f(\zeta)}{(\zeta - z)^2}\,\mathrm{d}\zeta$$

where we can choose C to be a circle of radius r centered at z. Then, evaluating the absolute value in the polar representation and then using the Darboux inequality:

$$|\mathrm{d}_z f(z)| \;=\; \frac{1}{2\pi} \left| \int_0^{2\pi} \frac{f(\zeta)}{r^2\,\mathrm{e}^{2\mathrm{i}\theta}}\,\mathrm{i}\,r\,\mathrm{e}^{\mathrm{i}\theta}\,\mathrm{d}\theta \right| \;\leq\; \frac{1}{2\pi\,r} \int_0^{2\pi} |f(\zeta)|\,\mathrm{d}\theta \;\leq \frac{M}{r}$$

Since f is entire, we can take r to be arbitrarily large, and is derivative indeed goes to 0, proving the theorem. This means that entire functions such as e^z and $\sin z$ are necessarily *not* bounded (as we have seen before in the case of the latter).

- **Fundamental Theorem of Algebra:**

If $P(z) = \sum a_n z^n$ is a polynomial of degree $N \geq 1$, then it has as at least one root, ie. there exists $z = z_1$ such that $P(z_1) = 0$.

To show this, it is sufficient to assume that $P(z) \neq 0$ everywhere, which leads to $1/|P(z)|$ being everywhere analytic (entire). Moreover, $1/P(z)$ is bounded, and thus, by Liouville's theorem, is a constant, which is impossible. Thus, there must exists at least one value z_1 of z such that $P(z) = (z - z_1)Q(z) = 0$, where $Q(z)$ is a poynomial of degree $N - 1$.

We can go on to argue in the same way that $Q(z)$ must also have at least one root, ie. a value z_2 where it vanishes. At the end of this process, we find that $P(z)$ must have exactly N roots.

- **Poisson's integral formulae**

Let $f(z)$ be holomorphic on and inside a circle C of radius R, and $z = r\,\mathrm{e}^{\mathrm{i}\theta}$ be any point inside the circle. Then, from the Cauchy-Gauss integral formula (4.11):

$$f(z) \;=\; \frac{1}{2\pi\,\mathrm{i}} \oint_C \frac{f(\zeta)}{\zeta \,-\, z}\,\mathrm{d}\zeta$$

Now the point $R^2/\overline{z}$ lies outside the circle, so that the function $f(\zeta)/(\zeta \,-\, R^2/\overline{z})$ is holomorphic *inside*, and its integral along the circle vanishes from the Cauchy theorem. Therefore, we can write:

$$f(z) \;=\; \frac{1}{2\pi\,\mathrm{i}} \left(\oint_C \frac{f(\zeta)}{\zeta \,-\, z}\,\mathrm{d}\zeta \,-\, \oint_C \frac{f(\zeta)}{\zeta \,-\, R^2/\overline{z}}\,\mathrm{d}\zeta \right) \;=\; \frac{1}{2\pi\,\mathrm{i}} \oint_C \frac{z \,-\, R^2/\overline{z}}{(\zeta - z)(\zeta - R^2/\overline{z})}\, f(\zeta)\,\mathrm{d}\zeta$$

Going to the polar representation $z = r\,\mathrm{e}^{\mathrm{i}\theta}$ and $\zeta = R\,\mathrm{e}^{\mathrm{i}\phi}$, we obtain:

$$\begin{aligned} f(r\,\mathrm{e}^{\mathrm{i}\theta}) \;&=\; \frac{1}{2\pi\,\mathrm{i}} \int_0^{2\pi} \frac{r\,\mathrm{e}^{\mathrm{i}\theta} - (R^2/r)\,\mathrm{e}^{\mathrm{i}\theta}}{(R\,\mathrm{e}^{\mathrm{i}\phi} - r\,\mathrm{e}^{\mathrm{i}\theta})(R\,\mathrm{e}^{\mathrm{i}\phi} - (R^2/r)\,\mathrm{e}^{\mathrm{i}\theta})}\, f(R\,\mathrm{e}^{\mathrm{i}\phi})\,\mathrm{i}\,R\,\mathrm{e}^{\mathrm{i}\phi}\,\mathrm{d}\phi \\ &=\; \frac{1}{2\pi} \int_0^{2\pi} \frac{(R^2 - r^2)}{(R\,\mathrm{e}^{\mathrm{i}\phi} - r\,\mathrm{e}^{\mathrm{i}\theta})(R\,\mathrm{e}^{-\mathrm{i}\phi} - r\,\mathrm{e}^{-\mathrm{i}\theta})}\, f(R\,\mathrm{e}^{\mathrm{i}\phi})\;\mathrm{d}\phi \\ &=\; \frac{1}{2\pi} \int_0^{2\pi} \frac{(R^2 - r^2)}{R^2 \,+\, r^2 \,-\, 2Rr\,\cos(\theta - \phi)}\, f(R\,\mathrm{e}^{\mathrm{i}\phi})\,\mathrm{d}\phi \end{aligned} \tag{4.13}$$

This is **Poisson's integral formula** on a circle.

There is a companion formula on the half-plane. Let $f(z)$ be holomorphic in the upper half-plane, and consider a semi-circle of radius R in the upper half-plane, centered on the origin, and with its base on the real axis. Let $z = a + \mathrm{i}b$ be a point inside the semi-circle. Then the Cauchy-Gauss formula and Cauchy's theorem give, respectively:

$$f(z) \;=\; \frac{1}{2\pi\,\mathrm{i}} \oint_C \frac{f(\zeta)}{\zeta \,-\, z}\,\mathrm{d}\zeta, \qquad\qquad 0 \;=\; \frac{1}{2\pi\,\mathrm{i}} \oint_C \frac{f(\zeta)}{\zeta \,-\, \overline{z}}\,\mathrm{d}\zeta$$

Subtracting and combining yields:

$$\begin{aligned} f(z) \;&=\; \frac{1}{2\pi\,\mathrm{i}} \oint_C \frac{z \,-\, \overline{z}}{(\zeta - z)(\zeta - \overline{z})}\, f(\zeta)\,\mathrm{d}\zeta \\ &=\; \frac{1}{\pi} \int_{-R}^{R} \frac{b}{(x - a)^2 \,+\, b^2}\, f(x)\,\mathrm{d}x \,+\, \frac{1}{\pi} \int_{\substack{|\zeta| = R \\ y > 0}} \frac{b}{(\zeta - z)(\zeta - \overline{z})}\, f(\zeta)\,\mathrm{d}\zeta \end{aligned}$$

This time, however, we impose the condition $|f(\zeta)/\zeta| \to 0$ as $R \to \infty$. Then the second integral on the semi-circle vanishes as $R \to \infty$, and we are left with another Poisson integral formula:

$$f(z) \;=\; \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{b}{(x - a)^2 \,+\, b^2}\, f(x)\,\mathrm{d}x \tag{4.14}$$

What can we do with these Poisson formulae? Well, you can be the life of the party by easily proving that:

$$\int_0^{2\pi} \frac{R^2 \,-\, r^2}{R^2 \,+\, r^2 \,-\, 2rR\,\cos(\theta - \phi)}\,\mathrm{d}\phi \;=\; 2\pi$$

4.4 Power-Series Expansions of Analytic Functions — Laurent and Taylor series (BF 6.7)

Power-series expansions provide an important application of the complex differentiability of functions of a complex variables.

Let $f(z)$ be holomorphic between and on circles C_1 and C_2 centered on z_0 in the complex plane. Then, everywhere in that region, $f(z)$ can be expanded in a *unique* series, the **Laurent series**:

$$f(z) \;=\; \sum_{n=-\infty}^{\infty} A_n\,(z\,-\,z_0)^n \qquad\qquad A_n \;=\; \frac{1}{2\pi\,\mathrm{i}}\oint_C \frac{f(z')}{(z'\,-\,z_0)^{n+1}}\,\mathrm{d}z' \qquad (n\in\mathbb{Z}) \tag{4.15}$$

where C is any contour that encloses z_0 in the annular region, and we take C_1 to be the outer circle. The series converges everywhere inside the annular region between C_1 and C_2. Outside this region, the series generally does not converge, and a new one must be found that converges.

To prove this theorem, consider a contour that traverses C_1 counterclockwise, but leaves C_1 at some point to go to C_2 and traverse it clockwise, coming back to C_1 along the path connecting the two circles. Integrating $f(z')/(z'-z)$ over this contour, it is clear that the interconnecting path makes no net contribution to the integral. And since f is holomorphic everywhere on and inside the contour, the value of f anywhere inside is given by the Cauchy-Gauss iItegral formula (4.11) which, applied to this contour, is:

$$\begin{aligned} f(z) \;&=\; \frac{1}{2\pi\,\mathrm{i}}\oint_{C_1}\frac{f(z')}{z'-z}\,\mathrm{d}z' \,-\, \frac{1}{2\pi\,\mathrm{i}}\oint_{C_2}\frac{f(z')}{z'-z}\,\mathrm{d}z' \\ &=\; \frac{1}{2\pi\,\mathrm{i}}\oint_{C_1}\frac{f(z')}{\big[(z'-z_0)\,-\,(z-z_0)\big]}\,\mathrm{d}z' \,+\, \frac{1}{2\pi\,\mathrm{i}}\oint_{C_2}\frac{f(z')}{\big[(z-z_0)\,-\,(z'-z_0)\big]}\,\mathrm{d}z' \end{aligned}$$

where the minus sign between the terms in the first line is due to C_2 being traversed clockwise.

Now we establish a useful identity:

$$\frac{1}{z_1-z_2} \;=\; \frac{1}{z_1} + \frac{z_2}{z_1}\,\frac{1}{z_1-z_2} \;=\; \sum_{n=0}^{N}\frac{z_2^n}{z_1^{n+1}} + \left(\frac{z_2}{z_1}\right)^{N+1}\frac{1}{z_1-z_2}$$

where the last equality is obtained by iterating the first equality (corresponding to $N=0$) N times. With this identity, $f(z)$ becomes:

$$\begin{aligned} f(z) \;=\; &\frac{1}{2\pi\,\mathrm{i}}\sum_{n=0}^{N}\oint_{C_1}\frac{(z-z_0)^n}{(z'-z_0)^{n+1}}\,f(z')\,\mathrm{d}z' \,+\, \frac{1}{2\pi\,\mathrm{i}}\sum_{m=0}^{N}\oint_{C_2}\frac{(z'-z_0)^m}{(z-z_0)^{m+1}}\,f(z')\,\mathrm{d}z' \,+ \\ &\frac{1}{2\pi\,\mathrm{i}}\oint_{C_1}\left(\frac{z-z_0}{z'-z_0}\right)^{N+1}\frac{f(z')}{z'-z}\,\mathrm{d}z' \,+\, \frac{1}{2\pi\,\mathrm{i}}\oint_{C_2}\left(\frac{z'-z_0}{z-z_0}\right)^{N+1}\frac{f(z')}{z-z'}\,\mathrm{d}z' \end{aligned}$$

Define $M_{1,2} = |f(z)|_{\max}$ on $C_{1,2}$, $d_{1,2} = |z-z'|_{\min}$ with $z'\in C_{1,2}$, and $l_{1,2} = |z-z_0|_{\max,\min}$, We note that $l_1 < r_1 = |z'-z_0|$ and $l_2 > r_2 = |z'-z_0|$, where $r_{1,2}$ is the radius of $C_{1,2}$. Then the absolute value of the last two terms is bounded by:

$$\frac{M_1\,r_1}{d_1}\left(\frac{l_1}{r_1}\right)^{N+1} + \frac{M_2\,r_2}{d_2}\left(\frac{r_2}{l_2}\right)^{N+1} \underset{N\to\infty}{\to} 0$$

Now, from the Cauchy-Goursat theorem, each ntegral in the first and second term is equal to an integral over an arbitrary contour C that encloses z_0 between C_1 and C_2. All that is left is to redefine the index m in the second integral as $m=-n-1$ and rewrite it as a sum over negative n. After letting $N\to\infty$, we merge the two sums and thus prove eq. (4.15). We can even deform C_1 and C_2 to other contours, so long as we do not cross any singular point on the way, ie., so long as $f(z)$ remains analytic within and on the deformed contours.

Definition 4.4. If $A_n \neq 0$ for $n > -N$ ($N > 0$), clearly $f(z)$ has a singularity at $z = z_0$. If $A_n = 0 \; \forall \, n < N$, then we say that $f(z)$ has a **pole of order** N at z_0. If $N \to \infty$, ie., if the series of negative powers is infinite, and if it is **isolated**—in the sense that it is not a branch point—this singularity is said to be **essential**. $f(z)$ is called **meromorphic** in regions where it has no *essential* singularity.

Because of their uniform convergence in the annular region, one can manipulate Laurent series as if they were *finite* power series, ie., polynomials: they can be added, multiplied, and even divided to represent the addition, the multiplication, or the ratio of the functions which they represent.

Now suppose further that $f(z)$ is holomorphic everywhere inside C_1. Then $A_n = 0 \; \forall \, n \leq -1$ so as to keep the Laurent series of $f(z)$ holomorphic at the origin. When only terms with $n \geq 0$ are present, we use eq. (4.12) to obtain directly **Taylor's theorem** for holomorphic functions:

$$f(z) \; = \; \sum_{n=0}^{\infty} \frac{1}{n!} \, \mathrm{d}_z^n f(z) \Big|_{z_0} (z \, - \, z_0)^n \tag{4.16}$$

valid anywhere inside a circle centered on z_0 where $f(z)$ is holomorphic. It also establishes that holomorphic functions are analytic in the sense of Definition 4.1, justifying the assertion that the two properties are equivalent.

Beyond its well-known usefulness, the Taylor series is also an accurate representation of $f(z)$ within a radius of convergence which extends all the way to its closest singularity. But since the function and all its derivatives are known at any point inside the circle of convergence, we can choose another point inside the circle around which to Taylor-expand, extending the domain in which we can calculate $f(z)$ further, until another singularity is encountered. This so-called **analytic continuation** can be repeated any number of times.

We conclude that knowledge of $f(z)$ in some region of the complex plane is sufficient to determine it *uniquely* everywhere it is analytic! This explains why we are able to start from the definitions and relations applying to many functions on the real axis, such as e^x or the trig functions, and extend them with impunity to the whole complex plane, or at least to any region in which the same function, with $x \to z$, is analytic.

Finally, if $f(z) = \sum a_n (z - z_0)^n$, where the series converges uniformly between C_1 and C_2, then $a_n = A_n$, that is, the series is a Laurent series (EXERCISE: multiply by $(z - z_0)^{-(k+1)}$, and integrate over an arbitrary contour between C_1 and C_2 that encloses z_0).

Example 4.5. The function $1/(1-z)$ can be written as the geometric series $S_1(z) = \sum_0^\infty z^n$, provided that $|z| < 1$. Similarly, if $|z + 1/4| < 5/4$:

$$S_2(z) \; = \; \frac{4}{5} \sum_{n=0}^{\infty} \left[\frac{4}{5} (z \, + \, \frac{1}{4}) \right]^n \; = \; \frac{4/5}{1 \, - \, \frac{4}{5} (z \, + \, 1/4)} \; = \; \frac{1}{1 \, - \, z}$$

The region in which S_2 converges is larger than the circle of convergence of S_1, and both represent the same function in the overlapping region. S_2 may be seen as an analytic continuation of S_1 beyond $|z| < 1$. Both S_1 and S_2 are Laurent series in their respective region of validity.

Example 4.6. The fact that the $2\pi\mathrm{i}\, A_{-1}$ coefficient of a Laurent series is equal to $\oint f(z)\mathrm{d}z$ can sometimes be used to evaluate the contour integral if A_{-1} can be found in some other way. Take $f(z) = 1/[z^2(z-3)]$ which has singularities at $z = 0$ and at $z = 3$, but is analytic in the open annular region $0 < |z| < 3$. Now write $f(z) = -1/[3z^2(1 - z/3)]$, and focus on the factor:

$$\frac{1}{1 - z/3} \; = \; \sum_{n=0}^{\infty} (z/3)^n$$

The right-hand side can be thought of either as a geometric series or as a Taylor expansion around $z = 0$ (since there is no singularity there). Then we reconstruct $f(z)$ and find that it can be written as the Laurent series:

$$f(z) \; = \; -\frac{1}{3z^2} \, - \, \frac{1}{9z} \, - \, \frac{1}{27} \, - \, \frac{z}{81} \, - \, \text{positive powers of } z \qquad\qquad |z| < 3$$

Therefore, $A_{-1} = -1/9$, and we arrive at:

$$\oint_C \frac{1}{z^2\,(z-3)}\,\mathrm{d}z \;=\; -\,\mathrm{i}\,\frac{2\pi}{9}$$

for *any* closed path that encloses $z = 0$ in the region $0 < |z| < 3$.

For the region $|z| > 3$, this Laurent series is not well defined because its positive-power terms make it diverge as $|z| \to \infty$. But we can write *for that region only*:

$$f(z) \;=\; \frac{1}{z^3}\,\frac{1}{1\,-\,3/z} \;=\; \frac{1}{z^3}\sum_{n=0}^{\infty}(3/z)^n$$

Example 4.7. As a highly non-trivial example of a Laurent series, consider:

$$f(z) \;=\; \frac{1}{\sqrt{z^2-1}} \;=\; \frac{1}{\sqrt{r_1\,r_2}}\,\mathrm{e}^{-\mathrm{i}\theta_1/2}\,\mathrm{e}^{-\mathrm{i}\theta_2/2} \qquad\qquad r_1 = |z-1|,\; r_2 = |z+1|$$

If we start at some z such that $|z-1| > 1$, for instance at $\theta_1 = 0$ and $\theta_2 = 0$ (positive real axis), it is clear in this polar parametrisation that when we come back to the same point after traversing a circle that encloses *both* singularities, both angles go from 0 to 2π, and we come back to the same value of $f(z)$, so that the function is single-valued and analytic on the real axis for $|z| > 1$. If, on the other hand, we start on the real axis between $z = 1$ and $z = -1$, and go full circle round *one only* of the two singular points, only *one angle* goes from 0 to 2π, and we come back to the negative of the value of $f(z)$ that we started from. Thus, there is a branch cut running between $z = 1$ and $z = -1$, and $f(z)$ is not analytic there.

Now $f(z)$ is analytic everywhere in the region (centered on the origin) $|z| > 1$, and this is where we look for a Laurent series. If we take the inner boundary C_2 of the region of existence of the Laurent series to be of radius $|z| = 1 + \epsilon$ around the origin, where ϵ can be as small as we wish but not zero, and if we further choose the outer boundary to have arbitrarily large radius, then $A_n = 0\;\forall\, n \geq 0$.

To find the A_n for $n < 0$, we choose a contour C made of circles of arbitrarily small (but non-zero) radius around $z = 1$ and $z = -1$, connected by a straight line hugging the real axis in the upper half-plane and another one very close to the real axis in the lower half-plane. Although this might seem to contradict the conditions $|z| > 1$ for the existence of a Laurent series, it does not: we use the path-independence of contour integration valid anywhere in a region of analyticity to say that the result is the same as if we had actually integrated on a contour with $|z| > 1$. The series itself, of course, is valid only for $|z| > 1$.

The contributions from the two circles can be seen to vanish. Indeed, for the circle C' around $z = 1$, we write:

$$\oint_{C'} \frac{(z'^2-1)^{-1/2}}{(z')^{n+1}}\,\mathrm{d}z' \;=\; \oint_{C'} \frac{\sqrt{z'-1}}{\sqrt{z'+1}\,(z')^{n+1}}\,\frac{1}{z'-1}\,\mathrm{d}z'$$

The integrand is of the form $f(z')/(z'-1)$, where $f(z')$ is analytic on and within C'. Therefore, from the Cauchy-Gauss integral formula (4.11), the integral is proportional to $f(z'=1) = 0$. Similar considerations apply to the circle around $z = -1$, and we are left with the contributions from the two straight lines above and below the real axis, arbitrarily close to it. Because of the discontinuity across the cut, these do not cancel one another. We write $\sqrt{x^2-1} = \pm\mathrm{e}^{\mathrm{i}\pi/2}\sqrt{1-x^2}$, where the + sign applies to the path from 1 to -1, in the upper half-plane, and the $-$ sign to the path from -1 to 1, in the lower half-plane. So long as we keep the sign distinction, we can then integrate on the actual x axis, and there comes for $A_n \neq 0$:

$$A_n \;=\; \frac{1}{2\pi\,\mathrm{i}}\left[-\int_{-1}^{1}\frac{x^{-(n+1)}\,\mathrm{e}^{-\mathrm{i}\pi/2}}{\sqrt{1-x^2}}\,\mathrm{d}x \;+\; \int_{1}^{-1}\frac{x^{-(n+1)}\,\mathrm{e}^{-\mathrm{i}\pi/2}}{\sqrt{1-x^2}}\,\mathrm{d}x\right] \;=\; \frac{1}{\pi}\int_{-1}^{1}\frac{x^{-(n+1)}}{\sqrt{1-x^2}}\,\mathrm{d}x \qquad (n\,\text{odd})$$

The integral can be evaluated by ***Maple/Mathematica*** or tables, and we finally get, with $\nu = -(n+1)/2$, the Laurent series:

$$\frac{1}{\sqrt{z^2-1}} \;=\; \sum_{\nu=0}^{\infty} \frac{(2\nu)!}{2^{2\nu}\,(\nu!)^2}\,\frac{1}{z^{2\nu+1}} \;=\; \frac{1}{z} + \frac{1}{2}\frac{1}{z^3} + \frac{3}{8}\frac{1}{z^5} + \ldots \qquad\qquad |z| > 1$$

Although this series has an infinite number of negative powers, the singularities at $z = \pm 1$ are not essential because, being branch points at the ends of the cut between $z = 1$ and $z = -1$, they are not isolated .

4.5 Cauchy Principal Value of an Integral (BF 6.5)

Consider the following integral over the real axis, with $-R < x_0 < R$:

$$\lim_{\delta\to 0}\left[\int_{-R}^{x_0-\delta} \frac{f(x)}{x-x_0}\,\mathrm{d}x \;+\; \int_{x_0+\delta}^{R} \frac{f(x)}{x-x_0}\,\mathrm{d}x\right] \;\equiv\; \mathcal{P}\int_{-R}^{R} \frac{f(x)}{x-x_0}\,\mathrm{d}x$$

where the symbol $\mathcal{P}$ means the **Cauchy principal value** of $f(x)/(x-x_0)$. Note that δ must be the same in both integrals. If we can evaluate this, we can make sense of an integral over the real axis whose integrand has a singularity at some point.

The principal value may always be written as:

$$\mathcal{P}\int_{-R}^{R} \frac{f(x)}{x-x_0}\,\mathrm{d}x \;=\; f(x_0)\,\mathcal{P}\int_{-R}^{R} \frac{1}{x-x_0}\,\mathrm{d}x \;+\; \mathcal{P}\int_{-R}^{R} \frac{f(x)\;-\;f(x_0)}{x-x_0}\,\mathrm{d}x$$

Now, if $-R < x_0 < R$,

$$\begin{aligned}\mathcal{P}\int_{-R}^{R} \frac{1}{x-x_0}\,\mathrm{d}x \;&=\; \lim_{\delta\to 0}\left[\int_{-R}^{x_0-\delta} \frac{1}{x-x_0}\,\mathrm{d}x \;+\; \int_{x_0+\delta}^{R} \frac{1}{x-x_0}\,\mathrm{d}x\right] \\ &=\; \lim_{\delta\to 0} \ln\left(\frac{R-x_0}{R+x_0}\right) \;=\; \ln\left(\frac{R-x_0}{R+x_0}\right)\end{aligned}$$

This Cauchy principal value vanishes when the limits of integration are taken to infinity ($R \to \infty$).

Thus, we obtain the useful relation, valid when $R > |x_0|$:

$$\mathcal{P}\int_{-R}^{R} \frac{f(x)}{x-x_0}\,\mathrm{d}x \;=\; f(x_0)\,\ln\left(\frac{R-x_0}{R+x_0}\right) \;+\; \mathcal{P}\int_{-R}^{R} \frac{f(x)\;-\;f(x_0)}{x-x_0}\,\mathrm{d}x \tag{4.17}$$

The integrand of the integral on the right is not singular at $x = x_0$ when $f(x)$ is differentiable at x_0; then the $\mathcal{P}$ symbol can be dropped.

Example 4.8. To illustrate this, let $f(x) = x/(x^2+1)$, whose derivative exists everywhere. The integrand on the right-hand side of eq. (4.17) becomes:

$$\frac{f(x)\;-\;f(x_0)}{x-x_0} \;=\; \left(\frac{x}{x^2+1} \;-\; \frac{x_0}{x_0^2+1}\right)\frac{1}{x-x_0} \;=\; \frac{1\;-\;x_0x}{(x^2+1)(x_0^2+1)}$$

The integrand is manifestly not singular at x_0, and we can drop the principal value symbol. Then:

$$\mathcal{P}\int_{-R}^{R} \frac{x}{(x^2+1)(x-x_0)}\,\mathrm{d}x \;=\; \frac{1}{x_0^2+1}\left[x_0\,\ln\left(\frac{R-x_0}{R+x_0}\right) \;+\; 2\,\tan^{-1}R\right] \qquad (R > |x_0|)$$

The restriction to $-R < x_0 < R$ becomes trivial if $R \to \infty$:

$$\mathcal{P}\int_{-\infty}^{\infty} \frac{x}{(x^2+1)(x-x_0)}\,\mathrm{d}x \;=\; \frac{\pi}{x_0^2+1}$$

4.6 Hilbert Transforms (BF 6.5)

I have already mentioned soon after introducing holomorphic functions how, given the real part of some holomorphic function $f(z)$, its imaginary part could be found up to a constant, and vice-versa. Let $|f(z)| \to 0$ as $|z| \to \infty$ in *either* the upper *or* lower half-plane, but not both (since then f would be entire and bounded, thus constant, ie. zero, by Liouville's theorem). We will show that with this condition a stronger result can be obtained for that same complex-valued function evaluated on the *real* axis: the real and imaginary parts of $f(x)$ are uniquely related via integrals over the *real* axis! Unlike when the Cauchy-Riemann conditions are used, we do not need detailed knowledge of $f(z)$ away fro the real axis, just that it is analytic in, say, the upper half-plane, and that it vanishes at infinity in that region.

First, we would like to integrate the function $f(z)/(z - x_0)$ over a closed counter-clockwise path (contour) consisting of a half-circle of infinite radius in the upper half-plane. centered on the origin, with its base along the real axis. $f(z)$ is analytic on and within the contour, but $f(z)/(z - x_0)$ has a singularity at x_0 on the real axis, so *it* is not analytic on the contour.

A strategy to make sense of the integral is to "avoid" x_0 along a semi-circle C_δ of radius δ in the upper half-plane, centered on x_0, whose diameter extends from $x_0 - \delta$ to $x_0 + \delta$. Since now $f(z)/(z - x_0)$ is analytic everywhere within and on the closed path we have chosen, the contour integral vanishes by the Cauchy-Goursat theorem.

The contribution to the contour integral from the upper arc of radius $R \to \infty$ is easily evaluated:

$$\lim_{R\to\infty} \int_{\substack{|z|=R \\ y>0}} \frac{f(z)}{z - x_0}\,\mathrm{d}z \;=\; \lim_{R\to\infty} \mathrm{i}\int_0^\pi \frac{f(z)}{\mathrm{e}^{\mathrm{i}\theta} - x_0/R}\,\mathrm{e}^{\mathrm{i}\theta}\,\mathrm{d}\theta$$

Now, $\left|\mathrm{e}^{\mathrm{i}\theta} - x_0/R\right| \geq \left|1 - x_0/R\right|$. Therefore, so long as $|f(z)| \to 0$ as $R \to \infty$, we have:

$$\left|\int_{R\to\infty} \frac{f(z)}{z - x_0}\,\mathrm{d}z\right| \;\leq\; \lim_{R\to\infty} \frac{1}{|1 - x_0/R|}\int_0^\pi |f(z)|\,\mathrm{d}\theta \;\longrightarrow\; 0$$

Another part of our closed path is the small semi-circular arc C_δ of radius δ, traversed *clockwise*:

$$\int_{-C_\delta} \frac{f(z)}{z - x_0}\mathrm{d}z \;\equiv\; f(x_0)\int_{-C_\delta} \frac{\mathrm{d}z}{z - x_0} + \int_{-C_\delta} \frac{f(z) - f(x_0)}{z - x_0}\mathrm{d}z \;=\; -\,\mathrm{i}\pi\, f(x_0) - \int_{C_\delta} \frac{f(z) - f(x_0)}{z - x_0}\mathrm{d}z$$

where $|z - x_0| = \delta$. The last integral vanishes as $\delta \to 0$. To see this, note that:

$$\lim_{\delta\to 0}\left|\int_{C_\delta} \frac{f(z) - f(x_0)}{z - x_0}\,\mathrm{d}z\right| \;\leq\; \lim_{\delta\to 0}\int_0^\pi |f(z) - f(x_0)|\,\mathrm{d}\theta$$

But since f is continuous (it is analytic), $|f(z) - f(x_0)|$ can always be made smaller than any positive number ϵ, no matter how small that number may be, by taking δ small enough, which the limit implies anyway. Then the integral, which is smaller than $\pi\epsilon$ from eq. (4.10) (Darboux), can be made to vanish by choosing ϵ vanishingly small. There is a much niftier way of calculating this integral over C_δ, but that is for a little later.

The remaining contribution is the principal value integral of $f(x)/(x - x_0)$. Thus, summing the non-zero contributions to the contour integral to zero, we have succeeded in evaluating the Cauchy principal value:

$$\mathcal{P}\int_{-\infty}^{\infty} \frac{f(x)}{x - x_0}\,\mathrm{d}x \;=\; \mathrm{i}\pi\, f(x_0) \tag{4.18}$$

(EXERCISE: Can this result be used to evaluate the Cauchy principal value of the function in example 4.8?) By taking the real and imaginary parts of this equation, we conclude that if $f(x)$ is a complex-valued function of a real variable such that $f(z)$ is analytic in the upper half-plane (or lower half-plane, by a similar argument) with $|f(z)| \to 0$ as $|z| \to \infty$, its real and imaginary parts are related by the **Hilbert transforms**:

$$\Re[f(x_0)] \;=\; \frac{1}{\pi}\,\mathcal{P}\int_{-\infty}^{\infty} \frac{\Im[f(x)]}{x - x_0}\,\mathrm{d}x, \qquad \Im[f(x_0)] \;=\; -\,\frac{1}{\pi}\,\mathcal{P}\int_{-\infty}^{\infty} \frac{\Re[f(x)]}{x - x_0}\,\mathrm{d}x \tag{4.19}$$

Also, with $R \to \infty$, the relation (4.17) we derived in the last section becomes:

$$\mathcal{P} \int_{-\infty}^{\infty} \frac{f(x)}{x - x_0}\,\mathrm{d}x \;=\; \mathcal{P} \int_{-\infty}^{\infty} \frac{f(x) \;-\; f(x_0)}{x - x_0}\,\mathrm{d}x$$

Now, because $f(z)$ is analytic on the real axis, it is differentiable at x_0, and we can drop the $\mathcal{P}$ symbol in the right-hand side. Then the Hilbert transforms become:

$$\Re[f(x_0] \;=\; \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\Im[f(x)] - \Im[f(x_0)]}{x - x_0}\,\mathrm{d}x, \qquad \Im[f(x_0)] \;=\; -\frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\Re[f(x)] - \Re[f(x_0)]}{x - x_0}\,\mathrm{d}x \tag{4.20}$$

4.7 Dispersion relations (BF 6.6)

4.7.1 Non-locality in time

The behaviour in time, $R(t)$, of a system in response to a stimulus $I(t)$ cannot be instantaneous. We shall now elucidate important consequences of this fact.

Suppressing any spatial dependence, the Fourier decompositions of $R(t)$ and $I(t)$ over the frequency domain are (*note the choice of sign in the exponentials!*):

$$R(t) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} r(\omega)\,\mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{d}\omega \qquad\qquad r(\omega) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} R(t')\,\mathrm{e}^{\mathrm{i}\omega t'}\,\mathrm{d}t'$$

$$I(t) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} i(\omega)\,\mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{d}\omega \qquad\qquad i(\omega) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} I(t')\,\mathrm{e}^{\mathrm{i}\omega t'}\,\mathrm{d}t'$$

We are interested in situations where $r(\omega) = g(\omega)\,i(\omega)$, with $g(\omega)$ bounded $\forall\,\omega$. Then:

$$\begin{aligned} R(t) \;&=\; \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{d}\omega\, g(\omega)\,\mathrm{e}^{-\mathrm{i}\omega t} \int_{-\infty}^{\infty} \mathrm{d}t'\,\mathrm{e}^{\mathrm{i}\omega t'}\,I(t') \;=\; \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} \mathrm{d}\omega\, g(\omega)\,\mathrm{e}^{-\mathrm{i}\omega(t-t')} \right] I(t')\,\mathrm{d}t' \\ &=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} G(\tau)\,I(t-\tau)\,\mathrm{d}\tau \qquad (\tau = t - t') \end{aligned}$$

where $G(\tau)$ is the **kernel**, or **response function**:

$$G(\tau) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(\omega)\,\mathrm{e}^{-\mathrm{i}\omega\tau}\,\mathrm{d}\omega \qquad\qquad g(\omega) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} G(\tau)\,\mathrm{e}^{\mathrm{i}\omega\tau}\,\mathrm{d}\tau \tag{4.21}$$

We see that $R(t)$ depends on the *whole history* of $I(t)$, and that the Fourier formalism provides non-locality in time in a natural way, as the convolution integral (see section 5.5 of Module V):

$$R(t) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} G(t-t')\,I(t')\,\mathrm{d}t'$$

But this is a little strange: not only does $R(t)$ depend on $I(t' < t)$, it also depends on what I does at times *later* than t ($t' > t$)!

4.7.2 Causality and Analyticity of $g(\omega)$

If we impose causality, however, only values of the stimulus before t can contribute to $R(t)$. Thus, $G(\tau \le 0) = 0$, and we can write a very general relation:

$$R(t) \;=\; \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} G(\tau)\,I(t-\tau)\,\mathrm{d}\tau \tag{4.22}$$

with the frequency dependence of the response function given by:

$$g(\omega) \;=\; \frac{1}{\sqrt{2\pi}} \int_0^\infty G(\tau)\, \mathrm{e}^{\mathrm{i}\omega\tau}\, \mathrm{d}\tau \tag{4.23}$$

Now extend ω to complex values and, on a contour enclosing the upper-half plane, evaluate:

$$\oint_{\mathrm{C}} g(\omega)\, \mathrm{e}^{-\mathrm{i}\omega\tau}\, \mathrm{d}\omega \;=\; \int_{-\infty}^{\infty} g(\omega)\, \mathrm{e}^{-\mathrm{i}\omega\tau}\, \mathrm{d}\omega \;+\; \int_{|\omega|\to\infty} g(\omega)\, \mathrm{e}^{-\mathrm{i}\omega\tau}\, \mathrm{d}\omega$$

Since we can choose any value of τ, let $\tau < 0$. Then the first term on the right, which is just $G(\tau)$ (up to $\sqrt{2\pi}$), vanishes because of causality. The integrand of the second term is bounded by $|g|\mathrm{e}^{\Im(\omega)\tau}$; because here $\Im(\omega) > 0$, the integrand vanishes at infinity when $\tau < 0$, and the contour integral is zero. This establishes[†] that $g(\omega)$ is analytic in the upper half-plane when causality is imposed. Thus, causality implies analyticity!

Other general information about the behaviour of the response function can be derived by further postulating that $G(\tau)$ goes to 0 at least as fast as $1/\tau$ as $\tau \to \infty$. Then, if we repeatedly integrate by parts the integral in eq. (4.23), we get, up to an irrelevant factor of $1/\sqrt{2\pi}$:

$$g(\omega) \;=\; \frac{\mathrm{i}}{\omega}\, G(0) \;-\; \frac{1}{\omega^2}\, \mathrm{d}_\tau G\big|_{\tau=0} \;+\; \ldots$$

The first term vanishes from causality, and we find that:

$$\Re[g(\omega)] \underset{\omega\to\infty}{\approx} \mathrm{O}(1/\omega^2) \qquad\qquad \Im[g(\omega)] \underset{\omega\to\infty}{\approx} \mathrm{O}(1/\omega^3) \tag{4.24}$$

Thus, $|g(\omega)| \to 0$ as $\omega \to \infty$, and we have shown that g satisfies the conditions under which Hilbert transforms may be written:

$$\Re[g(\omega)] \;=\; \frac{1}{\pi}\, \mathcal{P} \int_{-\infty}^{\infty} \frac{\Im[g(\omega')]}{\omega' - \omega}\, \mathrm{d}\omega', \qquad \Im[g(\omega)] \;=\; -\,\frac{1}{\pi}\, \mathcal{P} \int_{-\infty}^{\infty} \frac{\Re[g(\omega')]}{\omega' - \omega}\, \mathrm{d}\omega' \tag{4.25}$$

With a little extra manipulation, we can rewrite these as more useful integrals over positive frequencies. To do this, we must use the fact that reality of $I(t)$ and $R(t)$, and thus of $G(t)$, demands that $g(-\omega) = g^*(\omega)$. Indeed, $g(\omega)$ being a *complex* function of a real variable, take its complex conjugate. Because $R(t)$ and $I(t)$ are real, consistency demands (EXERCISE) that $g(-\omega) = g^*(\omega)$. This is a version of the **Schwarz reflection principle**.

In terms of real and imaginary parts, we have the **crossing relations**:

$$\Re[g(-\omega)] \;=\; \Re[g(\omega)] \qquad\qquad \Im[g(-\omega)] \;=\; -\,\Im[g(\omega)] \tag{4.26}$$

In other words, the real part of g is even, and the imaginary part odd, in their argument. Then:

$$\begin{aligned}
\Re[g(\omega)] \;&=\; \frac{1}{\pi}\, \mathcal{P} \left[\int_0^\infty \frac{\Im[g(-\omega')]}{-\omega' - \omega}\, \mathrm{d}\omega' \;+\; \int_0^\infty \frac{\Im[g(\omega')]}{\omega' - \omega}\, \mathrm{d}\omega' \right] \\
&=\; \frac{1}{\pi}\, \mathcal{P} \left[\int_0^\infty \frac{\Im[g(\omega')]}{\omega' + \omega}\, \mathrm{d}\omega' \;+\; \int_0^\infty \frac{\Im[g(\omega')]}{\omega' - \omega}\, \mathrm{d}\omega' \right] \\
&=\; \frac{2}{\pi}\, \mathcal{P} \int_0^\infty \frac{\omega'\, \Im[g(\omega')]}{\omega'^2 - \omega^2}\, \mathrm{d}\omega'
\end{aligned}$$

[†] It could be argued that this is not sufficient because there could exist singularities whose contributions to the contour integral cancel out. But it is easy to show that the n^{th}-order derivative of $g(\omega)$ always exists, simply by differentiating eq. (4.23) n times and noting that the resulting integrand remains bounded when $\tau > 0$ and $\Im(\omega) > 0$. On the real ω axis, $g(\omega)$ is bounded, and any branch point can be bypassed without changing anything.

$\Im[g(\omega)]$ can be rewritten (exercise) in the same fashion, and we end up with the **dispersion relations**:

$$\begin{aligned} \Re[g(\omega)] &= \frac{2}{\pi}\,\mathcal{P}\int_0^\infty \frac{\omega'\,\Im[g(\omega')]}{\omega'^2-\omega^2}\,\mathrm{d}\omega' \\ \Im[g(\omega)] &= -\frac{2\,\omega}{\pi}\,\mathcal{P}\int_0^\infty \frac{\Re[g(\omega')]}{\omega'^2-\omega^2}\,\mathrm{d}\omega' \end{aligned} \tag{4.27}$$

These were originally derived in 1926-7 for electrical susceptibility $\chi(\omega)$ by Kramers and Kronig.

This type of relation, which can be established with a minimum of assumptions (causality), can be very useful and exists in other areas of physics (it was often used in particle physics in the sixties, for instance).

If $\Im[g(\omega)]$ is sharply peaked around some frequency, say ω_p, we approximate it by $\Im[g(\omega')] = g_\mathrm{p}\delta(\omega'-\omega_\mathrm{p})$, in which case the integrand in the first relation is no longer singular at ω, and we can drop the $\mathcal{P}$ to get:

$$\Re[g(\omega)] = \frac{2}{\pi}\,g_\mathrm{p}\,\frac{\omega_\mathrm{p}}{\omega_\mathrm{p}^2-\omega^2}$$

Therefore, in a frequency range where there is a peak in $\Im[g(\omega)]$, $\Re[g(\omega)]$ decreases. In optics, this is known as anomalous dispersion.

On the other hand, we can ask what happens if there is a frequency range over which $\Im[g(\omega)]$ is very small. Taking $\Im[g(\omega)] \approx 0$ for $\omega_1 < \omega < \omega_2$ in the first dispersion relation of eq. (4.27), there should be little contribution from that frequency range. Then the rate of change of the real part of $g(\omega)$ is:

$$\mathrm{d}_\omega\Re[g(\omega)] = \frac{4\,\omega}{\pi}\int_0^{\omega_1}\frac{\omega'\,\Im[g(\omega')]}{(\omega'^2-\omega^2)^2}\,\mathrm{d}\omega' + \frac{4\,\omega}{\pi}\int_{\omega_2}^\infty\frac{\omega'\,\Im[g(\omega')]}{(\omega'^2-\omega^2)^2}\,\mathrm{d}\omega'$$

If $\Im[g(\omega)] > 0$, as is the case for the index of refraction $n(\omega)$ in an absorptive medium, the two terms are positive. Therefore, because of causality, the index of refraction in a medium increases with frequency, except in regions where $\Im[n(\omega)]$ (absorption) is significant! This is called normal dispersion, because it is the usual case which holds away from the absorption peaks.

We can also say something about $\Re[g(\omega)]$ in the high-frequency regime. We remember that at high frequency, eq. (4.24) says that the imaginary part of $g(\omega)$, goes like $1/\omega^3$. The integral in the dispersion relation for $\Re[g(\omega)]$ is then dominated by the low-frequency behaviour of $\Im[g(\omega)]$, in the sense that $\omega' \ll \omega$, and we can set $\omega'^2-\omega^2 \approx -\omega^2$ in the dispersion relation. We obtain a **sum rule** that is fully consistent with eq. (4.24):

$$\lim_{\omega\to\infty}\omega^2\,\Re[g(\omega)] = -\frac{2}{\pi}\int_0^\infty \omega'\,\Im[g(\omega')]\,\mathrm{d}\omega' \tag{4.28}$$

For much more detail about the application of the principle of causality to the response of a dispersive dielectric medium to electromagnetic waves, see sections 7.5 and 7.8 of Jackson's *Classical Electrodynamics*.

4.7.3 Dispersion relations with one subtraction

When deriving Hilbert transforms, it is possible to relax the condition that $|f(z)| \to 0$ at infinity, although there will be a (fairly modest) price to pay. Suppose that instead $|f(z)|$ goes to some constant which may not be known. The function $(f(z)-f(x_1))/(z-x_1)$ does go to zero as $|z|\to\infty$ in the upper-half plane because of its $1/z$ dependence; and if $f(z)$ is analytic, it is also analytic. Thus, when restricted to a real argument, it satisfies eq. (4.18):

$$\mathcal{P}\int_{-\infty}^\infty \frac{f(x')-f(x_1)}{(x'-x)(x'-x_1)}\,\mathrm{d}x' = \mathrm{i}\pi\,\frac{f(x)-f(x_1)}{x-x_1}$$

or, rearranging:

$$\begin{aligned} \mathrm{i}\pi\,\big(f(x)-f(x_1)\big) &= (x-x_1)\,\mathcal{P}\int_{-\infty}^\infty \frac{f(x')}{(x'-x)((x'-x_1)}\,\mathrm{d}x' - f(x_1)\,\mathcal{P}\int_{-\infty}^\infty \frac{x-x_1}{(x'-x)((x'-x_1)}\,\mathrm{d}x' \\ &= (x-x_1)\,\mathcal{P}\int_{-\infty}^\infty \frac{f(x')}{(x'-x)((x'-x_1)}\,\mathrm{d}x' - f(x_1)\left(\mathcal{P}\int_{-\infty}^\infty\frac{\mathrm{d}x'}{x'-x} - \mathcal{P}\int_{-\infty}^\infty\frac{\mathrm{d}x'}{x'-x_1}\right) \end{aligned}$$

In section 4.5 the last two principal-value integrals were evaluated and found to vanish when the limits of integration extend to infinity. Taking the real and imaginary parts as before, we arrive at the **once-subtracted** Hilbert transforms:

$$\begin{aligned}\Re[f(x)] &= \Re[f(x_1)] + \frac{x-x_1}{\pi}\,\mathcal{P}\int_{-\infty}^{\infty}\frac{\Im[f(x')]}{(x'-x)((x'-x_1)}\,\mathrm{d}x' \\ \Im[f(x)] &= \Im[f(x_1)] - \frac{x-x_1}{\pi}\,\mathcal{P}\int_{-\infty}^{\infty}\frac{\Re[f(x')]}{(x'-x)((x'-x_1)}\,\mathrm{d}x'\end{aligned} \tag{4.29}$$

The integrals can be split in two and once-subtracted dispersion relations involving only positive frequencies derived, analogous to eq. (4.27). We see that we need to know, eg. in the first relation, the real part of f at some point on top of its imaginary part everywhere.

4.7.4 Arrival of a Signal After Propagation

As another interesting physical application of these ideas, consider a plane electromagnetic wave train at normal incidence from vacuum to a medium of refraction index $n(\omega)$. Then the Fourier representation of the wave amplitude inside the medium ($x > 0$) is:

$$\psi(x,t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left(\frac{2}{1+n(\omega)}\right)A_{\mathrm{i}}(\omega)\,\mathrm{e}^{\mathrm{i}\omega[n(\omega)x/c-t]}\,\mathrm{d}\omega \qquad x>0$$

where:

$$A_{\mathrm{i}}(\omega) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\psi_{\mathrm{i}}(0,t')\,\mathrm{e}^{\mathrm{i}\omega t'}\,\mathrm{d}t' \qquad (\text{acts as } i(\omega))$$

is the Fourier transform of the real incident wave amplitude as it reaches the medium. The frequency-domain kernel function would be $g(\omega) = 2\mathrm{e}^{\mathrm{i}\omega n(\omega)x/c}/(1+n(\omega))$.

The integral for ψ can be evaluated by contour integration in the complex ω plane. The integral enclosing the whole upper half-plane vanishes: the integrand is analytic there, because electromagnetic considerations show that both $A(\omega)$ and $n(\omega)$ are analytic there. Also from electromagnetism, when $|\omega| \to \infty$, $n(\omega) \to 1$ and the argument of the exponential becomes $\mathrm{i}\omega[x-ct]/c$, so that the contribution from the semi-circle at infinity also vanishes if $x > ct$. Then the contribution along the real ω-axis must also vanish, and there is no wave amplitude for $x > ct$. This shows, without any detailed knowledge of $n(\omega)$, that no signal can propagate in any medium faster than c. We can say that analyticity implies causality (see also the more sophisticated argument at the end of BF 6.6).

4.8 Bessel Functions (BF 6.9)

The function $\mathrm{e}^{(w-1/w)z/2}$, which is analytic $\forall\, w \in \mathbb{C}$ except at $w = 0$, can be expanded in a Laurent series for any $|w| \neq 0$ in a region centered on, but not including, $w = 0$:

$$\mathrm{e}^{(w-1/w)z/2} = \sum_{n=-\infty}^{\infty} J_n(z)\,w^n$$

where, choosing the unit w-circle as contour and making the change of variable $w = \mathrm{e}^{\mathrm{i}\theta}$:

$$\begin{aligned}J_n(z) = \frac{1}{2\pi\,\mathrm{i}}\oint_{|w|=1}\frac{\mathrm{e}^{(w-1/w)z/2}}{w^{n+1}}\,\mathrm{d}w &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\mathrm{e}^{\mathrm{i}(z\sin\theta - n\theta)}\,\mathrm{d}\theta \\ &= \frac{1}{\pi}\int_0^{\pi}\cos(n\theta - z\sin\theta)\,\mathrm{d}\theta\end{aligned} \tag{4.30}$$

This is an integral representation for the n^{th} order **Bessel function of the first kind** $J_n(z)$ ($n \in \mathbb{Z}$), which satisfies (see pp. BF372-373) the **Bessel equation**:

$$z^2\,\mathrm{d}_z^2 J_n(z) + z\,\mathrm{d}_z J_n(z) + (z^2 - n^2)\,J_n(z) = 0 \tag{4.31}$$

$\mathrm{e}^{(w-1/w)z/2}$ is said to be a **generating function** for the $J_n(z)$; these are analytic at $z = 0$.

Bessel functions arise in the theory of diffraction, in the description of electromagnetic modes inside cavities, and also as solutions of Laplace's equation in cylindrical coordinates.

4.9 Calculus of Residues and Some Applications (BF 6.8)

4.9.1 Theorem of Residues

We know from the Cauchy-Goursat theorem that the contour integral of a function analytic on and inside a contour C vanishes. We now show how the integral can be evaluated even if there is a finite number of singularities within the contour.

The way to do this is to choose another contour within which the function is analytic, and therefore over which its integral still vanishes. This time the path leaves the original contour at a point closest to a singular point, goes straight to within a distance of this point, goes all the way around the point *clockwise* on a circle, and finally goes back to the original contour on the same interconnecting line, but in the opposite direction. The same procedure is repeated for each singularity. It should be obvious that the second path is the sum of the original contour, the incoming and outgoing paths to each singular point which generate cancelling contributions, and the sum of all the circular paths around each of the singular points. Since the original contour is traversed counterclockwise by convention, we have the **Residue Theorem** for n singularities:

$$\oint_C f(z)\,\mathrm{d}z \;=\; 2\pi\,\mathrm{i}\sum_{i=1}^{n}\mathrm{Res}[f(z_i)] \qquad\qquad \mathrm{Res}[f(z_i)] \;=\; \frac{1}{2\pi\,\mathrm{i}}\oint_{C_i} f(z)\,\mathrm{d}z \tag{4.32}$$

where each C_i is the circle around one, and only one, singular point, which is now traversed *counterclockwise*, and $\mathrm{Res}[f(z_i)]$ is called the **residue** at the singular point z_i. It should be clear from eq. (4.15) that a residue is just the A_{-1} coefficient of a Laurent series expansion of $f(z)$.

The Cauchy-Gauss integral formula (4.11) is often useful to evaluate residues, for instance, if the integrand in the residue integral only has simple (order 1) poles:

$$f(z) \;=\; \frac{g(z)}{(z-z_1)(z-z_2)...(z-z_n)}$$

Then we simply write:

$$\oint_{C_i}\frac{g(z)\,\mathrm{d}z}{(z-z_1)(z-z_2)...(z-z_n)} \;=\; \oint_{C_i}\frac{g(z)}{(z-z_1)\cdots(z-z_{i-1})(z-z_{i+1})\cdots(z-z_n)}\,\frac{1}{z-z_i}\,\mathrm{d}z$$

The sum of all the residues is just:

$$\sum\frac{1}{2\pi\mathrm{i}}\oint_{C_i} f(z)\,\mathrm{d}z \;=\; \frac{g(z_1)}{(z_1-z_2)...(z_1-z_n)} \;+\; \ldots \;+\; \frac{g(z_n)}{(z_n-z_1)(z_n-z_2)...(z_n-z_{n-1})}$$

If there are poles of higher order, we can always use eq. (4.12) to express the residue in terms of derivatives, eg.:

$$\frac{1}{2\pi\mathrm{i}}\oint\frac{g(z)}{(z-z_0)^n}\,\mathrm{d}z \;=\; \frac{1}{(n-1)!}\,\mathrm{d}_z^{n-1}g(z)\Big|_{z_0}$$

Another technique, illustrated by example 4.6, is to expand the integrand in a Laurent series. The contour integral of the $n \geq 0$ terms vanishes because of analyticity, and the remaining negative powers integrate to:

$$\frac{1}{2\pi\mathrm{i}}\oint_{C_0} f(z)\mathrm{d}z \;=\; \frac{1}{2\pi\mathrm{i}}\oint_{C_0}\sum_0^\infty\frac{A_{-n}}{(z-z_0)^n}\,\mathrm{d}z \;=\; \sum_0^\infty\frac{1}{(n-1)!}\,\mathrm{d}_z^{n-1}A_{-n}\Big|_{z_0} \;=\; A_{-1}$$

4.9.2 Real definite integrals

Many applications of the calculus of residues involve the evaluation of real definite integrals, which we would like to "complexify" so as to use contour integration. First, we establish a very useful result, generally referred to as **Jordan's lemma**:

Let $f(z)$ be a function whose absolute value is bounded in the upper half-plane, with that bound going uniformly to 0 as $|z| \to \infty$ for $0 \leq \text{phase}\,(z) \leq \pi$; also, let $\alpha > 0$. Then, on a semi-circle of radius R centered on the origin, we have:

$$\lim_{R\to\infty} \int_{\substack{|z|=R \\ y>0}} f(z)\,\mathrm{e}^{\mathrm{i}\alpha z}\,\mathrm{d}z \;=\; 0 \tag{4.33}$$

Indeed, in the polar representation with $z = R\mathrm{e}^{\mathrm{i}\theta}$, the absolute value of the integral is bounded:

$$\left| \int_{\substack{|z|=R \\ y>0}} f(z)\,\mathrm{e}^{\mathrm{i}\alpha z}\,\mathrm{d}z \right| \;\leq\; 2|f(R)|_{\max} \int_0^{\pi/2} R\,\mathrm{e}^{-\alpha R\sin\theta}\,\mathrm{d}\theta \;\leq\; 2|f(R)|_{\max} \int_0^{\pi/2} \mathrm{e}^{-2\alpha R\theta/\pi}\,\mathrm{d}(R\theta)$$

where the second inequality is justified by the fact that $\sin\theta \geq 2\theta/\pi$ over the interval of integration ($\sin\theta - 2\theta/\pi = 0$ at the end-points, is positive at $\pi/4$, and has no minimum in the interval, so is positive or zero everywhere). When $R \to \infty$, the integral on the right converges to a number and, because $|f(z)|_{\max} \to 0$ at infinity, the bound on the absolute value of the integral on the left-hand side also vanishes there. Actually, *Maple* easily evaluates the middle integral as α^{-1} in the limit $R \to \infty$. With $a = \alpha R$:

```
> limit(a*int(exp(-a*sin(theta)),theta=0..Pi/2),a=infinity)/alpha assuming
a>0;
```

$$1/\alpha$$

Whichever way, we obtain the result sought.

Under the same condition on $|f(z)|$ in the lower-half plane, Jordan's lemma holds for $\alpha < 0$, since $\sin\theta < 0$ in that region, provided we use a clockwise contour. From the above proof, the lemma also holds when $\alpha = 0$.

Integrals of the form: $\int_{-\infty}^{\infty} R(x)\,\mathrm{e}^{\mathrm{i}\alpha x}\,\mathrm{d}x$

- We consider first the case where the rational function $R(x)$ has no pole on the real axis. The Residue theorem allows us to evaluate the contour integral in the upper half of the complex plane, $y > 0$:

$$\oint_C R(z)\,\mathrm{e}^{\mathrm{i}\alpha z}\,\mathrm{d}z \;=\; \int_{-\infty}^{\infty} R(x)\,\mathrm{e}^{\mathrm{i}\alpha x}\,\mathrm{d}x \;+\; \lim_{\rho\to\infty} \int_{\substack{|z|=\rho \\ y>0}} R(z)\,\mathrm{e}^{\mathrm{i}\alpha z}\,\mathrm{d}z \;=\; 2\pi\,\mathrm{i} \sum_{y>0} \mathrm{Res}\big[R(z)\,\mathrm{e}^{\mathrm{i}\alpha z}\big]$$

  Provided that $\alpha \geq 0$ and that $|R(z)| \to 0$ uniformly in θ as $\rho \to \infty$, the integral over the semi-circle at infinity in the upper half-plane must vanish because of Jordan's lemma.

  [If $\alpha < 0$, simply take instead a *clockwise* contour in the lower half-plane ($y < 0$), and integrate from 2π to π so that $\sin\theta < 0$ in the exponential.]

  The result is then the following:

$$\int_{-\infty}^{\infty} R(x)\,\mathrm{e}^{\mathrm{i}\alpha x}\,\mathrm{d}x \;=\; \begin{cases} 2\pi\,\mathrm{i} \displaystyle\sum_{y>0} \mathrm{Res}\big[R(z)\,\mathrm{e}^{\mathrm{i}\alpha z}\big] & \alpha > 0 \\ -\,2\pi\,\mathrm{i} \displaystyle\sum_{y<0} \mathrm{Res}\big[R(z)\,\mathrm{e}^{\mathrm{i}\alpha z}\big] & \alpha < 0 \end{cases} \tag{4.34}$$

- Now we let $\alpha = 0$ and take some function $Q(x)$ with simple poles and $|Q(z)| \to 0$ as $|z| \to \infty$. As we did in our treatment of Hilbert transforms, we take a contour consisting of a semi-circle of infinite radius based on the real axis. We avoid the simple poles on the real axis by moving around them clockwise on a

small half-circle of radius ρ in the upper half-plane $y > 0$. Under our assumptions, the contribution from the semi-circle at infinity vanishes. If on the real axis $Q(z)$ has only one simple pole at $x = a$:

$$\mathcal{P}\int_{-\infty}^{\infty} Q(x)\,\mathrm{d}x \;+\; \lim_{\rho\to 0}\int_{\pi}^{0} \mathrm{i}\,Q(a+\rho\,\mathrm{e}^{\mathrm{i}\theta})\,\rho\,\mathrm{e}^{\mathrm{i}\theta}\,\mathrm{d}\theta \;=\; 2\pi\,\mathrm{i}\sum_{y>0}\mathrm{Res}\,Q(z)$$

using the Residue theorem. To make further progress, we first establish another useful result:

If $f(z)$ has a simple pole at $z = a$, and C_ρ is a counterclockwise arc of radius ρ centered on a and intercepting an angle α, then:

$$\lim_{\rho\to 0}\int_{C_\rho} f(z)\,\mathrm{d}z \;=\; \mathrm{i}\,\alpha\,\mathrm{Res}\,f(z)\Big|_{z=a} \tag{4.35}$$

To show this, note that because $f(z)$ must be of the form $g(z)/(z-a)$:

$$\int_{C_\rho}\frac{g(z)}{z-a}\,\mathrm{d}z \;=\; \int_0^{\alpha}\mathrm{i}\,\frac{g(a+\rho\,\mathrm{e}^{\mathrm{i}\theta})}{\rho\,\mathrm{e}^{\mathrm{i}\theta}}\,\rho\,\mathrm{e}^{\mathrm{i}\theta}\,\mathrm{d}\theta \;=\; \int_0^{\alpha}\mathrm{i}\,g(a+\rho\,\mathrm{e}^{\mathrm{i}\theta})\,\mathrm{d}\theta$$

Since $g(z)$ is analytic at a, it can be Taylor-expanded around a, so that $g(a+\rho\,\mathrm{e}^{\mathrm{i}\theta}) = g(a)+$ terms proportional to positive powers of ρ that do not contribute in the limit $\rho \to 0$. We are left with the result sought:

$$\lim_{\rho\to 0}\int_{C_\rho} f(z)\,\mathrm{d}z \;=\; \mathrm{i}\,\alpha\,g(a) \;=\; \mathrm{i}\,\alpha\,\mathrm{Res}\,f(z)\Big|_{z=a}$$

With $\alpha = -\pi$ (semi-circle traversed clockwise!), and for a finite number of simple poles on the real axis $y = 0$,

$$\mathcal{P}\int_{-\infty}^{\infty} Q(x)\,\mathrm{d}x \;=\; 2\pi\,\mathrm{i}\sum_{y>0}\mathrm{Res}\,Q(z) \;+\; \mathrm{i}\,\pi\sum_{y=0}\mathrm{Res}\,Q(z) \tag{4.36}$$

Integrals of the form: $\int_0^\infty x^{\lambda-1}\,R(x)\,\mathrm{d}x$, $(\lambda \notin \mathbb{Z})$

Here, $R(z)$ must be rational, ie. of the form $P_n(x)/Q_m(x)$, P and Q being polynomials of degree n and m, respectively. $R(z)$ must be analytic at the origin with no poles on the positive real axis. Then $Q_m(x)$ should have no zero or positive real root. As well, $\left|z^\lambda R(z)\right| \to 0$ both when $|z| \to \infty$ and $|z| \to 0$. These last conditions are satisfied if $\lambda > 0$ and if $\lambda + n < m$, which means that $m \geq 1$. Then $R(z)$ is guranteed to have poles somewhere (other than zero) in the complex plane.

Since λ is not an integer, we expect a branch cut in $z^{\lambda-1}$ that starts at the origin and which we take to be along the real axis. Then we only consider the branch:

$$z^{\lambda-1} \;=\; |z|^{\lambda-1}\,\mathrm{e}^{\mathrm{i}(\lambda-1)\theta} \qquad\qquad 0 < \theta < 2\pi$$

It means that, at an arbitrarily small but finite distance above the positive real axis, $z^{\lambda-1} = x^{\lambda-1}$, and $z^{\lambda-1} = x^{\lambda-1}\mathrm{e}^{2\pi\mathrm{i}(\lambda-1)}$ below (infinitesimally close to) the positive real axis.

We take a circular contour C at infinity centered on the origin, that runs back to the origin just below the positive real axis, encircles the origin at infinitesimally small radius, and goes back to infinity just above the positive real axis. The contributions to the contour integral from the two circles vanish, leaving those from the two sides of the cut. So, from the Residue theorem,

$$\begin{aligned}\int_C z^{\lambda-1}\,R(z)\,\mathrm{d}z \;&=\; \int_\infty^0 x^{\lambda-1}\,\mathrm{e}^{2\pi\mathrm{i}(\lambda-1)}\,R(x)\,\mathrm{d}x \;+\; \int_0^\infty x^{\lambda-1}\,R(x)\,\mathrm{d}x \;=\; -\,\frac{2\mathrm{i}\,\sin\lambda\pi}{\mathrm{e}^{-\mathrm{i}\pi\lambda}}\int_0^\infty x^{\lambda-1}\,R(x)\,\mathrm{d}x \\ &=\; 2\pi\,\mathrm{i}\sum_{\text{inside}\,C}\mathrm{Res}\big[z^{\lambda-1}R(z)\big]\end{aligned}$$

which can be rearranged to:

$$\int_0^\infty x^{\lambda-1}\,R(x)\,\mathrm{d}x \;=\; (-1)^{\lambda-1}\,\frac{\pi}{\sin\lambda\pi}\sum_{\text{inside}\,C}\mathrm{Res}\big[z^{\lambda-1}R(z)\big]$$

5 MODULE V — Approximation of Arbitrary Functions with Special Functions

There are some things that we do all the time in physics without quite realising how non-trivial they are. For instance, we often expand functions over *infinite* series of known functions. We write solutions of the Schrödinger equation as a linear combination of an often infinite number of eigenstates; or we expand in Fourier series. When the series have a finite number of terms, there is normally no problem. When they are infinite, however, we must answer the following question: Given some function $h(x)$ defined over $[a,\, b]$, under what conditions is it possible to express $h(x)$ as a linear combination of some possibly infinite set of functions defined over the same interval. Since such linear combinations are in general infinite series, we want to know under what conditions on $h(x)$ and on the set $\{f_n(x)\}$ a series $\displaystyle\sum_n^\infty c_n f_n(x)$:

- converges $\forall\, x\, \in [a,\, b]$;
- converges to $h(x)$ (almost) everywhere in $[a,\, b]$.

Once we have found what kind of functions admit this approximation, we shall have to worry about what happens when we take their derivative, and we will be led to a special kind of second-order differential operator with rich applications.

5.1 Hilbert Space Without Tears (BF 5.1)

5.1.1 Complete, normed vector spaces with inner product

Consider a vector space $\mathcal{V}$, ie. a set which is closed under vector addition and multiplication by a scalar:

$$\alpha\, f_1 \,+\, f_2 \,\in\, \mathcal{V} \qquad\qquad \forall f_1,\, f_2 \,\in\, \mathcal{V}, \qquad \alpha \in \mathbb{C} \tag{5.1}$$

Definition 5.1. $\mathcal{V}$ is said to be **normed** if, $\forall\, f \,\in\, \mathcal{V}$ there exists a real number, the **norm** $\|f\|$ of f, such that: $\|f\| \geq 0$, with $\|f\| = 0$ only if $f = 0$ "almost everywhere"; $\|\alpha f\| = |\alpha|\|f\|$ $\forall \alpha \in \mathbb{C}$; and $\|f + g\| \leq \|f\| + \|g\|$.

Definition 5.2. Let $\{h_n\}$ be a sequence in a normed space. If, for any arbitrarily small positive number ϵ, there is an integer $N(\epsilon)$ such that $\|h_j - h_k\| < \epsilon$ for $j > N$, $k > N$, we say that $\{h_n\}$ is a **Cauchy sequence**.

In an alternative and equivalent definition, the sequence is Cauchy when $\|h_j - h_k\| \to 0$ in the limit $j,\, k \,\to\, \infty$.

Definition 5.3. A vector space is said to be **complete** if there exists no Cauchy sequence of any of its elements which converges to an element which is not in the space.

The space of real numbers is complete in that sense, but the space of rational numbers is not, since, for instance, the (Cauchy) sequence $\{s_N\}$ of partial sums $s_N = \sum_0^N 1/n!$ converges to e, which is not rational.

Quite independently, we can also endow our vector space with a general **inner product** (or **complex sesquilinear form**, as mathematicians call it) that has the following properties:

1. $(f_1,\, f_2)$ is a complex number.
2. $(f,\, f)$ is a *real* number which is either zero or positive.
3. $(f_1,\, f_2) = (f_2,\, f_1)^*$.
4. $(f_1 + f_2,\, \alpha f_3) = \alpha(f_1,\, f_3) + \alpha(f_2,\, f_3)$, where α is a complex scalar.
5. $(\alpha f_1,\, f_2) = (f_2,\, \alpha f_1)^* = \alpha^*(f_2,\, f_1)^* = \alpha^*(f_1,\, f_2)$.

Properties (4) and (5) follow the convention, widely used in physics but not in mathematics, that the inner product is linear in its second argument and antilinear (or conjugate-linear) in its first argument.

5.1.2 Banach and Hilbert spaces of square-integrable functions

Definition 5.4. A vector space which is normed and complete is called a **Banach space**. If, *in addition*, it is endowed with an inner product, it is also a **Hilbert space** $\mathcal{H}$.

Any finite-dimensional vector space is a Banach space. In general, this is *not* the case for the infinite-dimensional spaces which are inevitably involved in the infinite series that we hope to use to approximate functions.

Consider the set $\{f(x)\}$ of all complex-valued functions of a real variable defined on a closed interval $[a,\, b]$ and which are **square-integrable**, in the sense that $\int_a^b |f(x)|^2 \mathrm{d}x$ is finite. Then we take $(f,\, f)$ as the norm of f, and we choose as the inner product :

$$(f_1,\, f_2) \;=\; \int_a^b f_1^*(x)\, f_2(x)\, w(x)\, \mathrm{d}x \tag{5.2}$$

where the real **weight function** $w(x) \geq 0$ everywhere in $[a,\, b]$.

A theorem by Riesz and Fischer guarantees that *the space of square-integrable functions is complete*, and therefore a Hilbert (and Banach) space.

There is a technicality which we should not gloss over. For eq. (5.2) to be an inner product such that $(f,\, f)$ is identified with the norm $\|f\|$, $(f,\, f) = 0$ should imply that f itself vanishes everywhere in $[a,\, b]$. But with our choice of inner product, f is allowed not to vanish at a finite number of points (or, more generally, a countably infinite one), and $(f,\, f)$ will still vanish. If this happens, the set of points at which f does not vanish is said to be a set of measure zero. Integrals involving sets of measure zero are usually handled by Lebesgue integration, as opposed to the usual Riemann kind. The long and the short is that if f is zero "almost everywhere" in $[a,\, b]$ (but not everywhere), we can still have $(f,\, f) = 0$.

With f and g square-integrable functions, it is straightforward to show the identity:

$$\big|(f,\, g)\big|^2 \;\equiv\; (f,\, f)(g,\, g) \;-\; \frac{1}{2} \iint \big|f(x)\, g(y) \;-\; f(y)\, g(x)\big|^2 \,\mathrm{d}x\, \mathrm{d}y$$

from which the **Schwartz inequality** immediately follows:

$$\big|(f,\, g)\big|^2 \;\leq\; (f,\, f)(g,\, g) \tag{5.3}$$

This makes it plain that square-integrability of two functions guarantees that their inner product is finite.

5.2 Orthonormality and Completeness of a Set of Functions (BF 5.2)

We will be especially concerned with the set of functions $f_i \in \mathcal{H}$ **orthonormal** with respect to a positive weight function $w(x)$ on $[a,\, b]$.

Definition 5.5. When

$$(f_i,\, f_j) \;=\; \int_a^b f_i^*(x)\, f_j(x)\, w(x)\, \mathrm{d}x \;=\; \delta_{ij}$$

we say that the f_i are orthonormal with respect to $w(x)$ on $[a,\, b]$.

Now suppose there exists such an infinite orthonormal set $\{f_j(x)\}$, with $j = 1,\, 2,\, \ldots$ Pick some arbitrary $h(x) \in \mathcal{H}$ over $[a,\, b]$, and compute a *finite* number of inner products $c_j = (f_j,\, h)$. Then construct the sequence of partial sums:

$$h_n(x) \;=\; \sum_j^n c_j f_j(x) \qquad n = 1,\, 2,\, \ldots$$

Then, for any given h_n, and using the orthonormality of the f_j, we can immediately write:

$$(h,\, h_n) \;=\; \sum_j^n |c_j|^2, \qquad\qquad (h_n,\, h_n) \;=\; \sum_j^n |c_j|^2$$

while, from the Schwartz inequality, eq. (5.3), we have:

$$(h_n,\, h)^2 \;=\; (h_n,\, h_n)^2 \;\leq\; (h,\, h)(h_n,\, h_n)$$

from which follows the inequality: $s_n = \sum_j^n |c_j|^2 = (h_n,\, h_n) \;\leq\; (h,\, h)$.

The sequence of s_n increases monotonically with n, always remaining bounded, yielding **Bessel's inequality**:

$$(h,\, h) \;\geq\; \lim_{n\to\infty} (h_n,\, h_n) \;=\; \sum_{j=1}^{\infty} \big|(f_j,\, h)\big|^2 \tag{5.4}$$

That the sum on the right converges in the limit $n \to \infty$ is important, because it means that our seemingly *ad hoc* initial construction, the sequence of partial sums h_n, has finite norm and thus also converges in the limit $n \to \infty$. The question is: to what?

Ask a slightly different question: given a square-integrable function $h(x)$ over the interval $[a,\, b]$, we would like to "fit" it with a linear combination of our orthonormal f_j, which are also square-integrable. The best fit is obtained by minimising the mean quadratic error:

$$M \;=\; \lim_{n\to\infty} \left(h - \sum_j^n a_j f_j \,,\; h - \sum_j^n a_j f_j \right) \;=\; (h,\, h) \;-\; \sum_j^{\infty} \big|(f_j,\, h)\big|^2 \;+\; \sum_j^{\infty} \big|a_j - (f_j,\, h)\big|^2 \tag{5.5}$$

By definition, $M \geq 0$, and we already know from Bessel's inequality that the sum of the first two terms on the right is also positive. We conclude that we can minimise M by setting the coefficients a_j equal to the inner products $c_j = (f_j,\, h)$ introduced previously.

Now, is it possible for the mean quadratic error to vanish? Before answering this, let us recall a couple of definitions for different types of convergence of series.

Consider the sequence of partial sums:

$$h_n(x) \;=\; \sum_i^n k_i(x) \qquad n = 1,\, 2,\, \ldots$$

Definition 5.6. If, for any arbitrarily small positive number ϵ, we can find an integer $n = N(\epsilon)$ for which $|h(x) - h_n(x)| < \epsilon$ *anywhere* in $[a,\, b]$ when $n > N$, then the sequence of partial sums $h_n(x)$ converges **uniformly** toward the value $h(x)$:

$$h(x) \;=\; \lim_{n\to\infty} h_n(x)$$

If N also depends on x, we have a weaker **point-wise** convergence.

It will be convenient to relax this criterion slightly by allowing it not to hold at a *finite* number of points in $[a,\, b]$. Then we say that we have **convergence in the mean** if:

$$\lim_{n\to\infty} (h - h_n,\, h - h_n) \;=\; 0 \tag{5.6}$$

Uniform convergence would force $h(x) = \lim h_n(x)$ everywhere in the interval. If this fails to occur at only a countable set of points, this set of points is of measure zero and contributes nothing to the integral in the inner product, so long as we treat the integral as a Lebesgue integral.

Definition 5.7. For a given set $\{f_i\}$ orthonormal on $[a,\, b]$, if it is possible by letting $n \to \infty$ to make the mean quadratic error vanish, ie. if the sequence $\{h_n\}$ with $h_n(x) = \sum c_i f_i(x)$ converges *in the mean* to $h(x)$, then the f_i form a **complete** orthonormal set.

Note that we wrote the coefficients as c_i, not c_{in}, because they do not change as we construct partial sums of higher n to improve the approximation.

When the f_i form a *complete* orthonormal set, the vanishing of the mean quadratic error (eq. (5.5)) as $n \to \infty$ transforms Bessel's inequality into the important **completeness relation** (**Parseval equality**):

$$(h,\, h) \;=\; \sum_j^{\infty} \left|c_j\right|^2 \;=\; \sum_j^{\infty} \left|(f_j,\, h)\right|^2 \tag{5.7}$$

which holds for *any* square-integrable function $h(x)$.

A more intuitive—and certainly more useful than Definition 5.7—criterion for completeness is provided by the following theorem (theorem BF5.2 on p. 222) which we quote without proof:

> A set of orthonormal functions in Hilbert space is complete if, and only if, there exists no function outside the set that is orthogonal to every function, in which case the set is said to be ***closed***.

Essentially, what we are saying here is that a complete, orthonormal set of functions forms a basis for the Hilbert space, in the sense that any function in that space can be expressed as a (possibly infinite) sum over the basis functions.

Note that completeness of $\{f_i\}$ does not allow us to assert that an arbitrary function can be expressed as a linear combination of the f_i *everywhere* in $[a,\, b]$; for this to be true, we need to establish *uniform* convergence. If only convergence in the mean can be proved, there will in general be a countable set of points in $[a, b]$ where we cannot say that $h = \sum c_i f_i$.

5.3 Approximation by Polynomials: the Theorem of Weierstrass (BF5.4)

According to the **Theorem of Weierstrass** (1885), for any function $f(q)$ continuous over $[a,\, b]$, there must exist a sequence of polynomials, $P_N(q)$, that converges uniformly to $f(q)\ \forall q \in [a,\, b]$:

$$f(q) \;=\; \lim_{N\to\infty} P_N(q) \tag{5.8}$$

where:

$$P_N(q) \;=\; \sum_{i=0}^{N} a_i^{(N)}\, q^i$$

Proofs of the theorem by "elementary" methods (see BF) are rather tedious, and more sophisticated proofs—the so-called Stone-Weierstrass Theorem (1937)— require concepts far beyond the scope of this course. One thing, however, should be clear from the outset about what the theorem asserts. Since the coefficients $a_i^{(N)}$ depend on N, we should not think of this representation of $f(q)$ as a power-series expansion, in which the coefficients would not change as more powers are included. As we improve the approximation with higher-order polynomials, coefficients of a given power of q will change!

The theorem may be extended so as to approximate functions of m variables :

$$f(q_1, q_2, \ldots, q_m) \;=\; \lim_{N\to\infty} \sum_{i_1,\ldots,i_m=0}^{N} a_{i_1,\ldots i_m}^{(N)}\, q_1^{i_1} \ldots q_m^{i_m}$$

Although the Theorem of Weierstrass is somewhat unwieldy in actual computations, it can help us discover various complete sets of polynomials over suitable *closed* 1-dim intervals, as well as other sets of complete functions.

Let us illustrate this with one very important case of approximation in terms of trigonometric functions, the **Fourier approximation**.

5.4 Fourier Series (BF 5.6)

5.4.1 Fourier theorem

We look at continuous functions $g(x, y)$ defined in two dimensions, for which the Weierstrass theorem guarantees the existence of functions:

$$g_N(x, y) \;=\; \sum_{j,k=0}^{N} a_{jk}^{(N)}\, x^j\, y^k$$

which approximate uniformly $g(x, y)$ over a finite interval in the $N \to \infty$ limit. Since we are really interested in a 1-dim case, we restrict ourselves to functions $g(x, y) = f(\theta)$ which live on the unit circle and we go to polar coordinates with $x = \cos\theta$, $y = \sin\theta$. Expressing the cosine and sine in terms of exponentials, the theorem of Weierstrass then requires that $f(x)$ can be written in the form:

$$\begin{aligned} f(x) \;&=\; \lim_{N\to\infty} \frac{1}{\sqrt{2\pi}} \sum_{-N}^{N} c_n^{(N)}\, \mathrm{e}^{\mathrm{i}nx} \\ &=\; \lim_{N\to\infty} \left[\frac{a_0^{(N)}}{2} + \sum_{n=1}^{N} \left(a_n^{(N)}\, \cos nx \;+\; b_n^{(N)}\, \sin nx \right) \right] \end{aligned}$$

where $f(x + 2\pi) = f(x)$. Actually, $f(x)$ need not be periodic so long as we are interested in approximating it *only* in the interval $[-\pi,\, \pi]$.

Though uniform, this approximation is not quite satisfactory since the coefficients depend on N. But it points to the set of functions $\mathrm{e}^{\mathrm{i}nx}/\sqrt{2\pi}$ (n any integer) as being worthy of investigation. All that has to be done now is show that the set is orthonormal and complete. Indeed, one easily shows that the set is orthonormal on $[-\pi,\, \pi]$; moreover, it is *complete* (p. BF241). The equivalent set $\left\{1/\sqrt{2\pi},\, (\cos nx)/\sqrt{\pi},\, (\sin nx)/\sqrt{\pi}\right\}$ is also orthonormal and complete.

It follows immediately that an *arbitrary* function can be approximated *in the mean* over $[-\pi,\, \pi]$ by the *expansions*:

$$\frac{1}{\sqrt{2\pi}} \sum_{-\infty}^{\infty} c_n\, \mathrm{e}^{\mathrm{i}nx} \;=\; \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n\, \cos nx \;+\; b_n\, \sin nx \right)$$

The convergence becomes uniform (theorems BF5.4 and 5.5 on pp. 242–245) if the first derivative of the function to be approximated is piecewise continuous over $[-\pi,\, \pi]$. Any point of discontinuity of the function requires special treatment.

A change of variable $x \to \pi x/l$ generalises BF's theorem 5.5 to any interval $[c,\, c + 2l]$, and we can now state the **Fourier-series theorem** in full:

> Let $f(x)$ and its first derivative be piecewise continuous on a *closed* interval $[c,\, c + 2l]$, where $c \in \mathbb{R}$. Then in every *closed* subinterval over which $f(x)$ is continuous, we have *uniform* convergence of the sequence of partial sums to $f(x)$:
>
> $$f(x) \;=\; \frac{1}{\sqrt{2l}} \sum_{-\infty}^{\infty} c_n\, \mathrm{e}^{\mathrm{i}n\pi x/l} \;=\; \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n\, \cos\frac{n\pi x}{l} \;+\; b_n\, \sin\frac{n\pi x}{l} \right) \tag{5.9}$$
>
> where $\left\{\mathrm{e}^{\mathrm{i}n\pi x/l}/\sqrt{2l}\right\}$ and $\left\{1/\sqrt{2\pi},\, (\cos nx)/\sqrt{l},\, (\sin nx)/\sqrt{l}\right\}$ are complete, orthonormal sets over $[c,\, c + 2l]$. The Fourier coefficients are given by:

$$
\begin{aligned}
c_n &= \frac{1}{\sqrt{2l}} \int_c^{c+2l} f(x)\, \mathrm{e}^{-\mathrm{i}n\pi x/l}\, \mathrm{d}x \\
a_n &= \frac{1}{l} \int_c^{c+2l} f(x)\, \cos\frac{n\pi x}{l}\, \mathrm{d}x \\
b_n &= \frac{1}{l} \int_c^{c+2l} f(x)\, \sin\frac{n\pi x}{l}\, \mathrm{d}x
\end{aligned}
\tag{5.10}
$$

At points of discontinuity of $f(x)$, its Fourier series converges to the average of the two limits, from the right and from the left, of the function.

If, and only if, $f(x+2l) = f(x)$, ie. if f is periodic, are these expressions for f valid outside $[c, c+2l]$. If f is non-periodic, we can still use the Fourier theorem in an open, finite interval whose length we identify with $2l$; outside this interval, we simply replace the actual f by an infinite succession of replicas of f over the interval so as to create a function of period $2l$.

The Fourier coefficients c_n, and $\{a_n,\, b_n\}$ satisfy the completeness relation (5.7):

$$
(f,\, f) \;=\; \sum_{-\infty}^{\infty} \left|c_n\right|^2 \;=\; l\left[\frac{a_0^2}{2} + \sum_{n=1}^{\infty}\left(a_n^2 + b_n^2\right)\right]
$$

The Fourier theorem also applies to periodic analytic functions $f(z)$.

5.4.2 Gibbs phenomenon

Note the emphasis on a *closed* interval in the Fourier theorem. If the (sub)interval should be open because $f(x)$ is discontinuous at its end-points, the sequence of partial (truncated) Fourier series converges point-wise, but *not uniformly*, and truncated series may not approximate a function correctly at points that depend on how many terms are kept. To see this, rewrite the truncated exponential Fourier series from eq. (5.9):

$$
\begin{aligned}
f_N(x) &= \frac{1}{2l}\int_0^{2l} \mathrm{d}x'\, f(x') \sum_{-N}^{N} \mathrm{e}^{\mathrm{i}n\pi(x-x')/l} \;=\; \frac{1}{2l}\int_0^{2l} \mathrm{d}x'\, f(x')\, \frac{\sin\left[(N+\frac{1}{2})\pi(x-x')/l\right]}{\sin\frac{\pi(x-x')}{2l}} \\
f_N(\alpha) &= \frac{1}{2\pi}\int_{-\alpha}^{2\pi-\alpha} \mathrm{d}\theta\, f\big(l(\theta+\alpha)/\pi\big)\, \frac{\sin(N+\frac{1}{2})\theta}{\sin\frac{\theta}{2}}
\end{aligned}
$$

where the last equality is obtained from the change of variable and redefinition: $\theta = \pi(x'-x)/l = \pi x'/l - \alpha$.

Now suppose there is a discontinuity in $f(x)$ at some point corresponding to $\alpha = \alpha_0 \in (0,\, 2\pi)$. At the discontinuity, the jump in the truncated sum can be written as: $\Delta f_N = f_N(\alpha_0+\epsilon) - f_N(\alpha_0-\epsilon)$, with ϵ arbitrarily small but not zero. Calling the sine factor in the integrand $F(\theta)$ for brevity, we have:

$$
\begin{aligned}
f_N(\alpha_0+\epsilon) &= \frac{1}{2\pi}\int_{-\alpha_0-\epsilon}^{2\pi-\alpha_0-\epsilon} \mathrm{d}\theta\, f\big[l(\theta+\alpha_0+\epsilon)/\pi\big]\, F(\theta) \\
&= \frac{1}{2\pi}\int_{-\alpha_0-\epsilon}^{-\alpha_0+\epsilon} \mathrm{d}\theta\, f\big[l(\theta+\alpha_0+\epsilon)/\pi\big]\, F(\theta) \;+\; \frac{1}{2\pi}\int_{-\alpha_0+\epsilon}^{2\pi-\alpha_0-\epsilon} \mathrm{d}\theta\, f\big[l(\theta+\alpha_0+\epsilon)/\pi\big]\, F(\theta)
\end{aligned}
$$

The first integral vanishes in the limit $\epsilon \to 0$ because the integrand is then continuous over the interval of integration. Similarly, we can split $f_N(\alpha_0-\epsilon)$ into an integral from $-\alpha_0+\epsilon$ to $2\pi-\alpha_0-\epsilon$ plus another integral from $2\pi-\alpha_0-\epsilon$ to $2\pi-\alpha_0+\epsilon$. Again, continuity of the integrand makes the latter integral vanish. We are left with:

$$
\Delta f_N \;=\; \frac{1}{2\pi}\int_{-\alpha_0+\epsilon}^{2\pi-\alpha_0-\epsilon} \mathrm{d}\theta\, \big[f\big(l(\theta+\alpha_0+\epsilon)/\pi\big) \;-\; f\big(l(\theta+\alpha_0-\epsilon)/\pi\big)\big]\, F(\theta)
$$

Now, for large N, the function $F(\theta)$ more or less vanishes outside the interval $[-\pi/(N+1/2),\, \pi/(N+1/2)]$; inside this interval, θ is very small, and we can approximate the jump of the truncated sum with:

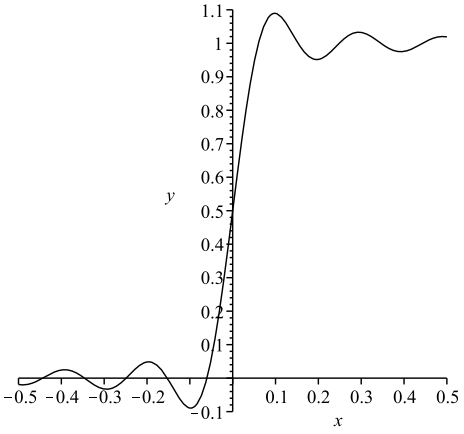
$$\Delta f_N \;\approx\; 2\,\frac{\Delta f}{\pi}\,\int_0^{\pi/(N+1/2)} \frac{\sin(N+1/2)\theta}{\theta}\,\mathrm{d}\theta \;=\; 1.18\,\Delta f$$

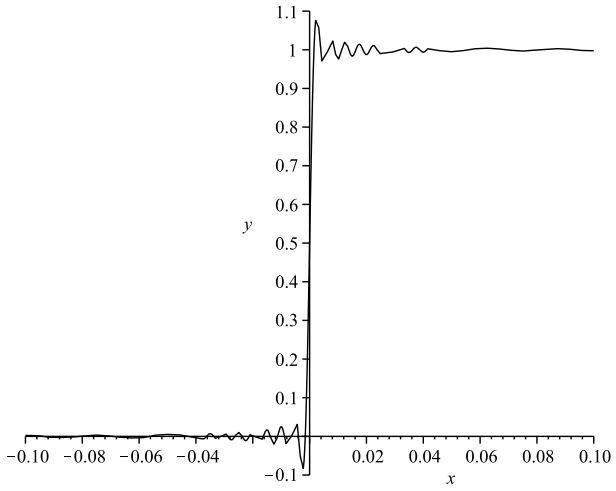
from which we conclude that, very close ($\approx \pi/N$) to a discontinuity, the Fourier approximation overshoots the function by about 18%. This is often called the **Gibbs phenomenon**.

Here is an example of the Gibbs phenomenon applied to a square wave with a discontinuity of 1 at $x = 0$. I have plotted the Fourier series:

$$y(x) \;=\; \frac{1}{2} \,+\, \frac{2}{\pi}\,\sum_{k=1}^{N} \frac{1}{2k+1}\,\sin\frac{(2k+1)\pi\,x}{10}$$

for $N = 50$ (on the left) and $N = 2000$ (on the right). As expected, the series overshoots 1 by about 0.09, independent of N so long as it is large enough. Note also the change of scale of the x-axis needed because the overshoot becomes narrower as N increases.





5.4.3 Multidimensional Fourier series

The Fourier theorem is readily extended to more than one dimension. Let $F(\mathbf{x})$ be a N-dimensional separable function, in the sense that it is a product of N 1-dim periodic functions f_i, each depending on one variable x_i, each having the Fourier decomposition:

$$f_i(x_i) \;=\; \frac{1}{\sqrt{2l_i}}\,\sum_{n_i=-\infty}^{\infty} c_{n_i}\,\mathrm{e}^{\mathrm{i}k_i x_i}$$

where $k_i = \pi n_i/l_i$. Then the Fourier expansion of F is:

$$F(\mathbf{x}) \;=\; \frac{1}{\sqrt{V}}\,\sum_{\mathbf{n}} c_{\mathbf{n}}\,\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{x}} \tag{5.11}$$

where $c_{\mathbf{n}} = c_{n_1}\cdots c_{n_N}$ and $V = 2^N(l_1\cdots l_N)$. We shall take this expression as the definition of the Fourier expansion of a function of N variables that is periodic in all the variables.

5.5 Fourier Integrals (BF 5.7) and Dirac Delta-Function (BF 5.3)

5.5.1 Fourier integrals

We would like to approximate a non-periodic function over the whole axis instead of just over some finite interval. Using the complex-exponential form of the Fourier theorem, we write:

$$f(x) \;=\; \frac{1}{\sqrt{2l}} \sum_{n=-\infty}^{\infty} \left(\frac{1}{\sqrt{2l}} \int_{-l}^{l} f(x')\, \mathrm{e}^{-\mathrm{i}n\pi x'/l}\, \mathrm{d}x' \right) \mathrm{e}^{\mathrm{i}n\pi x/l}$$

Now, in the summation, we redefine the index n so that $k_n = n\pi/l$, where the new running index jumps by values of $\Delta k = \pi/l$. Implementing this change in the summand gives:

$$f(x) \;=\; \frac{1}{\sqrt{2\pi}} \sum_{k_n=-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}} \int_{-l}^{l} f(x')\, \mathrm{e}^{-\mathrm{i}k_n x'}\, \mathrm{d}x' \right) \Delta k\, \mathrm{e}^{\mathrm{i}k_n x}$$

Letting $l \to \infty$, or $\Delta k \to \mathrm{d}k$, one transforms the summation into a **Fourier integral** for a piecewise *smooth* $f(x)$:

$$f(x) \;=\; \sqrt{\frac{|a|}{2\pi}} \int_{-\infty}^{\infty} g(k)\, \mathrm{e}^{\mathrm{i}akx}\, \mathrm{d}k \qquad\qquad g(k) \;=\; \sqrt{\frac{1}{2\pi|a|}} \int_{-\infty}^{\infty} f(x)\, \mathrm{e}^{-\mathrm{i}akx}\, \mathrm{d}x \tag{5.12}$$

where a is chosen *by convention*; some references choose $a = 2\pi$, others choose $a < 0$. Here we shall opt for another popular choice: $a = 1$. Also, we often write the Fourier transform of $f(x)$ as $f(k)$, or $\tilde{f}(k)$

$f(x)$ and $g(k)$ are called the **Fourier transforms** of each other[†] and provide equivalent representations of the same object. It is very easy to check (EXERCISE) that they inherit each other's symmetry (if any) in one dimension.

At points of discontinuity of $f(x)$, its Fourier transform converges to the average of the two limits, from the right and from the left, of the function.

It is tempting to see $f(x)$ in eq. (5.12) as an expansion over a continuous basis of functions $\mathrm{e}^{\mathrm{i}kx}$. But this is not quite like the discrete basis $\{\mathrm{e}^{\mathrm{i}n\pi x/l}/\sqrt{2l}\}$ whose elements certainly belong to the Hilbert space defined over the *finite* interval $(-l,\, l)$. Here, the interval is infinite, and the functions $\mathrm{e}^{\mathrm{i}kx}$ are not square-integrable over that interval. Therefore, they do not belong to the corresponding Hilbert space. Clearly, there must be a restriction on $f(x)$ which propagates to its Fourier transform $g(k)$. This restriction is that $f(x)$ must be piecewise smooth (its derivative may not be discontinuous at more than a finite number of points) and that the integral of $|f(x)|$ over the whole x axis must exist.

Example 5.1. A nice model for pulses is the Gaussian time-domain function $g(t) = \mathrm{e}^{-t^2/2\tau^2}$. This pulse is centered at $t = 0$, and τ controls its width at any particular time. Transforming to the frequency domain gives:

$$g(\omega) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-t^2/2\tau^2 - \mathrm{i}\omega t}\, \mathrm{d}t \;=\; \frac{\mathrm{e}^{-\omega^2\tau^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-(t+\mathrm{i}\omega\tau)^2/2\tau^2}\, \mathrm{d}t$$

where the last equality is obtained by completing the square in the argument of the exponential. The integral, of the form $\int \mathrm{e}^{-x^2/2}\, \mathrm{d}x$, is easily found by ***Maple/Mathematica*** or in tables, but it is instructive to see how we can evaluate it ourselves. Write in terms of two Cartesian variables and change to circular coordinates:

$$\begin{aligned} \iint_{-\infty}^{\infty} \mathrm{e}^{-(x^2+y^2)/2}\, \mathrm{d}x\, \mathrm{d}y \;&=\; \int_{\phi=0}^{2\pi} \mathrm{d}\phi \int_{\rho=0}^{\infty} \mathrm{e}^{-\rho^2/2}\, \rho\, \mathrm{d}\rho \;=\; 2\pi \int_{u=0}^{\infty} \mathrm{e}^{-u}\, \mathrm{d}u \\ &=\; 2\pi \end{aligned}$$

[†] As noted above, the sign of the argument of the exponentials is also a matter of convention, but it must change from one transform to the other.

But this integral is just the square of $\int \mathrm{e}^{-x^2/2}\,\mathrm{d}x$. We arrive at:

$$g(\omega) \;=\; \tau\,\mathrm{e}^{-\omega^2\tau^2/2}$$

Extending Fourier integrals to three dimensions is straightforward; with our conventions:

$$\begin{aligned} F(\mathbf{x}) \;&=\; \frac{1}{(2\pi)^{3/2}}\int_{\text{all space}} G(\mathbf{k})\,\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{x}}\,\mathrm{d}^3k \\ G(\mathbf{k}) \;&=\; \frac{1}{(2\pi)^{3/2}}\int_{\text{all space}} F(\mathbf{x})\,\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{x}}\,\mathrm{d}^3x \end{aligned} \tag{5.13}$$

EXERCISE: Calculate the 3-dim Fourier transform, $f(\mathbf{k})$, of $f(r) = \mathrm{e}^{-\alpha r}/r$, where $\alpha > 0$ and r is the radial coordinate. Then take the limit $\alpha \to 0$. What happens if you had tried to find the Fourier transform of $1/r$ directly?

5.5.2 Differentiating Fourier transforms

Differentiating Fourier transforms is not hard, and it can bring dividends:

$$\partial_n f(\mathbf{x}) \;=\; \frac{1}{(2\pi)^{3/2}}\int \partial_n\big[f(\mathbf{k})\,\mathrm{e}^{\mathrm{i}k_jx^j}\big]\,\mathrm{d}^3k \;=\; \frac{1}{(2\pi)^{3/2}}\int f(\mathbf{k})\,\mathrm{i}k_n\,\mathrm{e}^{\mathrm{i}k_jx^j}\,\mathrm{d}^3k \qquad \partial_n \;=\; \partial/\partial x^n$$

Then it is easy to see that a second-order ordinary differential equation with constant coefficients: $c_2\,\mathrm{d}_x^2 f(x) + c_1\,\mathrm{d}_x f(x) + c_0\,f(x) \;=\; F(x)$, becomes:

$$\frac{1}{\sqrt{2\pi}}\int f(k)\,\big(-c_2\,k^2 + \mathrm{i}\,c_1\,k + c_0\big)\,\mathrm{e}^{\mathrm{i}kx}\,\mathrm{d}k \;=\; \frac{1}{\sqrt{2\pi}}\int F(k)\,\mathrm{e}^{\mathrm{i}kx}\,\mathrm{d}k$$

Then, because the Fourier transform of the zero function vanishes everywhere, the differential equation is turned into the *algebraic* equation:

$$f(k) \;=\; \frac{F(k)}{-c_2\,k^2 + \mathrm{i}\,c_1\,k + c_0} \tag{5.14}$$

In principle, the solution $f(x)$ follows by taking the Fourier transform of $f(k)$, but that integral can be quite difficult to evaluate. We will look at some ways of doing this in a later module.

One important use for Fourier series and integrals is the decomposition of a suitable function into a superposition of oscillating functions in their trigonometric or complex exponential forms. This **spectral decomposition** is in terms of **harmonics**, or **modes**, of frequency $n\pi/l$ and **amplitude** $|c_n|$ [eq. (5.9)] when performed over a closed interval $[-l,\,l]$. Over an infinite interval, the spectrum is said to be **continuous** with **spectral density** given by $g(k)$ in eq. (5.13), where k is the frequency.

The 3-dim algebraic equation inherits (EXERCISE) the vector structure of the derivative: $\boldsymbol{\nabla} \to \mathrm{i}\mathbf{k}$, $\boldsymbol{\nabla}\cdot \to \mathrm{i}\mathbf{k}\cdot$, $\boldsymbol{\nabla}\times \to \mathrm{i}\mathbf{k}\times$.

5.5.3 Dirac delta function

A famous Fourier integral is of a finite wave train $f(x) = \mathrm{e}^{\mathrm{i}k_0x}$ over $[-l,\,l]$ and $f(x) = 0$ elsewhere. Its spectral density is:

$$g(k) \;=\; \frac{1}{\sqrt{2\pi}}\int_{-l}^{l}\mathrm{e}^{-\mathrm{i}(k-k_0)x}\,\mathrm{d}x \;=\; \sqrt{2\pi}\,\frac{\sin(k-k_0)l}{\pi(k-k_0)}$$

The longer the wave train, the more sharply peaked its spectral density around the single spatial frequency k_0. You can watch it as it happens by entering the following command in ***Maple***:

```
plots[animate](plot,sin((k-k0)*x)/(k-k0),k=-a..a,x=1..N);
```

where k0 is a number and a should be no more than ten percent of k0; N can be very large but about twenty times k0 should be sufficient. Right-click on the output plot and choose Animate → Play in the drop-down menu. You can vary the speed of the animation (FPS) at the top of the worksheet.

If we take the limit $l \to \infty$, we find that:

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}(k-k_0)x}\,\mathrm{d}x \;=\; \lim_{u\to\infty}\frac{\sin(k-k_0)u}{\pi(k-k_0)}$$

Another instructive manipulation consists in inserting $g(k)$ into $f(x)$, both in eq. (5.12):

$$f(x) \;=\; \int_{-\infty}^{\infty} f(x')\left[\frac{1}{2\pi}\int_{-\infty}^{\infty}\mathrm{e}^{\mathrm{i}k(x-x')}\,\mathrm{d}k\right]\mathrm{d}x'$$

Compare this to the defining equation for the **Dirac δ-function**:

$$f(a) \;=\; \int_{-\infty}^{\infty} f(x)\,\delta(a-x)\,\mathrm{d}x \tag{5.15}$$

to obtain the Fourier representation of the δ-function, perhaps the most useful:

$$\delta(x-x') \;=\; \frac{1}{2\pi}\int_{-\infty}^{\infty}\mathrm{e}^{\mathrm{i}k(x-x')}\,\mathrm{d}k \;=\; \lim_{u\to\infty}\frac{\sin(x-x')u}{\pi(x-x')} \tag{5.16}$$

The last equality makes it plain that there must be an implicit limit in the integral since its integrand is not square-integrable.

Other expressions (BF 5.3) for the δ-function which can be useful are the Gaussian and Lorentzian representations:

$$\delta(x-x') \;=\; \lim_{\tau\to 0}\frac{1}{\sqrt{2\pi}\,\tau}\,\mathrm{e}^{-(x-x')^2/2\tau^2} \;=\; \lim_{\epsilon\to 0}\frac{1}{\pi}\,\frac{\epsilon}{(x-x')^2+\epsilon^2} \tag{5.17}$$

The δ-function $\delta(x)$ vanishes everywhere except at $x = 0$, where it is singular. Because its integral over any interval containing $x = 0$ does not vanish, it is not a function in the usual sense. Although something like it was introduced by Weierstrass to prove his approximation theorem in 1885, only in 1947 was it rigorously interpreted by Laurent Schwartz (1915–2002) in his theory of distributions.

The δ-function can be used to extend the concept of orthonormality to a set $\{f_k(x)\}$ of functions with a *continuous* index k. The functions are said to be orthonormal if:

$$\int_{-\infty}^{\infty} f_k(x)\,f_{k'}(x)\,\mathrm{d}x \;=\; \delta(k-k') \tag{5.18}$$

It also allows us to prove the **Parseval's theorem** (or identity), for a Fourier integral:

$$\begin{aligned}\int_{-\infty}^{\infty}\left|f(x)\right|^2\mathrm{d}x \;&=\; \int \mathrm{d}k\,g^*(k)\int \mathrm{d}k'\,g(k')\left[\frac{1}{2\pi}\int \mathrm{e}^{\mathrm{i}(k'-k)x}\,\mathrm{d}x\right] \;=\; \int \mathrm{d}k\,g^*(k)\int \mathrm{d}k'\,g(k')\,\delta(k'-k)\\ &=\; \int_{-\infty}^{\infty}\left|g(k)\right|^2\mathrm{d}k\end{aligned} \tag{5.19}$$

This says that the norm of a function is invariant under a Fourier transform. See section 14.5 in Jackson's *Classical Electrodynamics* for an application to the calculation of the energy radiated per unit solid angle by a charge in arbitrary motion.

The generalisation of the δ-function's Fourier representation to three dimensions is straightforward:

$$\delta^3(\mathbf{x}-\mathbf{x}') \;=\; \frac{1}{(2\pi)^3}\int_{-\infty}^{\infty}\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}')}\,\mathrm{d}^3k \tag{5.20}$$

Some care must be exercised, however, when relating 3-dim to 1-dim δ-functions in curvilinear coordinates:

$$\delta^3(\mathbf{x}-\mathbf{x}') \;=\; \delta(x-x')\,\delta(y-y')\,\delta(z-z') \;=\; \frac{1}{r^2\,\sin\theta}\,\delta(r-r')\,\delta(\theta-\theta')\,\delta(\phi-\phi')$$

where the first expression is in Cartesian and the second line in spherical coordinates. This is so that, in *any* coordinate system, the integral of $\delta^3(\mathbf{x}-\mathbf{x}')$ over whole space is 1. There is some awkwardness about $\delta(r)$ since $r \geq 0$: normally, the limits of integration in $\int \delta(x)\,\mathrm{d}x = 1$ *contains* 0, where $\delta(x)$ is singular, but with r the lower limit lies at the singularity.. An exception has to be allowed for the radial variable, and $\delta(r)$ *defined* so that $\int \delta(r) f(r) \mathrm{d}r = f(0)$

Moreover, the δ-function can be expanded over *any* complete orthonormal set $\{f_k(x)\}$ (EXERCISE):

$$\delta(x-x') \;=\; \sum_k^{\infty} f_k(x)\, f_k^*(x') \tag{5.21}$$

When k is *continuous*, replace the summation by an integral over k. The δ-function has lots of representations!

5.5.4 Convolution

Definition 5.8. The **convolution** $[f_1 \star f_2](x)$ of two functions f_1 and f_2 is defined by:

$$[f_1 \star f_2](x) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f_1(x')\, f_2(x-x')\,\mathrm{d}x' \tag{5.22}$$

This can be viewed as a weighted average of the input function f_1 at the point x with weighting given by $f_2(x-x')$. As x varies, the weight function emphasises different regions of the input function. The convolution is commutative, associative, distributive, and linear.

If we express the right-hand side of eq. (5.22) in terms of the Fourier transforms, g_1 and g_2, of f_1 and f_2:

$$\begin{aligned}
[f_1 \star f_2](x) \;&=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f_1(x') \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g_2(k)\, \mathrm{e}^{\mathrm{i}k(x-x')}\,\mathrm{d}k\right] \mathrm{d}x' \\
&=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f_1(x')\, \mathrm{e}^{-\mathrm{i}kx'}\,\mathrm{d}x'\right] g_2(k)\, \mathrm{e}^{\mathrm{i}kx}\,\mathrm{d}k \qquad (5.23) \\
&=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g_1(k)\, g_2(k)\, \mathrm{e}^{\mathrm{i}kx}\,\mathrm{d}k \qquad (5.24)
\end{aligned}$$

the **convolution theorem** then says that the Fourier transform $G(k)$ of $[f_1 \star f_2](x)$ is the product $g_1(k)g_2(k)$ of the Fourier transforms of f_1 and f_2. In other words, in Fourier space, the convolution integral becomes multiplication.

Convolution is an important tool in signal processing.

5.5.5 Discrete and Fast Fourier Transforms

The Fourier transform is invaluable for the study of functions whose analytical form is known. In numerical calculations, however, all we have is a table with a set of points in, say, time, and values of the function at those discrete points. How can we adapt our Fourier formalism to deal with such cases?

We assume that N values of this function $f(t_j)$, ($t_j = j\Delta t$), are given at equal intervals $\Delta t = T/N$ over a total period T. When we try to write the Fourier transform $f(\omega)$, we should not expect to be able to construct a continuous transform: it too will be discretised at frequencies $\omega_k = k\Delta\omega$, where the frequency interval $\Delta\omega$ is to be determined. Letting $\int \longrightarrow \sum$ and $\mathrm{d}t \longrightarrow \Delta t$, we write:

$$f(\omega_k) \;=\; \frac{1}{\sqrt{2\pi}} \frac{T}{N} \sum_{j=0}^{N-1} f(t_j)\, \mathrm{e}^{-\mathrm{i}\omega_k t_j}$$

We will also want to reconstruct $f(t_j)$ from $f(\omega_k)$. Our final result will look prettier if we multiply this by $\sqrt{2\pi N}/T$ and redefine $f(\omega_k)$ so that $\sqrt{2\pi N} f(\omega_k)/T \longrightarrow f(\omega_k)$. Then, multiplying by $\mathrm{e}^{\mathrm{i}\omega_k t_n}$ and summing over k, there comes:

$$\sum_{k=0}^{N-1} f(\omega_k)\, \mathrm{e}^{\mathrm{i}\omega_k t_n} \;=\; \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} f(t_j) \sum_{k=0}^{N-1} \mathrm{e}^{\mathrm{i}\omega_k (t_n - t_j)} \tag{5.25}$$

With $\omega_k = k\Delta\omega$, we recognise that the last sum over k is a geometric series of the form: $\sum_{k=0}^{N-1} r^k = \frac{1-r^N}{1-r}$ with $r = \mathrm{e}^{\mathrm{i}\Delta\omega(t_n - t_j)}$, and we obtain:

$$\sum_{k=0}^{N-1} \mathrm{e}^{\mathrm{i}k\Delta\omega(t_n-t_j)} = \frac{1-\mathrm{e}^{\mathrm{i}N\Delta\omega(t_n-t_j)}}{1-\mathrm{e}^{\mathrm{i}\Delta\omega(t_n-t_j)}}$$

Now, if we can demand that this expression vanish when $n \neq j$, we will have an expression for $f(t_j)$ that looks usable. This will happen if $N\Delta\omega(t_n - t_j) = N\Delta\omega\,\Delta t(n-j)$ is equal to $2\pi(n-j)$. Indeed, when $n \neq j$, the numerator vanishes, but the denominator does not. And when $n = j$, the sum is simply N. Also, since $\Delta t = T/N$, we find the frequency interval we had to determine: $\Delta\omega = 2\pi/T$. Eq. (5.25) now equals just $\sqrt{N} f(t_n)$, and we arrive at the discrete transforms:

$$f(\omega_k) = \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} f(t_j)\,\mathrm{e}^{-\mathrm{i}\omega_k t_j} \qquad f(t_j) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} f(\omega_k)\,\mathrm{e}^{\mathrm{i}\omega_k t_j} \tag{5.26}$$

where $t_j = j\Delta t = jT/N$ and $\omega_k = 2\pi k/T$. Other properties, such as the convolution theorem and Parseval's identity, are readily rewritten in their discretised form.

Discrete Fourier transforms can be used for the numerical solution of differential equations as well as for the spectrum analysis of signals, but they can lead to rather heavy computations. Because k and j each take N values, our calculation contains a $N \times N$ matrix, and it will take a time of order N^2. Fortunately, there exists a ridiculously simple way of speeding up such calculations tremendously: the **Fast Fourier Transform** (FFT).

To get some idea of how it works, we shall require that $N = 2^l$, with l a positive integer. Now split the sum for $f(\omega_k)$ in eq. (5.26) into two sums, one over even, and one over odd, values of j. Since $\omega_k t_j = 2\pi kj/N$, we have:

$$f(\omega_k) = \sum_{j'=0}^{N'-1} f(t_{2j'})\,\mathrm{e}^{-\mathrm{i}2\pi kj'/N'} + \mathrm{e}^{-\mathrm{i}2\pi k/N} \sum_{j'=0}^{N'-1} f(t_{2j'+1})\,\mathrm{e}^{-\mathrm{i}2\pi kj'/N'}$$

Each sum looks exactly like the original sum, except that N has been replaced by $N' = N/2$. Now, under $k \to k + N'$, each sum is invariant and the exponential factor in front of the odd sum switches sign, so that only $N' = N/2$ values of k need computing. Therefore, each sum corresponds to a $\frac{N}{2} \times \frac{N}{2}$ matrix, and here is the magic: the computation's time is now of order $2(N/2)^2 = N^2/2$, half of what it was before.

Since $N/2$ is an even number, we can repeat the process, dividing each sum into two, ending up with a computational time of order $4(N/4)^2 = N^2/4$. After l such steps, what remains is a sum of N complex products! The computational time is now lN, ie. $N\log_2 N$. For 1024 points, a FFT computation would be about 100 times faster than a direct calculation.

5.5.6 The Sampling Theorem (Whittaker,..., Shannon)

This important theorem asserts that a frequency band-limited function is completely determined by sampling at a frequency (the **Nyquist frequency**) that is at least twice the highest frequency occurring in the spectrum of the function. The sampling theorem also tells us how to reconstruct the function from its samples. Analog to digital conversion relies heavily on it.

With Δt the period between samplings, we sample the continuous function $f(t)$ at $t_n = n\Delta t$ with an operator called the **Dirac comb**:

$$\mathrm{comb}(t) = \sum_{n=-\infty}^{\infty} \delta(t - t_n)$$

There is a corresponding sampling frequency ω_s. The result is a sampled function: $f_\mathrm{s}(t) = f(t)\,(\Delta t)\,\mathrm{comb}(t)$.

Now the comb-"function" admits a Fourier representation (EXERCISE):

$$\mathrm{comb}(t) \;=\; \frac{1}{\Delta t}\sum_{k=-\infty}^{\infty} \mathrm{e}^{\mathrm{i}k\omega_\mathrm{s}t}$$

Therefore:

$$f_\mathrm{s}(t) \;=\; \sum_{k=-\infty}^{\infty} f(t)\,\mathrm{e}^{\mathrm{i}k\omega_\mathrm{s}t}$$

What is the Fourier transform of $f_\mathrm{s}(t)$? By linearity, just the sum of the transforms of $f(t)\,\mathrm{e}^{\mathrm{i}k\omega_\mathrm{s}t}$, ie.:

$$\tilde{f}_\mathrm{s}(\omega) \;=\; \sum_{k=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(t)\,\mathrm{e}^{\mathrm{i}(k\omega_\mathrm{s}-\omega)t}\mathrm{d}t = \sum_{k=-\infty}^{\infty} \tilde{f}(\omega - k\omega_\mathrm{s})$$

where we have used a tilde to make it clearer when we are talking about Fourier transforms.

What we have just found is that the sampled frequency spectrum is made of an infinite number of replicas of the spectrum of the original function around integer multiples of ω_s.

Now suppose the frequency spectrum of f has **limited bandwith**, in the sense that $\tilde{f}(\omega) = 0$ when $\omega > \omega_\mathrm{max}$. The sampled spectrum still consists of copies of this original spectrum. In practice, this means that for some sampled value we cannot disentangle the contribution from $\tilde{f}(\omega < \omega_\mathrm{max})$, which is the one that carries information, from contributions at $\omega + k\omega_\mathrm{s}$ *which are spurious since there is no spectrum there*! If, however, we sample at a frequency larger than the **Nyquist frequency**, $\omega_\mathrm{N} = 2\omega_\mathrm{max}$, then the copies all lie above ω_max, and we can pick out the actual spectrum by **filtering** $\tilde{f}_\mathrm{s}(\omega)$:

$$\tilde{f}(\omega) \;=\; \mathrm{rect}(\omega/\omega_\mathrm{s})\tilde{f}_\mathrm{s}(\omega)$$

where we have introduced the **rectangular function**:

$$\mathrm{rect}(u) \;=\; \begin{cases} 1 & |u| < \frac{1}{2} \\ 0 & |u| > \frac{1}{2} \end{cases} \tag{5.27}$$

According to the convolution theorem, $f(t)$ can be seen as the convolution of $\widetilde{\mathrm{rect}}$ and f_s:

$$f(t) \;=\; \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \widetilde{\mathrm{rect}}(t')\,f_\mathrm{s}(t-t')\,\mathrm{d}t'$$

The Fourier transform of the rectangle function is easily found (EXERCISE): $\widetilde{\mathrm{rect}}(t) \;=\; \frac{\sqrt{2\pi}}{\Delta t}\,\mathrm{sinc}\,(t/\Delta t)$, where:

$$\mathrm{sinc}(x) \;\equiv\; \frac{\sin \pi x}{\pi x} \tag{5.28}$$

is the **sine cardinal** function[†]. Inserting this in the integral expression for $f(t)$ yields:

$$f(t) \;=\; \int_{-\infty}^{\infty} \frac{\mathrm{sinc}(t'/\Delta t)}{\Delta t}\,f_\mathrm{s}(t-t')\,\mathrm{d}t' \;=\; \sum_{n=-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{sinc}(t'/\Delta t)\,f(t-t')\,\delta(t-t_n-t')\,\mathrm{d}t'$$

Thus, we arrive at:

$$f(t) \;=\; \sum_{n=-\infty}^{\infty} f(t_n)\,\mathrm{sinc}\left(\frac{t-t_n}{\Delta t}\right) \tag{5.29}$$

If the sampling frequency is above the Nyquist frequency, this exactly reconstructs f, *even between samples*!

Important as they may be, Fourier transforms fail to provide useful and accurate representations of many "real-world" functions. Also, even the FFT method may be too slow for very large sets of data. More recently, a new type of method, called **wavelet** analysis has been developed that obviates these problems. This topic lies well outside the scope of this course.

[†]Sometimes, $\mathrm{sinc}(x)$ is defined as $(\sin x)/x$.

5.6 Some Special Functions and their Defining Differential Equations

In the second half of this module we introduce a unified approach to the study of functions which although called "special" are ubiquitous in mathematical physics. The reason for the name has more to do with the fact that they cannot be expressed in closed form in terms of elementary functions. Historically, they often appeared first as solutions to second-order differential equations. In fact, these equations, along with appropriate boundary conditions, contain all the information about the functions.

5.6.1 Self-adjoint differential operators (BF5.10)

Before we discuss ordinary differential equations, there are some things we should know about their main ingredient: the differential operator $\mathrm{d}_x \equiv \mathrm{d}/\mathrm{d}x$. This will be acting on functions living in a Hilbert space endowed with the extended inner product introduced in Definition 5.2:

$$(f,\, g) \;=\; \int_{-\infty}^{\infty} f^*(x)\, g(x)\, w(x)\, \mathrm{d}x \tag{5.30}$$

with $w(x) \geq 0$. The weight function can be used to restrict the effective interval to a finite one when appropriate.

In general, however, the outcome of differentiation will *not* be an element of the Hilbert space of which the function being differentiated is an element. It is easy to find examples of this, for instance the function $\sqrt{x-a}$ which is square-integrable over $[a,\, b]$ while its derivative is not. Because of this, we say that d_x is not bounded. This means that in a given interval we will have to restrict the set of functions on which the operator may act—its **domain**— by specifying boundary conditions.

The kind of differential operator we are interested in takes the general form:

$$L \;=\; \alpha(x)\, \mathrm{d}_x^2 \,+\, \beta(x)\, \mathrm{d}_x \,+\, \gamma(x) \tag{5.31}$$

where $\alpha(x)$, $\beta(x)$ and $\gamma(x)$ are real. If eigenvalue equations of the type $L[f_n] = \lambda f_n$ are to be consistent, we would like $L[f_n]$ to be square-integrable if the eigenfunction f_n is square-integrable.

Such operators have an **adjoint** $L^\dagger$ which by definition satisfies:

$$(L^\dagger[f],\, g) \;=\; (f,\, L[g]) \tag{5.32}$$

We shall demand that L be **self-adjoint** (Hermitian), ie. those which satisfy the condition $\big(L[f],\, g\big) = \big(f,\, L[g]\big)$ for any functions in Hilbert space on a given interval. Then the eigenfunctions f_n that correspond to different eigenvalues are *orthogonal* and these eigenvalues are *real*. Indeed:

$$\big(f_n,\, L[f_m]\big) \,-\, (L[f_n],\, f_m) \;=\; (\lambda_m \,-\, \lambda_n^*)\, (f_n,\, f_m) \;=\; 0$$

The case $m = n$ constrains λ_n to be real, and $m \neq n$ yields $(f_n,\, f_m) = 0$. This easily derived result is very important. When we can convince ourselves that the orthonormal f_n form a complete set, we will have constructed a basis for any function $f \in \mathcal{H}$ functions in $[a,\, b]$. Then we will have:

$$L[f] \;=\; \sum a_n\, L[f_n] \;=\; \sum a_n\, \lambda_n\, f_n$$

which will guarantee that $L[f] \in \mathcal{H}$! We shall show in the next section how this can be done for Sturm-Liouville operators with polynomial solutions.

What restrictions does all this put on the coefficients in L, and perhaps also on the functions on which it acts? Before we find out, let us bring in a very useful object:

Definition 5.9. Let f_1 and f_2 be differentiable functions, and $f'(x) \equiv \mathrm{d}_x f(x)$. Their **Wronskian** is:

$$W_{f_1 f_2}(x) \;=\; f_1(x)\, f_2'(x) \,-\, f_2(x)\, f_1'(x) \tag{5.33}$$

$W(x)$ can be written as a determinant, which allows us to extend the definition to n functions f_i:

$$W_{f_1 \ldots f_n}(x) \;=\; \det \begin{pmatrix} f_1 & \mathrm{d}_x f_1 & \cdots & \mathrm{d}_x^{n-1} f_1 \\ f_2 & \mathrm{d}_x f_2 & \cdots & \mathrm{d}_x^{n-1} f_2 \\ \vdots & & \vdots & \\ f_n & \mathrm{d}_x f_n & \cdots & \mathrm{d}_x^{n-1} f_n \end{pmatrix} \tag{5.34}$$

The Wronskian provides a very useful and easy test of linear independence: Let f_1 and f_2 be differentiable and non-vanishing everywhere in the interval $[a,\, b]$; their Wronskian, $W_{f_1 f_2}(x)$, vanishes $\forall\, x \in [a,\, b]$ if, and only if, f_1 and f_2 are linearly dependent everywhere (f_2 is the *same* multiple of $f_1\ \forall\, x$) in $[a,\, b]$ (EXERCISE).

Now we are ready to compute $\big(f,\, L[g]\big) - \big(L[f],\, g\big)$. The result is:

$$\big(f,\, L[g]\big) \;-\; \big(L[f],\, g\big) \;=\; w\alpha\, W_{f^* g}(x)\Big|_{-\infty}^{\infty} \;\; - \int_{-\infty}^{\infty} \big[(w\alpha)' \;-\; w\beta\big]\, W_{f^* g}(x)\, \mathrm{d}x$$

The right-hand side vanishes and L is Hermitian (self-adjoint) if, an only if:

$$w\alpha\, W_{f^* g}\Big|_{-\infty}^{\infty} \;=\; 0 \tag{5.35}$$

$$(w\alpha)' \;=\; w\beta \qquad \Longleftrightarrow \qquad w\alpha \;=\; C \exp\left[\int^x \frac{\beta(x')}{\alpha(x')}\, \mathrm{d}x'\right] \tag{5.36}$$

where C is an arbitrary constant.

Our generic second-order operator (5.31) is not self-adjoint in general. But we can always use a weight function $w(x)$ that satisfies eq. (5.36) to make it so! Indeed, simply multiply L by w and apply $w\beta = (w\alpha)'$:

$$L' \;= wL \;=\; \mathrm{d}_x\left[(w\alpha)\, \mathrm{d}_x\right] \;+\; \gamma\, w \tag{5.37}$$

Provided condition (5.35) is met, $\big(f,\, L'[g]\big) = \big(L'[f],\, g\big)$, and L' is self-adjoint. Then we say that L' is the self-adjoint form of L. It is clear from condition (5.36) that if $\beta = \alpha'$, then w is a constant which can be set to 1, and L is already in self-adjoint form, acquiring full self-adjoint status when, from eq. (5.35), either α vanishes outside a finite interval (or fast enough at infinity), or the Wronskian vanishes fast enough at infinity.

Example 5.2. One form of the equation satisfied by Bessel functions of order n is:

$$L[J_n(x)] \;=\; \big(\mathrm{d}_x^2 \;+\; \frac{1}{x}\, \mathrm{d}_x \;-\; \frac{n^2}{x^2} \;+\; k^2\big)\, J_n(x) \;=\; 0$$

A quick computation using the solution for $w(x)$ in eq. (5.36) yields $w(x) = x$, As $\beta = 1/x \neq \mathrm{d}_x \alpha$, L is not in self-adjoint form. We can, however, multiply L by x to obtain a self-adjoint operator:

$$L[J_n] \;=\; \mathrm{d}_x\left(x\, \mathrm{d}_x\, J_n(x)\right) \;+\; \big(k^2\, x \;-\; \frac{n^2}{x}\big)\, J_n(x)$$

subject (since $w\alpha = x$) to appropriate boundary conditions on the $J_n(x)$.

Suppose now that we find that $w(x)$ vanishes on the boundaries of an interval of interest. Since the constant C in eq. (5.36) is arbitrary, we can match $w(x)$ ($x \in [a,\, b]$) at the boundaries to other functions $w(x)$ *outside* the interval for which the arbitrary constant C in eq. (5.36) vanishes. $w(x)$ can then be defined as:

$$w(x) \;=\; \begin{cases} \frac{1}{\alpha} \exp\left[\int^x \frac{\beta(x')}{\alpha(x')}\, \mathrm{d}x'\right] & a \leq x \leq b \\ 0 & x > b,\; x < a \end{cases}$$

This allows us to keep using the whole axis as the interval of integration in inner products if this makes the integral easier to calculate.

Definition 5.10. Over the interval $[a,\, b]$, the self-adjoint operator:

$$\mathrm{d}_x\left[(w\alpha)\,\mathrm{d}_x\right] + \gamma\, w$$

together with either the **separated boundary conditions** for the functions $f \in \mathcal{H}$ on which it acts:

$$c_1\, f(a) + c_2\, \mathrm{d}_x f\Big|_a = 0, \qquad d_1\, f(b) + d_2\, \mathrm{d}_x f\Big|_b = 0$$

(with at least one constant in each equation being non-zero) or the **periodic boundary conditions**:

$$f(a) = f(b), \qquad \mathrm{d}_x f\Big|_a = \mathrm{d}_x f\Big|_b$$

for the functions on which it acts, form[†] a **Sturm-Liouville system**.

With appropriate boundary conditions on its domain, the operator $L = \mathrm{d}_x^2 + q(x)$ is already self-adjoint and in Sturm-Liouville form, since $\alpha = 1$. It turns out that it is always possible to redefine the functions on which L in eq. (5.31) acts so as to transform it into the form with only a second-order derivative with a constant coefficient. Some of you have done precisely this when you eliminated the first-order derivative in the radial Schrödinger equation for a spherically-symmetric potential by redefining the radial solutions: $R(r) = u(r)/r$.

Many differential equations in physics are **eigenvalue equations** of the form $L[f] = \lambda f$. If $L[f]$ is in its self-adjoint form, and if f is subject to separated or periodic boundary conditions as defined above, we say that $L[f] = \lambda f$, where λ is the eigenvalue associated with eigenfunction f, is a **Sturm-Liouville** problem.

We note that if there were two eigenfunctions, f_1 and f_2, differentiable everywhere and corresponding to the same eigenvalue, that satisfy the same eigenvalue equation and *separated* boundary condition, say at $x = a$, then the conditions could be trivially rewritten as:

$$\begin{pmatrix} f_1(a) & f_1'(a) \\ f_2(a) & f_2'(a) \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = 0$$

But since c_1 and c_2 cannot both vanish, the Wronskian $W_{f_1 f_2}(x)$ must vanish at the boundary and therefore (differentiate its definition and eliminate second-order derivatives using $L[f] = \lambda f$, then integrate to see that if $W_{f_1 f_2}(x) = 0$ anywhere in the interval, $W_{f_1 f_2}(x) = 0$ everywhere) and the functions are linearly dependent. This shows that the eigenvalues of a Sturm-Liouville operator with *separated* boundary conditions are non-degenerate.

It is important to remember that the eigenfunctions of a Sturm-Liouville operator do not form a basis for the *whole* of Hilbert space, ie. for all square-integrable functions, but only for the subspace of square-integrable functions that satisfy the appropriate boundary conditions for the corresponding Sturm-Liouville problem. Nevertheless, from Sturm-Liouville problems we can extract several bases composed of known functions over which to expand many functions of interest.

Example 5.3. In this example, we illustrate the results of choosing different intervals and boundary conditions for the same Sturm-Liouville operator, d_x^2, with eigenvalue equation:

$$\mathrm{d}_x^2 f_n + \lambda_n\, f_n = 0 \qquad \lambda_n > 0$$

defined over either $[0,\, L]$ or $[-L,\, L]$. The eigenfunctions depend on the interval we choose. Over $[0,\, L]$ and with the separated boundary conditions $f_n(0) = 0$, $f_n(L) = 0$, the eigenfunctions are $\sin(n\pi x/L)$. Note that these conditions are *not* periodic! Indeed, the boundary conditions satisfied by the derivatives are $\mathrm{d}_x f_n|_{x=0} = (n\pi/L)$ and $\mathrm{d}_x f_n|_{x=L} = (-1)^n(n\pi/L)$ which are not identical.

If instead we take $[-L,\, L]$ as our interval and *periodic* boundary conditions, $f_n(-L) = f_n(L)$ and $\mathrm{d}_x f_n|_{x=-L} = \mathrm{d}_x f_n|_{x=L}$, the allowed eigenfunctions are now $\sin(n\pi x/L)$ *and* $\cos(n\pi x/L)$.

We now have doubly degenerate eigenvalues, allowed since the boundary conditions are not separated.

The eigenvalues of the operator, however, are the same in all cases: $\lambda_n = \omega_n^2 = (n\pi/L)^2$ ($n \in \mathbb{Z}^+$).

[†]Here we shall not bother with the distinction between regular and singular Sturm-Liouville systems.

5.6.2 Orthonormal polynomial solutions of Sturm-Liouville problems (BF 5.10)

We now focus on a particular class of solutions to a Sturm-Liouville problem: orthonormal *polynomials*, $Q_n(x)$, of *finite* degree n. Weierstrass guarantees that functions can be approximated by polynomials only on *finite*, not semi-infinite or infinite, intervals, but if we can establish that a set of orthonormal polynomials is complete, we have a basis for our Hilbert space. There will in general also exist non-polynomial solutions which fall outside the scope of the following discussion.

We begin by establishing the *completeness* of the *infinite* set $\{Q_n\}$ of any polynomial eigenfunctions of a Sturm-Liouville problem over the whole real axis. This can be done by showing that $\{Q_n\}$ is closed, ie. that there exists no function orthogonal to all the Q_n. Using our extended inner product, and for any real k:

$$(\mathrm{e}^{\mathrm{i}kx},\, f) \;=\; \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}kx}\, f(x)\, w(x)\, \mathrm{d}x \;=\; \sum_{m=0}^{\infty} \frac{(-\mathrm{i}k)^m}{m!} \int_{-\infty}^{\infty} f(x)\, x^m\, w(x)\, \mathrm{d}x \;=\; \sum_{m=0}^{\infty} \frac{(-\mathrm{i}k)^m}{m!}\, (x^m, f)$$

Now it is always possible to take the first polynomials of degree $\leq m$ and solve for x^m as a linear combination of th. This means that, if there exists a function f such that $(f,\, Q_n) = 0\ \forall\, n$, then $(x^m,\, f) = 0\ \forall\, m$. Note that since n is not bounded, neither is m, and we find that $(\mathrm{e}^{\mathrm{i}kx},\, f) = 0$.. This means that the Fourier transform of $f(x)w(x)$ vanishes, which can only be true if $f(x)w(x) = 0$. Now, by definition, $w(x) = 0$ only outside the interval of interest, so $f(x) = 0$ in that interval—except perhaps at a countable number of points. This is precisely the criterion for completeness. In regions where $w(x) = 0$, $f(x)$ is arbitrary because it does not contribute to the inner product; the Hilbert space is restricted to the interval in which $w(x) \neq 0$.

Restricting to the polynomial solutions of a Sturm-Liouville problem has many consequences. Firstly, we can not rely on the vanishing of the Wronskian in condition 5.35 to get a self-adjoint differential operator L. There exists no finite interval at both boundaries of which polynomials of *any* degree all vanish; and polynomials do not vanish at infinity. We must therefore demand that $w\alpha$ vanish at the boundaries of any finite interval we choose. If the interval is infinite, $w\alpha$ must go to zero faster than $1/x^m$ for *any* $m > 0$.

This in turns means that we must exclude operators with $\beta(x) = 0$, otherwise condition 5.35 would force $w\alpha$ to be constant, and therefore zero, everywhere! Put another way, Sturm-Liouville operators of the form $\mathrm{d}_x^2 + \gamma(x)$ do not have polynomial eigenfunctions.

Now, demanding that the eigenfunctions of any second-order operator L be polynomials $Q_n(x)$ constrains its coefficients to be themselves polynomials, whose degree is easily determined. Since it cannot vary, we can nail it down just by inserting in turn Q_0, Q_1 and Q_2 in $L[Q_n] = \lambda_n Q_n$, transforming it into an *algebraic* system.

When $n = 0$, the derivative terms in L contribute nothing, leaving $\gamma(x) = \lambda_0$. The other two equations are:

$$\begin{aligned} \beta(x)\, \mathrm{d}_x Q_1 \,+\, \lambda_0\, Q_1 \;&=\; \lambda_1\, Q_1 \qquad (n \,=\, 1) \\ \alpha(x)\, \mathrm{d}_x^2 Q_2 \,+\, \beta(x)\, \mathrm{d}_x Q_2 \,+\, \lambda_0\, Q_2 \;&=\; \lambda_2\, Q_2 \qquad (n \,=\, 2) \end{aligned}$$

Inserting $\beta(x)$ from the first equation into the second, there comes:

$$\begin{aligned} \beta(x) \;&=\; (\lambda_1 - \lambda_0)\, \frac{Q_1}{Q_1'} \;\equiv\; \beta_1 x \,+\, \beta_0 \\ \alpha(x) \;&=\; (\lambda_2 - \lambda_0)\, \frac{Q_2}{Q_2''} \,-\, (\lambda_1 - \lambda_0)\, \frac{Q_1}{Q_1'}\, \frac{Q_2'}{Q_2''} \;\equiv\; \alpha_2 x^2 \,+\, \alpha_1 x \,+\, \alpha_0 \end{aligned} \tag{5.38}$$

where we have used the prime notation for the derivatives so as to minimise clutter.

Sturm-Liouville Eigenvalues

We can solve for the eigenvalues of a Sturm-Liouville operator $L = \alpha \mathrm{d}_x^2 + \beta\, \mathrm{d}_x + \gamma(x)$ with polynomial eigenfunctions in terms of a given α and β to find that they depend only on the coefficient of the highest power of x in α and β (EXERCISE):

$$\lambda_n \;=\; n(n-1)\, \alpha_2 \,+\, n\, \beta_1 \tag{5.39}$$

where $\gamma = \lambda_0$ has been absorbed into λ_n. This shows that the eigenvalues have a multiplicity of 1, ie., they are non-degenerate.

The form of eq. (5.39) suggests that we shift all the eigenvalues by λ_0, which will not change the eigenfunctions. This is equivalent to taking $\gamma(x) = \gamma_0 = 0$. Now the five parameters α_0, α_1, α_2, β_0 and β_1 will determine the eigenfunctions, but the eigenvalues depend only on α_2 and β_1.

But some of these parameters can be set arbitrarily. This arbitrariness manifests itself in the fact that we can make the change of variable $x \to ax + b$, where a and b ar real constants, without changing the eigenvalues, only the polynomials which now are functions of $ax + b$. Also, any change of the coefficients that results in scaling the eigenvalues (multiplying all of them by the same constant), will not be considered as fundamentally distinct.

Parity of the Sturm-Liouville Polynomials

In one dimension, the parity operator P reverses the sign of x in an expression. Therefore, it is idempotent: $P^2 = \mathsf{I}$, the identity operator. By operating twice on an arbitrary 1-dim eigenfunction of P, its eigenvalues are easily found to be ± 1.

Now suppose that L is invariant under P, ie. $\alpha_1 = \beta_0 = 0$ in eq. (5.38). Then $PL[Q_n(x)] = L[P\,Q_n(x)] = \lambda_n \big(PQ_n(x)\big)$, and $P\,Q_n(x)$ is *also* an eigenfunction of L with the same eigenvalue λ_n. Because there is only one eigenfunction for a given λ_n, we must have $P\,Q_n(x) = c_n\,Q_n(x)$. Q_n is thus an eigenfunction of P as well, with *well-defined parity* ± 1. We conclude that when $\alpha_1 = \beta_0 = 0$, the Sturm-Liouville polynomials contain only even *or* odd powers of x, not both.

5.6.3 The classical Sturm-Liouville operators

We are now ready to find *all* the polynomials that solve a Sturm-Liouville problem. Since, as we know, $\beta(x) \neq 0$, we must calculate a weight function $w(x)$ that can turn L into its standard Sturm-Liouville form.

- $\alpha = \alpha_0$ *constant*

  The eigenvalues are then $n\beta_1$. We choose $\alpha_0 = 1$, noting that if it isn't, we can simply multiply L by an appropriate constant, leaving the eigenfunctions the same and re-scaling the eigenvalues. Then, choosing $C = \mathrm{e}^{\beta_0^2/2\beta_1}$ in eq. (5.36):

  $$w(x) \;=\; C \exp\left[\int^x \frac{\beta(x')}{\alpha(x')}\,\mathrm{d}x'\right] \;=\; \exp\left[\frac{\beta_1}{2}\left(x + \frac{\beta_0}{\beta_1}\right)^2\right]$$

  $w(x)$ is non-zero over the whole x-axis. We are at liberty to choose the origin as we wish without loss of generality, so we consider only $\beta_0 = 0$. We also remember that $w\alpha = w$ must go to zero fast enough at infinity, because the Wronskian in eq. (5.35), being made of polynomials, will not. For this it is sufficient that β_1 be negative. In fact, we can make it any negative number we wish with the knowledge that different choices merely correspond to a rescaling of x. Choosing $\beta_1 = -2$ leads to $w(x) = \mathrm{e}^{-x^2}$ and $\lambda_n = -2n$, yielding the differential equation defined over $(-\infty,\,\infty)$:

  $$f'' \,-\, 2x\,f' \,+\, 2n\,f \;=\; 0 \qquad \Longleftrightarrow \qquad \left[\mathrm{e}^{-x^2}\,f'(x)\right]' \,+\, 2n\,\mathrm{e}^{-x^2}\,f(x) \;=\; 0 \tag{5.40}$$

- $\alpha = \alpha_1 x + \alpha_0$

  Again, the eigenvalues are $n\beta_1$. We can let $\alpha_1 = 1$ since, if it isn't, we can divide L by α_1, the only effect being a rescaling of all eigenvalues. And we can shift the origin so as to absorb α_0. Then:

  $$w(x)\alpha(x) \;=\; \exp\left[\int^x \frac{\beta(x')}{\alpha(x')}\,\mathrm{d}x'\right] \;=\; \exp\left[\int^x \frac{\beta_1 x' + \beta_0}{x'}\,\mathrm{d}x'\right] \;=\; C\,x^{\beta_0}\,\mathrm{e}^{\beta_1 x}$$

  Provided that $\beta_0 \geq 0$ so as to prevent a divergence at $x = 0$, we can ensure that $w\alpha$ goes to zero at $+\infty$ faster than any finite power of x by taking $\beta_1 = -1$ (any other negative number is equivalent via a rescaling of x). Then we must exclude the interval $(-\infty, 0)$ by demanding that $w(x) = 0$ for $x < 0$. For $x \geq 0$, $w(x) = x^s\,\mathrm{e}^{-x}$, with $s = \beta_0 - 1$. We could choose $\beta_1 > 0$ if the interval is restricted to $(-\infty, 0]$. The eigenvalues are $\lambda_n = -n$, and we arrive at a differential equation defined this time only over $[0,\,\infty)$:

  $$x\,f'' \,-\, \big[x \,-\, s \,-\, 1\big]\,f' \,+\, nf \;=\; 0 \qquad \Longleftrightarrow \qquad \left[x^{s+1}\,\mathrm{e}^{-x}\,f'(x)\right]' \,+\, n\,x^s\,\mathrm{e}^{-x}\,f(x) \;=\; 0 \tag{5.41}$$

In the two cases we have so far considered, the intervals over which the polynomials are defined can extend all the way to infinity. This puts them outside the reach of the Theorem of Weierstrass which asserts the existence of polynomial approximations only over finite intervals. Our third case, though, involves a finite interval.

- $\alpha = \alpha_2 x^2 + \alpha_1 x + \alpha_0$

  First, let $x \to ax + b$ (rescaling of x together with a shift of origin). Then $\alpha = \alpha_2 a^2 x^2 + a(2\alpha_2 b + \alpha_1)x + \alpha_2 b^2 + \alpha_1 b + \alpha_0$. The linear term is eliminated by choosing $b = -\alpha_1/2\alpha_2$, so that $\alpha = \alpha_2 a^2 x^2 + (3/4)\alpha_1^2/\alpha_2 + \alpha_0$. Now we divide L by $-\alpha_2 a^2$, rescaling all the eigenvalues by the same factor, and we adjust a^2 so that $\alpha = -x^2 \pm 1$.

  Our manipulations have also changed β, but it is still a linear polynomial. From eq. (5.36) we have, using *Maple*:

$$w\alpha \;=\; C\exp\left[\int^x \frac{\beta(x')}{\alpha(x')}\,\mathrm{d}x'\right] \;=\; C\exp\left[\int^x \frac{\beta_1\,x' \,+\, \beta_0}{\pm 1 \,-\, x'^2}\,\mathrm{d}x'\right]$$
$$= \;(\sqrt{\alpha})^{-\beta_1}\begin{cases}\exp\,[\beta_0\,\mathrm{arctanh}\,x] & \alpha = 1 - x^2\\[2ex] \exp\,[-\,\beta_0\arctan x] & \alpha = -1 - x^2\end{cases}$$

  Since this time the exponential factor is of no help, and $(\sqrt{\alpha})^{-\beta_1}$ cannot vanish faster than any power of x so as to counteract the bad asymptotic behaviour of the Wronskian in eq. (5.35), we must demand that the interval be finite and that $(\sqrt{\alpha})^{-\beta_1} = 0$ at the end-points, and thus that α have real roots. Therefore, choose $\alpha = 1 - x^2$. But then $\beta_1 < 0$ so as to prevent $w\alpha$ from diverging at $x = \pm 1$. Then $w(x) \neq 0$ on $(-1,\,1)$, and since $w(\pm 1) = 0$, we can match it to another solution $w = 0$ valid outside $[-1,\,1]$.

  With $\alpha = 1 - x^2$, we can ask *Maple* again for w, taking care to specify that $x \in [-1,\,1]$:

$$w(x) \;=\; (1 \,-\, x)^{-(\beta_0+\beta_1)/2-1}\,(1 \,+\, x)^{(\beta_0-\beta_1)/2-1}$$

  It is usual to redefine the exponents so that:

$$w(x) \;=\; (1 \,-\, x)^p\,(1 \,+\, x)^q$$

  where $\beta_0 = q - p$ and $\beta_1 = -(p + q + 2)$. The real principal root is implied if p or q are not integers. Since $\beta_1 < 0$ so as to guarantee proper behaviour of $w\alpha$ as $x \to \pm 1$, then we have: $q > -1$, $p > -1$.

  The eigenvalues are $\lambda_n = n(n-1)\,\alpha_2 \,+\, n\,\beta_1 = -n(n + p + q + 1)$.

  Instead of setting $\alpha_1 = 0$ in L, we could transform x to $1 - 2x$, in which case we would have $\alpha = x(1 - x)$ and $w = x^p(1 - x)^q$ over $[0,\,1]$ and $w = 0$ elsewhere.

5.6.4 Generating formula for Sturm-Liouville polynomials

A general **Rodrigues formula** uniquely determines the Q_n for a given weight function:

$$Q_n(x) \;=\; K_n\,\frac{1}{w(x)}\,\frac{\mathrm{d}^n}{\mathrm{d}x^n}\big[\alpha^n(x)\,w(x)\big] \tag{5.42}$$

where the K_n are normalisation constants which are not determined by the Sturm-Liouville eigenvalue equation since it is linear. They are chosen according to which application we are interested in.

To see how this can be true, we first show that the functions produced by the Rodrigues formula are indeed polynomials of degree n, and then that these are orthogonal with respect to the weight function $w(x)$.

First, given a function $\alpha^k w p_l(x)$, with $p_l(x)$ a polynomial of degree l, we establish that because:

$$\mathrm{d}_x\big(\alpha^k w p_l(x)\big) \;=\; \alpha^{k-1} w\,\big[\beta\,p_l(x) \,+\, (k-1)\,\mathrm{d}_x\alpha\,p_l(x) \,+\, \alpha\,\mathrm{d}_x p_l(x)\big]$$

then, for $k \geq 1$, and since $\beta_1 < 0$ in all cases and $\alpha_2 < 0$ ($\alpha_2 \neq 0$ needed for the last two terms to be of degree $l+1$), there can be no cancellations of powers $l+1$, and the term in square brackets must be a polynomial of degree $l+1$.

This means that if we start with $\alpha^n\, w r_l(x)$, with $l = 0$ and differentiate it $k \leq n$ times, at each step we lose a power of α and raise the degree of r_l by one, so that after the k^{th} differentiation we are left with:

$$\mathrm{d}_x^k(\alpha^n w) \;=\; \alpha^{n-k} w\, r_k(x)$$

After differentiating n times and inserting in eq. (5.42), we get (up to a constant) $r_n(x)$. Then the functions Q_n coming out of the Rodriguez formula have the form $r_n(x)$ and are thus polynomials of degree n.

Also, being aware that $w(x)$ can vanish outside a finite interval, we have, up to a constant:

$$\begin{aligned}\int_{-\infty}^{\infty} Q_m(x)\, Q_n(x)\, w(x)\, \mathrm{d}x \;&=\; \int_{-\infty}^{\infty} Q_m(x)\, \mathrm{d}_x^n(\alpha^n\, w)\, \mathrm{d}x \\ &=\; \sum_{j=0}^{n-1} \left[(-1)^j\, (\mathrm{d}_x^j Q_m)\, \mathrm{d}_x^{n-1-j}(\alpha^n\, w)\right]\Big|_{-\infty}^{\infty} \;+\; (-1)^n \int_{-\infty}^{\infty} (\mathrm{d}_x^n Q_m)\, \alpha^n\, w\, \mathrm{d}x\end{aligned}$$

If $n > m$, $\mathrm{d}_x^n Q_m = 0$ and the integral vanishes; if it isn't, we start with the Rodrigues formula for Q_m instead of Q_n. As for the boundary terms, we know that $\mathrm{d}^{n-1-j}(\alpha^n\, w) = \alpha^{j+1} w\, r_{n-1-j}(x)$. But $w\alpha$ vanishes at the boundaries faster than any power of x, and so, therefore, will this term. This proves that the Rodrigues-generated polynomials are orthogonal. Also, since we have assumed $\alpha(x)$ and $w(x)$ to be real, it is not hard to see from the Rodrigues formula that the Q_n themselves will be real polynomials.

Now, because they form a complete set, the polynomials that satisfy the orthogonality relation for a given weight function and over a given interval are unique, We conclude that the Rodrigues formula provides a (nice and compact!) representation of the polynomial eigenfunctions of *any* Sturm-Liouville operator. All that is needed is w and α.

Here is a list of all the Sturm-Liouville polynomials:

- **Hermite polynomials**:

$$H_n(x) \;=\; K_n\, \frac{1}{w(x)}\, \frac{\mathrm{d}^n}{\mathrm{d}x^n} w(x) \;=\; (-1)^n\, \mathrm{e}^{x^2}\, \mathrm{d}_x^n \mathrm{e}^{-x^2} \tag{5.43}$$

  are the (not normalised to 1) polynomial eigenfunctions corresponding to the Sturm-Liouville equation over $(-\infty,\, \infty)$:

$$H'' \;-\; 2x\, H' \;+\; 2n\, H \;=\; 0$$

  Since $\alpha_1 = \beta_0 = 0$, the H_n have parity $(-1)^n$. They satisfy the relation:

$$\int_{-\infty}^{\infty} H_m(x)\, H_n(x)\, \mathrm{e}^{-x^2}\, \mathrm{d}x \;=\; 2^n\, n!\, \sqrt{\pi}\, \delta_{mn}$$

  The functions $H_n(x)\mathrm{e}^{-x^2/2}/[2^n\, n!\, \sqrt{\pi}]^{1/2}$ are the basis wave-functions for the 1-dim quantum harmonic oscillator.

- The **associated Laguerre polynomials** (or just Laguerre polynomials when $s = 0$):

$$L_m^s(x) \;=\; K_m\, \frac{1}{w(x)}\, \mathrm{d}_x^m\big(x^m\, w(x)\big) \;=\; x^{-s}\, \mathrm{e}^x\, \mathrm{d}_x^m\big(x^{m+s}\, \mathrm{e}^{-x}\big) \qquad x \in [0, \infty) \tag{5.44}$$

  are the (not normalised to 1) polynomial eigenfunctions of degree m corresponding to the Sturm-Liouville equation over $[0,\, \infty)$:

$$x\, L_m^{s\,\prime\prime} \;-\; \big[x \;-\; s \;-\; 1\big]\, L_m^{s\,\prime} \;+\; m L_m^s \;=\; 0$$

Since $\alpha_1 \neq 0$, the $L_m^s(x)$ do not have definite parity. The Laguerre polynomials ($s = 0$) satisfy:

$$\int_0^\infty L_m^0(x)\, L_n^0(x)\, \mathrm{e}^{-x}\, \mathrm{d}x \;=\; m!\, \delta_{mn}$$

The functions $x^{2l} L_{n+l}^{2l+1}(x)\mathrm{e}^{-x/2}$, where n is the energy quantum number and l the orbital quantum number, occur in the radial wave-functions which solve the Schrödinger equation for the Coulomb potential. Showing the equivalence of the associated radial Schrödinger equation and of the above forms does require some non-trivial manipulation (see your favourite quantum mechanics textbook for details).

- **Jacobi polynomials**:

$$J_n^{p,q}(x) \;=\; K_n^{p,q}\, \frac{1}{w(x)}\, \mathrm{d}_x^n(\alpha^n\, w) \;=\; (1-x)^{-p}\,(1+x)^{-q}\, \mathrm{d}_x^n\big[(1-x^2)^n\,(1-x)^p\,(1+x)^q\big] \tag{5.45}$$

are the most general (not normalised to 1) polynomial eigenfunctions corresponding to the Sturm-Liouville eigenvalue equation over $[-1,\, 1]$:

$$(1-x^2)\, J_n'' \;+\; \big[(q\;-\;p)\;-\;(2\;+\;p\;+\;q)\, x\big]\, J_n' \;+\; n\,(1\;+\;p\;+\;q\;+\;n)\, J_n \;=\; 0$$

The most interesting subclass of the **Jacobi polynomials**, $J_n^{p,q}$, are the **Gegenbauer polynomials**, sometimes called **ultraspherical**, for which $p \;=\; q \;=\; m \;>\; -1$. They all have well-defined parity. Their Rodrigues generating formula is the more tractable:

$$G_n^m(x) \;=\; K_n^m\, \frac{1}{w(x)}\, \mathrm{d}_x^n\big[(1-x^2)^n\, w)\big] \;=\; (1-x^2)^{-m}\, \mathrm{d}_x^n\big[(1-x^2)^{n+m}\big] \tag{5.46}$$

Two subclasses of the Gegenbauer polynomials are more famous and useful:

  - The **Chebyshev** (Tchebycheff, Tschebyscheff) polynomials which have $m = \pm 1/2$ and satisfy either:

$$(1-x^2)\, f'' \;-\; x\, f' \;+\; n^2\, f \;=\; 0 \qquad (m = -1/2)$$

    or

$$(1-x^2)\, f'' \;-\; 3x\, f' \;+\; n\,(n\;+\;2)\, f \;=\; 0 \qquad (m = 1/2)$$

    With $m = -1/2$ the Chebyshev polynomials of the first kind are generated by:

$$T_n(x) \;=\; K_n\, \sqrt{(1-x^2)}\, \mathrm{d}_x^n\big[(1-x^2)^{n-1/2}\big] \tag{5.47}$$

    With the change of variable $x = \cos\theta$, the equation for the T_n can be rewritten as $\mathrm{d}_\theta^2 T_n + n^2 T_n = 0$. The solutions in complex form are:

$$\mathrm{e}^{\mathrm{i}n\theta} \;=\; \cos n\theta \;+\; \mathrm{i}\, \sin n\theta \;=\; (\cos\theta \;+\; \mathrm{i}\, \sin\theta)^n$$

    Expressing the right-hand side as its binomial expansion and taking the real part, there comes:

$$\cos n\theta \;=\; \sum_{\substack{k=0 \\ k\ \text{even}}}^{n} (-1)^{-k/2}\, \tbinom{n}{k}\, \cos^{n-k}\theta\, (1\;-\;\cos^2\theta)^{k/2}$$

    By inspection, this is a polynomial in $\cos\theta$. The imaginary part $\sin n\theta$, however, is a polynomial multiplied by $\sin\theta$, which is not a polynomial in $\cos\theta$, and therefore cannot be a Chebyshev polynomial of the first kind. Its polynomial part, $(\sin n\theta)/\sin\theta$, is known as a Chebyshev polynomial of the second kind.

    So we can now assert that $T_n(\cos\theta) = \cos\, n\theta$. This can be used, eg., to express $\cos n\theta$ in terms of powers of $\cos\theta$. More important, Chebyshev expansions are related to Fourier series with which they share most properties, except that they can converge faster.

– The restriction $p = q = 0$, or $\beta_1 = -2$, $\beta_0 = 0$, leads to $w(x) = 1$, and the eigenvalues are $\lambda_l = -l(l+1)$. We obtain the differential equation:

$$(1-x^2)\,P_l'' - 2x\;P_l' + l(l+1)\,P_l \;=\; 0 \qquad \Longleftrightarrow \qquad \big[(1-x^2)\,P_l'(x)\big]' + l(l+1)\,P_l(x) \;=\; 0 \tag{5.48}$$

with the **Legendre polynomials** as (*not normalised to 1*) polynomial solutions:

$$P_l(x) \;=\; K_l\,\frac{1}{w(x)}\,\frac{\mathrm{d}^l}{\mathrm{d}x^l}\big[(1-x^2)^l\,w(x)\big] \;=\; \frac{(-1)^l}{2^l\,l!}\,\mathrm{d}_x^l(1-x^2)^l \qquad x \in [-1,\,1] \tag{5.49}$$

The $P_l(x)$ have parity $(-1)^l$ and satisfy:

$$\int_{-1}^{1} P_{l'}(x)\,P_l(x)\,\mathrm{d}x \;=\; \frac{2}{2l+1}\,\delta_{l'l}$$

Quite often, the parametrisation $x = \cos\theta$ is used. Legendre polynomials provide a very useful expansion of the function $1/|\mathbf{x}-\mathbf{x}'|$; therefore they allow potentials with a $1/r$ dependence to be expressed in terms of a so-called **multipole expansion**. They also arise as the angular part of the solution of the Laplace equation, $\nabla^2 f = 0$, in spherical coordinates, in cases of azimuthal symmetry, ie. when the solution is independent of the azimuthal angle that rotates around a suitably chosen z-axis.

All the Sturm-Liouville polynomials satisfy recurrence (or recursion) relations found in many references (see course webpage). There are general ones that cover all the polynomials, but they are complicated. Often, it is easiest to derive them directly.

Example 5.4. Using the trigonometric representation of the Chebyshev polynomials, $T_n(x = \cos\theta) = \cos\,n\theta$, one almost immediately obtains (EXERCISE):

$$T_{n+1}(x) \;=\; 2x\,T_n(x) \;-\; T_{n-1}(x)$$

Example 5.5. We also derive (EXERCISE) the relations:

$$H_n' \;=\; 2n\,H_{n-1}, \qquad\qquad H_{n+1} \;=\; 2x\,H_n \;-\; 2n\,H_{n-1}$$

From these examples one can see that recurrence relations can provide a quick way of generating classical orthogonal polynomials. They also come in very handy in some integrals.

Although this discussion exhausts all possibilities for complete orthonormal sets of *polynomials* as solution of a Sturm-Liouville problem, they are not necessarily the only complete orthonormal *functions* on the same interval.

5.7 Associated Legendre Functions and Spherical Harmonics

To discover a complete set of orthonormal functions on the unit sphere in three dimensions, we appeal again to the Theorem of Weierstrass. This asserts the existence of partial sums:

$$g_M(\theta,\phi) \;=\; \sum_{\alpha,\beta,\gamma=0}^{M} A^{(M)}_{\alpha\beta\gamma}\,u^\alpha\,v^\beta\,w^\gamma \;=\; \sum_{l=0}^{3M}\;\sum_{\substack{\alpha,\beta,\gamma=0\\ \alpha+\beta+\gamma=l}}^{M} A^{(M)}_{\alpha\beta\gamma}\,\mathrm{e}^{\mathrm{i}m\phi}\,\sin^{\alpha+\beta-|m|}\theta\,\sin^{|m|}\theta\,\cos^\gamma\theta \qquad\qquad l>0$$

such that the $g_M(\theta,\phi)$ converge uniformly to the continuous function $g(\theta,\phi)$ in the limit $M \to \infty$. We have written $u = x + \mathrm{i}y = \sin\theta\,\mathrm{e}^{\mathrm{i}\phi}$, $v = x - \mathrm{i}y = \sin\theta\,\mathrm{e}^{-\mathrm{i}\phi}$, $w = z = \cos\theta$, and $m = \alpha - \beta$. Note that the last

term on the right is merely a rearrangement of the summations. Note also that the second sum is restricted to $\alpha + \beta + \gamma = l$.

The powers in the summand can now be written in terms of l and m, so that $g_M(\theta, \phi)$ becomes:

$$g_M(\theta, \phi) = \sum_{l=0}^{3M} \sum_{m=-l}^{l} B_{lm}^{(M)} Y_{lm}(\theta, \phi)$$

where: $Y_{lm}(\theta, \phi) = \mathrm{e}^{\mathrm{i}m\phi} \sin^{|m|}\theta\, f_{lm}(\cos\theta)$, and: $f_{lm}(\cos\theta) \equiv (1 - \cos^2\theta)^{(l-|m|-\gamma)/2} \cos^\gamma\theta$ is a polynomial in powers of $\cos\theta$, with maximum degree $l - |m|$, because $\alpha + \beta - |m| \geq 0$ is even.

Then it is enough to find $f_{lm}(\cos\theta)$ such that:

$$\int_{\phi=0}^{2\pi} \int_{\theta=0}^{\pi} Y_{l'm'}^*(\theta, \phi)\, Y_{lm}(\theta, \phi)\, \sin\theta\, \mathrm{d}\theta\, \mathrm{d}\phi = \delta_{l'l}\, \delta_{m'm}$$

Following the same kind of reasoning as with the Fourier expansion, this will guarantee that the Y_{lm} form a complete set.

The ϕ dependence of Y_{lm} already guarantees part of the orthogonality. The following functions do satisfy (not proved) the full orthonormality relation:

$$Y_{lm}(\theta, \phi) = (-1)^m \sqrt{\frac{2l+1}{4\pi} \frac{(l-m)!}{(l+m)!}}\, P_l^m(x)\, \mathrm{e}^{\mathrm{i}m\phi} \qquad (x = \cos\theta,\ 0 \leq m \leq l) \tag{5.50}$$

where the unnormalised **associated Legendre functions** of order m, $P_l^m(x)$, are given in terms of the Legendre polynomials $P_l(x) = P_l^0(x)$ by:

$$\begin{aligned} P_l^m(x) &= \frac{(-1)^l}{2^l l!} (1 - x^2)^{m/2} \mathrm{d}_x^{l+m} (1 - x^2)^l \\ &= (1 - x^2)^{m/2} \mathrm{d}_x^m P_l(x) \qquad (0 \leq m \leq l,\ -1 \leq x \leq 1) \end{aligned} \tag{5.51}$$

The spherical harmonics also exist for $m < 0$ and are then defined as:

$$Y_{l,-m} = (-1)^m\, Y_{lm}^*$$

The $\sqrt{(l + 1/2)(l - m)!/(l + m)!} P_l^m(x)$ are themselves orthonormal functions on $[-1, 1]$, just as the $\sqrt{l + 1/2} P_l(x)$, *but they are not polynomials* and thus do not contradict the uniqueness of the Legendre polynomials. They too satisfy a Sturm-Liouville equation. Differentiating the Legendre equation (5.48) m times and redefining $\mathrm{d}_x^m P_l(x)$ with eq. (5.51), we obtain:

$$\mathrm{d}_x \left[(1 - x^2)\, \mathrm{d}_x P_l^m(x)\right] + \left[l(l + 1) - \frac{m^2}{1 - x^2}\right] P_l^m(x) = 0 \tag{5.52}$$

The $P_l^m(x)$ are the only solutions of this equation which are regular at $x = \pm 1$.

The completeness relation satisfied by the spherical harmonics reads:

$$\sum_{l=0}^{\infty} \sum_{m=-l}^{l} Y_{lm}^*(\theta', \phi') Y_{lm}(\theta, \phi) = \delta(x - x')\delta(\phi - \phi') \qquad (x = \cos\theta) \tag{5.53}$$

The spherical harmonics satisfy other useful relations which can be found in several references (Arfken's *Mathematical Methods for Physicists* and *Handbook of Mathematical Functions* by Abramowitz and Stegun are two popular ones). They occur in the solution to many problems in physics: in quantum mechanics for instance, when the potential in the Schrödinger equation is spherically-symmetric, the angular dependence of the wave functions is always given by spherical harmonics. They are also extremely useful in electrostatics.

6 MODULE VI — Solution of Differential Equations with Green Functions

Physical quantities are generally represented by functions of up to four (three spatial and one time) variables and therefore satisfy partial differential equations (PDE). More precisely, let $y(x^1, \ldots, x^n)$ be a variable **dependent** on the **independent** variables $x^1, \ldots, x^n$, then y may have to satisfy equations of the form:

$$f\left(y, \frac{\partial y}{\partial x^i}, \ldots, \frac{\partial^m y}{\partial x^i \partial x^j \ldots}, x^i\right) = 0 \tag{6.1}$$

where $0 \leq i, j, \ldots \leq m$, with the constraint: $i + j + \ldots = m$.

If this equation can be split into:

$$g\left(y, \frac{\partial y}{\partial x^i}, \ldots, \frac{\partial^m y}{\partial x^i \partial x^j \ldots}, x^i\right) = F(x^i)$$

it is said to be **inhomogeneous**. If $F(x^i) = 0$, $g = 0$ is said to be a **homogeneous** equation.

You may be relieved to know that in physics we almost never have to go beyond $m = 2$. Still, PDEs can be extremely challenging, and most have to be solved numerically. Very thick books have been written on techniques for numerically solving PDEs, and we will not even attempt to broach the topic. In some important cases, PDEs in n independent variables can be converted into n ordinary differential equations (ODE).via the technique of **separation of variables**. To test whether a PDE has separable solutions, insert $y(x^1, \ldots, x^n) = X_1(x^1)X_2(x^2)\ldots X_n(x^n)$ into it, and see if it can be written as a sum of terms, each of which depends on one x^i only. If that happens, the PDE can be satisfied only if each term is equal to a constant, called the **separation constant**, with all the constants summing to zero. Then we are left with n ODEs, one for each $X_i(x^i)$. If the solution to each of these ODEs is unique, this solution to the PDE will also be unique.

In the next few sections, we shall discuss ODEs of first and second order, returning to PDEs later.

6.1 Ordinary Differential Equations

An ODE of order n for $y(x)$ is said to be **linear** if $g(y, y', y'', \ldots)$ is linear in y and all its derivatives occurring in the equation. Then, with the notation $L[y] \equiv g(y, y', y'', \ldots)$, linearity means that $L[af_1+bf_2] = aL[f_1]+bL[f_2]$, where f_1 and f_2 solve the ODE; a and b are constants. We see immediately that adding the sum of the general solution of $L[y] = 0$ and a solution to $L[y] = F$ also solves the latter and is its general solution.

Many important natural phenomena are described by **non-linear** differential equations, but there exists no general method for solving them, only tricks that may work in a limited number of cases. One such trick is to try a clever redefinition. Take for instance the first-order Bernoulli equation:

$$\mathrm{d}_x y + \beta(x)\, y = y^n\, F(x)$$

which is readily rewritten in the inhomogeneous form:

$$y^{-n}\, \mathrm{d}_x y + \beta(x)\, y^{1-n} = F(x)$$

If we notice that $(1-n)y^{-n}\, \mathrm{d}_x y = \mathrm{d}_x(y^{1-n})$ then we should redefine $u = y^{1-n}$ to convert the ODE into the linear equation

$$\mathrm{d}_x u + \big[(1-n)\,\beta(x)\big]\, u = (1-n)\, F(x)$$

A whole lore also exists on finding so-called **integrating factors**, but again this is very much a hit-and-miss affair, and we shall (with some regret) leave the study of non-linear ODEs for another time.

6.1.1 First-order ODEs

The most general form for a first-order ODE is:

$$y' \;=\; g\big(x, y(x)\big) \tag{6.2}$$

When supplemented by an **initial condition**, such as $y(x_0) = y_0$, this becomes an **initial-value problem (IVP)**. Provided that g is continuous in x, a theorem by Peano guarantees the existence of a solution, at least over some interval around x_0. But such a solution may not be unique! Indeed, let $g\big(x, y(x)\big) = y^n$ for $0 < n < 1$ when $x < 0$ and zero otherwise, with $y(0) = 0$. Then $y(x) = \pm[(1-n)x]^{1/(1-n)}$ for $x < 0$ and zero otherwise both solve the IVP (EXERCISE), and so does $y(x) = 0\;\forall\, x$, all with the *same* initial condition, so that the solution is not unique.

If, however, one also requires continuity of $\mathrm{d}_y g$ in y (which was not true at $x = 0$ in our example), the Picard-Lindelöf theorem guarantees unicity. Briefly, to derive this solution one first rewrites eq. (6.2) in the form:

$$y(x) \;=\; y_0 \,+\, \int_{x_0}^{x} g\big(x, y(x')\big)\,\mathrm{d}x'$$

Then, starting with $y_0 = 0$, one constructs the sequence of functions:

$$y_{n+1} \;=\; y_0 \,+\, \int_{x_0}^{x} g\big(x, y_n(x')\big)\,\mathrm{d}x'$$

This iterative process is called **Picard's iteration method**. Then one shows that the sequence is Cauchy (see module V) to prove that it converges toward $y(x)$.

On the other hand, it is not difficult to obtain an *explicit* general solution to a first-order *linear* ODE. First, it is convenient to recast it into its **normal form**:

$$\mathrm{d}_x f \,+\, \beta(x)\, f \;=\; F(x)$$

Then one shows (EXERCISE) that if $\beta(x)$ is continuous over an interval $(a,\, b)$, the general solution of this ODE is, $\forall\, x_0 \in (a,\, b)$:

$$f(x) \;=\; \frac{i(x_0)}{i(x)}\, f(x_0) \,+\, \int_{x_0}^{x} F(x')\,\frac{i(x')}{i(x)}\,\mathrm{d}x' \qquad\qquad i(x) \;=\; \mathrm{e}^{\int^x \beta(x')\,\mathrm{d}x'} \tag{6.3}$$

where $f(x_0)$ is arbitrary.

6.1.2 Second-order linear ODEs

Consider first the most general form for a homogeneous second-order equation:

$$L[f] \;=\; \alpha(x)\,\mathrm{d}_x^2 f \,+\, \beta(x)\,\mathrm{d}_x f \,+\, \gamma(x)\, f \;\;=\; 0 \tag{6.4}$$

Note that since γ is known, this is not an eigenvalue equation for L.

It can be shown (the proof is technical and not very illuminating) that in a finite interval $[a,\, b]$, the only solution for which f and $\mathrm{d}_x f$ *both* vanish at the initial point $x = a$ is the trivial solution $f = 0$. Consequently, if there exist two solutions f and g that satisfy the initial conditions $f(a) - g(a) = 0$ and $f'(a) - g'(a) = 0$, then $f = g$ over the whole interval. In other words, the two initial conditions uniquely specify the solution to a homogeneous second-order differential equation.

Let $f_1(x)$ and $f_2(x)$ be two solutions of eq. (6.4). If there exists no constant C such that $f_2 = C f_1\;\forall\, x$, f_1 and f_2 are linearly independent and any linear combination of them is also a solution of the homogeneous equation (principle of linear superposition). One cannot have both $f_1 = f_2$ and $f_1' = f_2'$ at any point.

The Wronskian of two functions $f_1(x)$ and $f_2(x)$ is defined as: $W(x) \coloneqq f_1 f_2' - f_2 f_1'$. It is easy to obtain a first-order differential equation for $W(x)$,whose solution (EXERCISE) is called **Abel's formula**:

$$W(x) \;=\; W(x_0)\,\mathrm{e}^{-\int_{x_0}^{x}[\beta(x')/\alpha(x')]\mathrm{d}x'} \tag{6.5}$$

where x_0 is any point in the interval $[a, b]$. This form of the Wronskian makes no direct reference to homogeneous solutions of eq. (6.4). We see that eq. (6.4) with $\beta(x) = 0$ leads to constant Wronskians. We also see, because the exponential cannot vanish in a finite interval, that if the Wronskian vanishes anywhere, it vanishes everywhere.

And now comes a surprising fact: given *one* solution of the *homogeneous* equation, a general solution of the *inhomogeneous* equation valid over $[a, b]$:

$$\alpha(x)\,\mathrm{d}_x^2 f \;+\; \beta(x)\,\mathrm{d}_x f \;+\; \gamma(x)\,f \;=\; F(x)$$

can be generated. We give a sketch of the procedure, also known as the method of **variation of parameters**:

1. Given f_1, we seek a second linearly independent solution of the homogeneous equation. Noticing that:

$$\frac{W(x)}{f_1^2(x)} \;=\; \left(\frac{f_2}{f_1}\right)'$$

and integrating, we find with eq. (6.5) that:

$$f_2(x) \;=\; f_1(x)\int_a^x \frac{W(x_0)}{f_1^2(x')}\,\mathrm{e}^{-\int^{x'}(\beta/\alpha)\mathrm{d}x''}\,\mathrm{d}x' \tag{6.6}$$

is also a solution. If the result still contains a part that is proportional to f_1, discarding it leaves a solution that is linearly independent from f_1. Note that here $W(x_0)$ is an *arbitrary* non-zero constant.

2. Having generated a second solution to the homogeneous equation, let us obtain a *particular* solution $h(x)$ to the inhomogeneous equation. The key step is to insert $h(x) = f_1(x)g(x)$ to obtain (EXERCISE) a first-order equation for g'. Using the general first-order equation solution eq. (6.3) and (6.5) again, there comes:

$$g'(x) \;=\; \left(\frac{f_2}{f_1}\right)' \int_a^x \frac{f_1(x')\,F(x')}{\alpha(x')\,W(x')}\,\mathrm{d}x'$$

A straightforward integration then leads to:

$$h(x) \;=\; f_1\,g \;=\; f_2(x)\int_a^x \frac{f_1(x')\,F(x')}{\alpha(x')\,W(x')}\,\mathrm{d}x' \;-\; f_1(x)\int_a^x \frac{f_2(x')\,F(x')}{\alpha(x')\,W(x')}\,\mathrm{d}x' \tag{6.7}$$

The general solution of the inhomogeneous equation is then a linear combination of f_1 and f_2 (homogeneous solution) plus this particular solution. The coefficients of the linear combination are fixed by boundary (2-point) or initial (1-point) conditions which we shall soon discuss.

Let us note also that provided $\int \beta(x)/\alpha(x)\mathrm{d}x$ exists within the interval of interest, it is always possible to eliminate the first-order derivative term in any linear second-order ODE, by a redefinition of the form $f(x) = g(x)\mathrm{e}^{\mu(x)}$ (the substitution $f(x) = \mu(x)g(x)$ also works), to arrive at (EXERCISE):

$$\alpha(x)\,g''(x) \;+\; \left(\gamma(x) \;-\; \frac{\beta'}{2} \;+\; \frac{\beta}{2}\frac{\alpha'}{\alpha} \;-\; \frac{1}{4}\frac{\beta^2}{\alpha}\right) g(x) \;=\; F(x)\exp\left[\int^x \frac{\beta}{2\alpha}\mathrm{d}x'\right] \tag{6.8}$$

as determined by the requirement that the transformed ODE have no first-order derivative in g.

One would do well to remember, however, that making an ODE look simpler does not always make it easier to solve!

6.2 Solving One-dimensional Equations with Green's Functions (BF 7.3)

6.2.1 Solutions in terms of Green's Functions and Boundary Conditions

We focus on inhomogeneous second-order linear equations of the type:

$$L\big[f(t)\big] \;=\; \alpha(t)\,\mathrm{d}_t^2 f(t) \;+\; \beta(t)\,\mathrm{d}_t f(t) \;+\; \gamma(t)\,f(t) \;=\; F(t)$$

with α, β and γ continuous all over the interval of interest. Such ordinary (one-dimensional) differential equations often have time as the independent variable, and in physics $F(t)$ is then called a **driving** term. Of course, its solutions will be twice (and most often more than twice) differentiable.

Obtaining a very useful expression for a solution to this equation is not so difficult if L admits a complete set of orthonormal eigenfunctions ϕ_j on the interval (which is the case if L is self-adjoint), with real eigenvalues $\lambda_j \neq 0$. Then f can be expanded over the set (with unknown coefficients a_j), and so can F, with known coefficients b_j. Both sets of coefficients are, as usual, projections of f and F on the eigenfunctions. The eigenvalue equation then yields a relation between them, and there comes (EXERCISE), assuming that integral and summation signs can be interchanged) the inhomogeneous solution:

$$f(t) \;=\; \int \left[\sum_j \frac{\phi_j(t)\,\phi_j^*(t')}{\lambda_j}\right] F(t')\,\mathrm{d}t' \;=\; \int G(t,t')\,F(t')\,\mathrm{d}t'$$

where we have identified a two-point function:

$$G(t,t') \;=\; \sum_j \frac{\phi_j(t)\,\phi_j^*(t')}{\lambda_j} \qquad\qquad \lambda_j \neq 0 \tag{6.9}$$

Acting on $G(t,t')$ with L yields an associated differential equation:

$$L\big[G(t,t')\big] \;=\; \sum_j \frac{L\big[\phi_j(t)\big]\,\phi_j^*(t')}{\lambda_j} \;=\; \sum_j \phi_j(t)\,\phi_j^*(t') \;=\; \delta(t-t')$$

where the last equality comes from the completeness relation for the eigenfunctions. The defining equation for a **Green function** $G(t,t')$ is:

$$L\big[G(t,t')\big] \;=\; \delta(t-t') \tag{6.10}$$

Suppose that we know two linearly independent solutions, $f_1(t)$ and $f_2(t)$, to the associated homogeneous equation. Then, *if boundary conditions allow the homogeneous solution to exist*, the general solution is the sum of the general solution to the homogeneous equation plus a solution to the inhomogeneous equation.

In 1-dim problems, boundary conditions are generally of two types:

(1) **One-Point (Initial) conditions, aka IVP, or Initial-Value Problem:** f and its first-order derivative are known at some time, usually called the initial time t_0. Then the general solution of our inhomogeneous equation can be written:

$$f(t) \;=\; A\,f_1(t) \;+\; B\,f_2(t) \;+\; \int_{t_0}^{t} G_{\mathrm{ivp}}(t,t')\,F(t')\,\mathrm{d}t' \tag{6.11}$$

where A and B are two constants to be determined from the initial conditions, and the Green function $G_{\mathrm{ivp}}(t,t')$ associated with a given differential operator L corresponds to this *specific type of boundary conditions*. We see that: $f(t_0) \;=\; A\,f_1(t_0) \;+\; B\,f_2(t_0)$, and $\dot{f}(t_0) \;=\; \big[A\,\dot{f}_1(t) \;+\; B\,\dot{f}_2(t)\big]_{t_0}$, so that A and B do not depend on the inhomogeneous term $F(t)$. With the (always non-zero) Wronskian of f_1 and f_2: $W \equiv f_1\dot{f}_2 - \dot{f}_1 f_2$, they are found to be:

$$A \;=\; \frac{\dot{f}_2(t_0)\,f(t_0) \;-\; f_2(t_0)\,\dot{f}(t_0)}{W(t_0)}, \qquad\qquad B \;=\; -\,\frac{\dot{f}_1(t_0)\,f(t_0) \;-\; f_1(t_0)\,\dot{f}(t_0)}{W(t_0)}$$

In time-evolution problems with initial conditions, we most often demand that $G_{\mathrm{ivp}}(t,t') = 0$ for $' > t$ so as to preserve causality. This allows us, if we wish, to extend the upper bound of the integral all the way to infinity.

(2) **Two-point boundary conditions, or Boundary-Value Problem:** In one example of this case, $f(t_0)$ and $f(t_1)$ ($t_1 > t_0$) are known (the case for which the derivatives of f are known will be addressed in section 6.2.3). Then the general solution would be written:

$$f(t) \;=\; A\, f_1(t) \;+\; B\, f_2(t) \;+\; \int_{t_0}^{t_1} G_{\rm bvp}(t,t')\, F(t')\, {\rm d}t' \tag{6.12}$$

Note the different upper limit of integration, with the integral still a function of t. This time, A and B do not depend on the inhomogeneous term if the appropriate Green function always obeys the homogeneous boundary conditions: $G_{\rm bvp}(t_0,t') = G_{\rm bvp}(t_1,t') = 0$; we shall prove that this is indeed the case below in section 6.2.3. If such a $G_{\rm bvp}$ is to exist, we see from eq. (6.9) that L cannot have a zero eigenvalue; in other words, *there can be no non-trivial solution of the homogenous equation $L[f] = 0$ that satisfies homogeneous B.C.*

Moreover, symbolically writing the inhomogeneous solution as $f(t) = L_t^{-1}F(t) = \int G(t,t')F(t'){\rm d}t'$, we see that knowledge of the Green function amounts to inverting the inhomogeneous equation, and gives us an inhomogeneous solution *in the form of an integral*. Our interest in Green functions entirely resides in this property.

6.2.2 A General Expression for 1-d Green Functions for $L[f(t)] = F(t)$

What restrictions does eq. (6.10) impose on $G(t,t')$? Three, in fact:

(a) $G(t,t')$ is a continuous function of t everywhere, including at $t = t'$, otherwise its second derivative at $t = t'$ would be the derivative of a δ-function, and the differential equation would not be satisfied. Note, however, that the Green function for a *first-order* operator can be discontinuous, eg., $L = -{\rm id}_t$ has as Green function the step-function ${\rm i}\theta(t-t')$.

(b) $\dot G$ must have a discontinuity at $t = t'$. To see this, integrate eq. (6.10) from $t = t' - \epsilon$ to $t = t' + \epsilon$. Since the coefficients in L are continuous, they hardly vary when the interval is taken to be arbitrarily small ($\epsilon \to 0$). In that limit, the integrals of G and $\dot G$ both vanish because G is continuous, and we are left with the contribution of the integral of $\ddot G$:

$$\lim_{\epsilon\to 0} \dot G(t,t')\Big|_{t=t'-\epsilon}^{t=t'+\epsilon} \;=\; \frac{1}{\alpha(t')}$$

So, when the coefficient of the second-order derivative in L is 1, the derivative of G must jump by 1 at $t = t'$.

(c) When $t \neq t'$, G must satisfy the homogeneous equation $L\big[G(t,t')\big] = 0$, and can be written in terms of f_1 and f_2:

$$G(t,t') \;=\; \begin{cases} a_1(t')\, f_1(t) \;+\; a_2(t')\, f_2(t) & t > t' \\ b_1(t')\, f_1(t) \;+\; b_2(t')\, f_2(t) & t < t' \end{cases}$$

The continuity of G and the expression for the discontinuity in $\dot G$ at $t = t'$ then yield:

$$\begin{aligned} a_1\, f_1(t') \;+\; a_2\, f_2(t') \;-\; b_1\, f_1(t') \;-\; b_2\, f_2(t') \;&=\; 0 \\ a_1\, \dot f_1(t') \;+\; a_2\, \dot f_2(t') \;-\; b_1\, \dot f_1(t') \;-\; b_2\, \dot f_2(t') \;&=\; \frac{1}{\alpha(t')} \end{aligned}$$

These can be solved for a_1 and a_2 in terms of b_1 and b_2:

$$\begin{aligned} a_1(t') \;&=\; b_1(t') \;-\; \frac{f_2(t')}{\alpha(t')\, W(t')} \\ a_2(t') \;&=\; b_2(t') \;+\; \frac{f_1(t')}{\alpha(t')\, W(t')} \end{aligned}$$

where the Wronskian of f_1 and f_2 is $W \equiv f_1\dot{f}_2 - \dot{f}_1 f_2 \neq 0$ since f_1 and f_2 are linearly independent. Then the Green function for L must take the form:

$$G(t,t') \;=\; \begin{cases} b_1(t')\,f_1(t) \;+\; b_2(t')\,f_2(t) \;-\; \dfrac{f_1(t)\,f_2(t') \;-\; f_2(t)\,f_1(t')}{\alpha(t')\,W(t')} & t > t' \\ b_1(t')\,f_1(t) \;+\; b_2(t')\,f_2(t) & t < t' \end{cases} \tag{6.13}$$

The term with the Wronskian vanishes at $t = t'$, guaranteeing the continuity of G as required.

The adjustable parameters b_1 and b_2 can now be chosen so that G satisfies suitable boundary conditions. In the case (IVP) where initial conditions are specified on $f(t)$ and $\dot{f}(t)$, requiring $G(t,t') = 0$ for $t < t'$ immediately leads to $b_1 = b_2 = 0$, and there comes the simple but very general expression:

$$G_{\rm ivp}(t,t') \;=\; \theta(t-t')\,\frac{f_2(t)f_1(t') \;-\; f_1(t)\,f_2(t')}{\alpha(t')W(t')} \tag{6.14}$$

where the **step-function**, or **Heaviside function**, $\theta(u)$ is defined* by:

$$\theta(u) \;=\; \begin{cases} 1 & u > 0 \\ 0 & u \leq 0 \end{cases} \tag{6.15}$$

If the initial conditions are homogeneous, ie., if $f(0) = \dot{f}\big|_0 = 0$, there is no non-trivial homogeneous solution for f, and $\int G_{\rm ivp}(t,t')F(t')\mathrm{d}t'$ in eq. (6.11), which can always be evaluated, either analytically or numerically, is the general solution to the inhomogeneous equation $L[f] = F$. Then $G_{\rm ivp}$ can be viewed as the inverse of L.

On the other hand, for two-point boundary conditions (BVP) at t_0 and t_1, $G(t_0,t') = 0$ ($t_0 < t'$) immediately leads to: $b_2(t') \;=\; -\,b_1(t')f_1(t_0)/f_2(t_0)$, whereas $G(t_1,t') = 0$ ($t_1 > t'$) gives:

$$b_1(t') = \frac{f_2(t_0)}{\alpha(t')\,W(t')}\,\frac{f_1(t_1)\,f_2(t') \;-\; f_2(t_1)\,f_1(t')}{f_1(t_1)\,f_2(t_0) \;-\; f_2(t_1)\,f_1(t_0)} \quad\Longrightarrow\quad b_2(t') = \frac{f_1(t_0)}{\alpha(t')\,W(t')}\,\frac{f_2(t_1)\,f_1(t') \;-\; f_1(t_1)\,f_2(t')}{f_1(t_1)\,f_2(t_0) \;-\; f_2(t_1)\,f_1(t_0)}$$

The BVP Green function factorises (EXERCISE) in t and t':

$$G_{\rm bvp}(t,t') \;=\; \frac{1}{\alpha(t')\,W(t')}\,\frac{\big[f_1(t_1)\,f_2(t_>) \;-\; f_2(t_1)\,f_1(t_>)\big]\big[f_1(t_<)\,f_2(t_0) \;-\; f_1(t_0)\,f_2(t_<)\big]}{f_1(t_1)\,f_2(t_0) \;-\; f_2(t_1)\,f_1(t_0)}$$

where $t_> \coloneqq \max(t,t')$ and $t_< \coloneqq \min(t,t')$. If $f_1(t_0) = 0$, then:

$$G_{\rm bvp}(t,t') \;=\; \frac{1}{\alpha(t')\,W(t')}\left[f_2(t_>)f_1(t_<) \;-\; \frac{f_2(t_1)}{f_1(t_1)}\,f_1(t_>)f_1(t_<)\right]$$

The most simple case occurs when $f_1(t_0) = f_2(t_1) = 0$; then: $G_{\rm bvp}(t,t') = f_1(t_<)f_2(t_>)/\alpha(t')W(t')$.

Note that, unlike for the IVP problem, a Green function for the one-dim BVP for a given L exists only if b_1 and b_2 are finite, ie. if $f_1(t_1)\,f_2(t_0) \;-\; f_2(t_1)\,f_1(t_0) \neq 0$. Also, if the boundary conditions allow the existence of a homogeneous solution, ie. an eigenfunction of L with eigenvalue zero satisfying these same conditions, eq. (6.9) forbids a Green function.

Example 6.1. A Simple Example: the Harmonic Oscillator

Consider the undamped oscillator operator $\mathrm{d}_t^2 + \omega_0^2$ with initial conditions on f and $\dot{f}$ at a single point (IVP). We choose $f_1 = \sin\omega_0 t$ and $f_2 = \cos\omega_0 t$. Also, noting that $\alpha = 1$ and $W = -\omega_0$, eq. (6.13) yields, with $b_1 = b_2 = 0$, the IVP Green function:

$$G_{\rm ivp}(t,t') \;=\; G_{\rm ivp}(t-t') \;=\; \theta(t-t')\,\frac{\sin\big[\omega_0(t-t')\big]}{\omega_0}$$

*See pp. BF399-401 for useful properties of the step-function.

Note that the dependence of the IVP Green function is on the *difference*[†] $t-t'$. Indeed, it can be shown (EXERCISE) that for the second-order linear differential equation: $L[f(t)] = F(t)$ with *constant* coefficients, Green functions for a one-dimensional IVP must satisfy $G(t,t') = G(t-t')$, just by using the general form of the homogeneous solutions: $f_\pm(t) = \mathrm{e}^{\lambda_\pm t}$. This is a manifestation of the invariance of the differential operator with constant coefficients under translations of the variable t (eg. time). $G(t-t')$ can be viewed as a convolution operator or a response function in the sense of section 4.7.

By contrast, for the same L but with a BVP at $t_0 = 0$ and t_1, we immediately obtain:

$$b_1(t') \;=\; \frac{\cot\omega_0 t_1\,\sin\omega_0 t' \;-\; \cos\omega_0 t'}{\omega_0} \;=\; -\,\frac{\sin\omega_0(t_1-t')}{\omega_0\sin\omega_0 t_1}$$

and, with $t_> \coloneqq \max(t,t')$, $t_< \coloneqq \min(t,t')$, eq. (6.13) gives (EXERCISE):

$$G_{\mathrm{bvp}}(t,t') \;=\; \frac{1}{\omega_0\sin\omega_0 t_1}\,\sin\big[\omega_0(t_>-t_1)\big]\,\sin\omega_0 t_< \tag{6.16}$$

and a particular solution to $(\mathrm{d}_t^2+\omega_0^2)f(t) = F(t)$ is:

$$f(t) \;=\; \frac{\sin\big[\omega_0(t-t_1)\big]}{\omega_0\sin\omega_0 t_1}\int_{t_0}^{t}\sin\omega_0 t'\,F(t')\,\mathrm{d}t' \;+\; \frac{\sin\omega_0 t}{\omega_0\sin\omega_0 t_1}\int_{t}^{t_1}\sin\big[\omega_0(t'-t_1)\big]\,F(t')\,\mathrm{d}t'$$

If $\omega_0 t_1 = n\pi$ ($n\in\mathbb{Z}$), ie. if t_1 is an integer multiple of the half-period, the condition for the existence of a BVP Green function, $f_1(t_1)\,f_2(t_0) \;-\; f_2(t_1)\,f_1(t_0) \neq 0$, is violated.

Example 6.2. An interesting feature of eq. (6.13) is that the *explicit* dependence of the Green functions on the form of L is only through the coefficient of the second-order derivative. This can be exploited to minimise the work required to find the Green functions for a damped oscillator operator, $L = \mathrm{d}_t^2 + 2\gamma\mathrm{d}_t + \omega_0^2$, by using instead the appropriate homogeneous solutions: $f_1(t) = \mathrm{e}^{-\gamma t}\sin\big[\sqrt{\omega_0^2-\gamma^2}\,t\big]$, $f_2(t) \;=\; \mathrm{e}^{-\gamma t}\cos\big[\sqrt{\omega_0^2-\gamma^2}\,t\big]$. Now $W \;=\; -\sqrt{\omega_0^2-\gamma^2}\,\mathrm{e}^{-2\gamma t'}$, and a straightforward substitution into eq. (6.13) for an IVP ($b_1 = b_2 = 0$) gives:

$$G_{(3)}(t,t') \;=\; \theta(t-t')\,\mathrm{e}^{-\gamma(t-t')}\,\frac{\sin\big[\sqrt{\omega_0^2-\gamma^2}(t-t')\big]}{\sqrt{\omega_0^2-\gamma^2}} \tag{6.17}$$

Example 6.3. While we are talking about the damped harmonic oscillator, let us use it to illustrate another way to solve differential equations that combines Fourier and Green techniques. The idea is to write the equation:

$$\ddot{f}(t) \;+\; 2\gamma\,\dot{f}(t) \;+\omega_0^2\,f(t) \;=\; F(t)$$

in the frequency domain, assuming that the driving force dies at $t\to\pm\infty$ or, alternatively, is turned on at, say, $t=0$, and then off at some later time. In this case the Fourier transform of $F(t)$ exists, and eq. (5.14) can be used to obtain:

$$f(\omega) \;=\; \frac{F(\omega)}{-\omega^2 \;+\; 2\mathrm{i}\,\gamma\,\omega \;+\omega_0^2}$$

The differential equation in the time domain becomes an algebraic equation in the frequency domain! To go back to the time domain, we just write a solution to the inhomogeneous equation:

$$\begin{aligned} f(t) \;&=\; \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(\omega)\,\mathrm{e}^{\mathrm{i}\omega t}\,\mathrm{d}\omega \;=\; \int\left[\frac{1}{2\pi}\int\frac{\mathrm{e}^{\mathrm{i}\omega(t-t')}\,\mathrm{d}\omega}{-\omega^2 \;+\; 2\mathrm{i}\,\gamma\,\omega \;+\omega_0^2}\right]F(t')\,\mathrm{d}t' \\ &=\; \int_{-\infty}^{\infty} G(t,t')\,F(t')\,\mathrm{d}t' \end{aligned}$$

[†]The response function in eq. (4.22)is now seen to be a Green function.

where:

$$G(t,t') \;=\; G(t-t') \;=\; -\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{\mathrm{e}^{\mathrm{i}\omega(t-t')}}{(\omega-\omega_+)(\omega-\omega_-)}\,\mathrm{d}\omega$$

with $\omega_\pm = \pm\sqrt{\omega_0^2-\gamma^2}+\mathrm{i}\gamma$.

To calculate G for $t > t'$, we use contour integration in the complex ω plane, with the contour C chosen to be counterclockwise around the upper infinite half-plane. Both poles $\omega = \omega_\pm$ lie in the upper half-plane. Breaking up the contour into the real axis plus the semi-circle at infinity, we have:

$$-\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{\mathrm{e}^{\mathrm{i}\omega(t-t')}}{(\omega-\omega_+)(\omega-\omega_-)}\,\mathrm{d}\omega \;=\; -\frac{1}{2\pi}\oint_C\frac{\mathrm{e}^{\mathrm{i}\omega(t-t')}}{(\omega-\omega_+)(\omega-\omega_-)}\,\mathrm{d}\omega+\frac{1}{2\pi}\int_{|\omega|\to\infty}\frac{\mathrm{e}^{\mathrm{i}\omega(t-t')}}{(\omega-\omega_+)(\omega-\omega_-)}\,\mathrm{d}\omega$$

With $t-t' > 0$, the numerator in the second integral on the right goes to zero as $|\omega| \to \infty$, and the integral vanishes. The contour integral is evaluated with the Residue theorem (4.32). Then:

$$\begin{aligned} G(t-t') \;&=\; -\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{\mathrm{e}^{\mathrm{i}\omega(t-t')}}{(\omega-\omega_+)(\omega-\omega_-)}\,\mathrm{d}\omega \;=\; 2\pi\mathrm{i}\,\frac{-1}{2\pi}\left(\frac{\mathrm{e}^{\mathrm{i}\omega_+(t-t')}}{\omega_+-\omega_-}-\frac{\mathrm{e}^{\mathrm{i}\omega_-(t-t')}}{\omega_+-\omega_-}\right) \\ &=\; \mathrm{e}^{-\gamma(t-t')}\,\frac{\sin\left[\sqrt{\omega_0^2-\gamma^2}(t-t')\right]}{\sqrt{\omega_0^2-\gamma^2}} \end{aligned}$$

When $t-t' < 0$, we must use a contour enclosing the lower infinite half-plane. But the integrand in the contour integral is analytic in this region, and the integral vanishes by the Cauchy-Goursat theorem. Thus, $G(t,t') = 0$ for $t < t'$, and we have recovered the result obtained in eq. (6.17). Unlike that approach, here no knowledge of the homogeneous solutions was needed to find the Green function!

The general solution follows by adding a linear combination of homogeneous solutions.

6.2.3 Green's second 1-dim identity and general solution in terms of Green functions

Consider the expression, quadratic and symmetric in $u(t)$ and $v(t)$: $Q(u,v) \;=\; \alpha(t)\,\dot{u}\,\dot{v} \;+\; \lambda(t)\,(\dot{u}\,v \;+\; u\,\dot{v}) \;-\; \mu(t)\,u\,v$. Then, subtracting $Q(v,u)$, one derives after a few manipulations (EXERCISE) **Lagrange's identity**:

$$v\,L[u] \;-\; u\,L[v] \;=\; \mathrm{d}_t\big[v\,(\alpha\,(v\,\dot{u} \;-\; u\,\dot{v})\big]$$

where $L[u] = \mathrm{d}_t(\alpha\dot{u}) + (\dot{\lambda}+\mu)u$. Integrate this identity over an interval $[t_0,\,t_1]$ to obtain **Green's second identity** (1-d):

$$\int_{t_0}^{t_1}\big(v\,L[u] \;-\; u\,L[v]\big)\,\mathrm{d}t' \;=\; \alpha\,(v\,\dot{u} \;-\; u\,\dot{v})\Big|_{t_0}^{t_1} \tag{6.18}$$

Now suppose that $u = G(t,t')$ and that $v = f(t')$ satisfies the inhomogeneous equation $L[f(t')] = F(t')$. Then one easily shows from Green's identity that for $t \in [t_0,\,t_1]$,

$$f(t) \;=\; \int_{t_0}^{t_1} G(t,t')\,F(t')\,\mathrm{d}t' \;-\; \Big[\alpha\,\big(G\,\partial_{t'}f \;-\; f\,\partial_{t'}G\big)\Big]_{t'=t_0}^{t'=t_1} \tag{6.19}$$

where $G(t,t')$ is a Green function for the self-adjoint differential operator L. We are already familiar with the first (inhomogeneous) term, but the second one warrants careful examination. Obviously, it must be related to the homogeneous solutions. But wait—is $f(t)$ actually the general solution? Not yet! It is still just an identity. The second term is evaluated at the end-points of the interval, so it depends on the boundary conditions for f. We cannot freely specify f *and* $\dot{f}$ at both t_0 and t_1 as this would be in general inconsistent. If f if specified at the end-points, then we must find the solution for f in order to know what its derivatives are at the end-points.

But we can use the fact that any function $\tilde{G}(t,t')$ which satisfies $L[\tilde{G}] = 0$ can always be added to G to choose different types of boundary conditions for the Green function. For instance, specifying f at t_0 and t_1, we can in

principle find a $\tilde{G}$ such that the new G vanishes at t_0 and t_1. This kind of boundary-value problem, where the solution (not its derivative) is specified on the boundary, is called a **Dirichlet** problem, with the corresponding Green function denoted $G_{\rm D}$. Incidentally, this justifies our earlier setting of G to zero at the end-points. Now we can write the general *solution* for the two-point (Dirichlet) problem:

$$f(t) \;=\; \int_{t_0}^{t_1} G_{\rm D}(t,t')\,F(t')\,{\rm d}t' \;+\; \Big[\alpha\, f\, \partial_{t'} G_{\rm D}\Big]_{t'=t_0}^{t'=t_1} \qquad\qquad G_{\rm D}(t,t_0) \;=\; G_{\rm D}(t,t_1) \;=\; 0 \tag{6.20}$$

Compare this form of the general solution, which explicitly depends only on $F(t)$ and $G_{\rm D}$, plus $f(t_0)$ and $f(t_1)$, to the solution of eq. (6.12) in terms of the homogeneous solutions. It is a very instructive EXERCISE to show their equivalence. We also see that if f happens to obey homogeneous B.C., $f(t_0) = f(t_1) = 0$, there is no homogeneous part, consistent with our earlier statement near the end of section 6.2.1 that the existence of $G_{\rm bvp}$ is conditional on this absence.

We could instead want to specify $\dot{f}$ at t_0 and t_1. This **Neumann** problem has a solution if it is possible to set $\dot{G}_{\rm N}\big|_{t_0,t_1} = 0$ (or it has a solution up to a constant if $\dot{G}_{\rm N}\big|_{t_0,t_1}$ are constants). It is not nearly as frequent in physics as Dirichlet problems. Then:

$$f(t) \;=\; \int_{t_0}^{t_1} G_{\rm N}(t,t')\,F(t')\,{\rm d}t' \;-\; \Big[\alpha\, G_{\rm N}\, \partial_{t'} f\Big]_{t'=t_0}^{t'=t_1} \qquad\qquad \dot{G}_{\rm N}(t,t')\Big|_{t'=t_0} \;=\; \dot{G}_{\rm N}(t,t'))\Big|_{t'=t_1} \;=\; 0$$

For this to be possible, it will often be necessary to add an extra term to the defining equation of the Green function (more about this when we look at 3-dim Green functions). Much more exotic, we might want to specify f at one end-point and $\dot{f}$ at the other (so-called Robin B.C.); the B.C. on Green's function can be worked out by using Green's second identity.

I would argue that eq. (6.20) is a much nicer representation of the general solution to a Dirichlet problem than the more standard eq. (6.12). First, it has contributions from both volume and boundary terms. Second, the Dirichlet boundary conditions on f (which are really on the homogeneous solution since $G_{\rm D}(t,t') = 0$ at $t = t_0$ and $t = t_1$) appear explicitly in the solution. Third, unlike eq. (6.12), there is no explicit dependence on f_1 and f_2, and thus no need to calculate integration constants. It is true that f_1 and f_2 are still needed if $G_{\rm D}(t,t')$ is found via eq. (6.13), but we have seen in example 6.3 that Green functions can be calculated directly. Also, contrary to eq. (6.12), eq. (6.20) can be generalised to higher dimensions. Note, however, that this particular approach is not suited to an IVP, in which B.C. are specified at *one* point.

One important property of Dirichlet Green functions may be derived by letting $v = G_{\rm D}(t',t)$ and $u = G_{\rm D}(t',t'')$ in Green's second 1-dim identity (6.18), which holds for differential operators of the form $L = {\rm d}_t(\alpha {\rm d}_t) + \gamma$. Because $G_{\rm D} = 0$ at the end-points and $L\big[G(t,t')\big] = \delta(t-t')$, we immediately find that $G_{\rm D}$ for such operators is symmetric in its arguments:

$$G_{\rm D}(t,t'') \;=\; G_{\rm D}(t'',t) \tag{6.21}$$

The differential operator for the undamped harmonic oscillator is indeed of the right form (with $\alpha = 1$); we expect that $G_{\rm D}$ will be symmetric, and $G_{(1)}$ in eq. (6.16) is indeed symmetric. This property can provide a useful check on calculations.

Problems in More than One Dimension (BF 7.4)

In one dimension, Green's function for a second-order linear differential operator L always exists and is *unique* for an IVP. If it exists for a BVP (no zero eigenvalue for L), it is unique. This is closely related to the fact that boundary conditions are specified at one or two points only. In two or more dimensions, the boundaries contain an infinite number of points, and Green functions are no longer guaranteed to exist, even for an IVP, But they do exist in important cases of physical interest.

6.3 Differential Equations with Partial Derivatives

Unless you are working on superstrings, you will find that it is sufficient to study PDEs in no more than four dimensions[†].

In accordance with modern usage, we shall use Greek indices in four-dimensional (three spatial and one time) problems, and roman indices in three spatial dimensions. We also implement the Einstein summation convention according to which repeated indices in factors are to be summed over; in any such pair, we will try to write one index as a superscript and one as a subscript so as to spot them more easily. Then the form of a second-order linear differential operator that we shall use is:

$$L \;=\; \alpha^{\mu}(\mathbf{x})\partial_{\mu}^{2} \;+\; \beta^{\nu}(\mathbf{x})\partial_{\nu} \;+\; \gamma(\mathbf{x}) \tag{6.22}$$

where $\mathbf{x}$ is the generalised position and it should be emphasised that Cartesian coordinates are implied. The coefficients are assumed to be continuous in $\mathbf{x}$.

We follow Hadamard (1923) and classify such equations according to the coefficients of the second-order derivatives:

Definition 6.1.
- If at least one of the α^{μ} vanishes at some point, the operator (and corresponding homogeneous PDE will be said to be **parabolic** (eg. heat equation, Schrödinger equation, in which there is no second-order time-derivative).
- If the sign of one α^{μ} coefficient is different from all others, we say that L is **hyperbolic** (expected in a pseudo-Euclidean spacetime, eg. with the wave equation).
- If all α^{μ} coefficients have the same sign (expected in a Euclidean space), L is **elliptic** (eg. Laplace and Helmholtz operators — static 3-dim problems).

6.4 Separation of Variables in Elliptic Problems

Since the Laplacian operator occurs in all elliptic problems, it is worth taking a closer look at it. Our first task is to separate it into two convenient parts; at the same time this will get us acquainted with a very powerful technique.

6.4.1 An Important and Useful 3-dim Differential Operator

To do this, we introduce the self-adjoint vector operators $-\mathrm{i}\boldsymbol{\nabla}$ and $\mathbf{L} = -\mathrm{i}\,\mathbf{x}\times\boldsymbol{\nabla}$, or $L_i = -\mathrm{i}\epsilon_{ijk}x^j\partial^k$, where ϵ_{ijk} is the completely antisymmetric Levi-Civita symbol, and summation over repeated indices is implied. With the identity: $\epsilon_{ijk}\epsilon^{imn} = \delta_j{}^m\delta_k{}^n - \delta_j{}^n\delta_k{}^m$, the scalar product of $\mathbf{L}$ with itself is, in Cartesian coordinates:

$$\begin{aligned}\mathbf{L}\cdot\mathbf{L} \;&=\; -\,\epsilon_{ijk}\epsilon^{imn}\,x^j\partial^k\,x_m\partial_n \\ &=\; -\,x^j\left(\partial_j \;+\; x_j\,\partial^k\partial_k \;-\; 3\partial_j \;-\; x_k\partial^k\partial_j\right) \;=\; -\,x^jx_j\,\partial^k\partial_k \;+\; 2x^j\,\partial_j \;-\; x^j\,\partial_j \;+\; x^j\partial_j x^k\partial_k\end{aligned}$$

Extracting the Laplacian and reverting to coordinate-free notation, there comes:

$$\nabla^2 \;=\; -\frac{\mathbf{L}^2}{r^2} \;+\; \frac{1}{r}\left[\partial_r \;+\; \partial_r(r\,\partial_r)\right] \tag{6.23}$$

[†] Anyway, it is straightforward to generalise our discussion to any number of spatial dimensions plus one time dimension.

The distance r to the origin can be expressed in any coordinates we wish, yet this expression obviously wants to single out the direction along $\mathbf{x} = r\,\hat{\mathbf{n}}$ from the other two. Also, it would be nice if $\mathbf{L}$ only involved derivatives in directions perpendicular to $\hat{\mathbf{n}}$. This is most easily realised in a spherical coordinate system, since its radial coordinate naturally correpsonds to the direction along $\mathbf{x}$; the other two coordinates are *angular*. By transforming the Cartesian components of $\mathbf{L}$ to spherical coordinates $(r,\,\theta,\,\phi)$, we obtain (the calculation is rather tedious, but ***Maple/Mathematica*** will readily do it for us):

$$\begin{aligned} L_x &= -\mathrm{i}\,(y\partial_z - z\partial_y) = -\mathrm{i}\,(-\sin\phi\,\partial_\theta - \cot\theta\,\cos\phi\,\partial_\phi) \\ L_y &= -\mathrm{i}\,(z\partial_x - x\partial_z) = -\mathrm{i}\,(\cos\phi\,\partial_\theta - \cot\theta\,\sin\phi\,\partial_\phi) \\ L_z &= -\mathrm{i}\,(x\partial_y - y\partial_x) = -\mathrm{i}\,\partial_\phi \end{aligned}$$

The derivatives with respect to r have cancelled out! We also find that:

$$\mathbf{L}^2 = -\left[\frac{1}{\sin\theta}\,\partial_\theta\,(\sin\theta\,\partial_\theta) + \frac{1}{\sin^2\theta}\,\partial_\phi^2\right] \tag{6.24}$$

So $\mathbf{L}^2$ depends only on the angular coordinates. Also, eq. (6.23) makes it obvious that the **commutator** $[\nabla^2,\,\mathbf{L}^2] \coloneqq \nabla^2\mathbf{L}^2 - \mathbf{L}^2\nabla^2 = 0$, so that $[\nabla^2,\,\mathbf{L}] = 0$.

Now one readily shows that the following important relations hold:

$$[L_x,\,L_y] = \mathrm{i}\,L_z, \qquad [L_y,\,L_z] = \mathrm{i}\,L_x, \qquad [L_z,\,L_x] = \mathrm{i}\,L_y \tag{6.25}$$

Using these **commutator rules**, we find that:

$$\begin{aligned} [\mathbf{L}^2,\,L_z] &= [L_x^2,\,L_z] + [L_y^2,\,L_z] + \cancelto{0}{[L_z^2,\,L_z]} \\ &= L_x\,[L_x,\,L_z] + [L_x,\,L_z]\,L_x + L_y\,[L_y,\,L_z] + [L_y,\,L_z]\,L_y \\ &= -\mathrm{i}\,\cancel{L_xL_y} - \mathrm{i}\,\cancel{L_yL_x} + \mathrm{i}\,\cancel{L_yL_x} + \mathrm{i}\,\cancel{L_xL_y} \\ &= 0 \end{aligned}$$

By symmetry, we have immediately that $[\mathbf{L}^2, \mathbf{L}] = 0$.

6.4.2 Eigenvalues of $\mathbf{J}^2$ and J_z

The importance of eq.(6.25) cannot be overstated. It says that $\mathbf{L}$ belongs to the class of self-adjoint operators $\mathbf{J}$ whose Cartesian components satisfy the canonical commutation relations:

$$[J_i,\,J_j] = \mathrm{i}\,\epsilon_{ijk}\,J_k, \tag{6.26}$$

Just from these properties, it is possible to derive the eigenvalues λ of $\mathbf{J}^2$, ie. such that $\mathbf{J}^2 g = \lambda\,g$, and the eigenvalues m of J_z, such that $J_z f = m\,f$, where λ and m are expected to be *real* numbers since the operators are self-adjoint.

1. Introduce the **ladder** operators $J_\pm = J_x \pm \mathrm{i}J_y$. Then $[\mathbf{J}^2,\,J_\pm] = 0$. Since J_x and J_y are self-adjoint, we also have $J_+^\dagger = J_-$, and $J_-^\dagger = J_+$. Therefore:

$$[J_z,\,J_\pm] = [J_z,\,J_x] \pm \mathrm{i}\,[J_z,\,J_y] = \mathrm{i}\,J_y \pm \mathrm{i}\,(-\mathrm{i}\,J_x) = \pm\,J_\pm$$

Now, using $[\mathbf{J}^2,\,J_\pm] = 0$ and $[J_z,\,J_\pm] = \pm J_\pm$, we apply first $J_\pm$ and then J_z on f:

$$J_z(J_\pm f) \equiv [J_z,\,J_\pm]\,f + J_\pm J_z f = \pm\,J_\pm f + m\,J_\pm f = (m \pm 1)\,(J_\pm f)$$

Now f is also an eigenfunction of $\mathbf{J}^2$. Indeed, $J_z\mathbf{J}^2 f = \mathbf{J}^2 J_z f = m\mathbf{J}^2 f$, and $\mathbf{J}^2 f$ is an eigenfunction of J_z with the same eigenvalue as f. Since the eigenvalues of J_z are non-degenerate, $\mathbf{J}^2 f$ must be a multiple of f: $\mathbf{J}^2 f = \lambda f$. Also:

$$\mathbf{J}^2(J_\pm f) \;=\; J_\pm(\mathbf{J}^2 f) \;=\; \lambda\,(J_\pm f)$$

These results tell us what $J_\pm$ do for a living: they raise (J_+) or lower (J_-) the eigenvalues of J_z by 1, whence their name. In other words, if f is an eigenfunction of J_z with eigenvalue m, so is $J_\pm f$, but with eigenvalue $(m \pm 1)$. We also have found that *all* the eigenfunctions of J_z reachable with the ladder operators are eigenfunctions of $\mathbf{J}^2$ as well, with the *same* eigenvalue λ.

It is also reassuring to find that applying $J_\pm$ to eigenfunctions of J_z and J^2 gives a result that belongs to a Hilbert space; because $J_\pm$ are not self-adjoint, this was not guaranteed.

2. Another expression will be needed. Use the definition of $J_\pm$ to write:

$$J_\pm\, J_\mp \;=\; J_x^2 \,+\, J_y^2 \,\mp\, \mathrm{i}\,[J_x,\, J_y] \;=\; J_x^2 \,+\, J_y^2 \,\pm\, J_z \;=\; \mathbf{J}^2 \,-\, J_z^2 \,\pm\, J_z$$

so that:

$$\mathbf{J}^2 \;=\; J_\pm\, J_\mp \,+\, J_z^2 \,\mp\, J_z \tag{6.27}$$

3. For a given value of λ, we expect that m should have a maximum value, $m_{\max} \equiv j$, as well as a minimum value, $m_{\min} \equiv j'$.

   Now act with $\mathbf{J}^2$ on the eigenfunction of J_z with the maximum value of m, which we call f_j. Then $\mathbf{J}_+ f_j = 0$ and, from the identity (6.27), we find:

$$\mathbf{J}^2 f_j \;=\; J_z^2 f_j \,+\, J_z f_j \;=\; j(j+1)\, f_j \;=\; \lambda f_j$$

   Similarly, act with $\mathbf{J}^2$ on the eigenfunction of $\mathbf{J}_z$ with the minimum value of m, $f_{j'}$, keeping in mind that $f_{j'}$ is also an eigenfunction of $\mathbf{J}^2$ with the *same* eigenvalue, λ, as f_j:

$$\mathbf{J}^2 f_{j'} \;=\; (j')^2 f_{j'} \,-\, j' f_{j'} \;=\; j'(j'-1)\, f_{j'} \;=\; \lambda f_{j'}$$

   Comparing yields $\lambda \,=\, j(j+1) \,=\, j'(j'-1)$, and thus $j' = -j$. It follows that m goes from $-j$ to j in N integer steps, ie, $j = -j + N$, so $j = N/2$.

We conclude that:

- The eigenvalues of $\mathbf{J}^2$ are $j(j+1)$, where j is a positive integer or a half-integer.
- For a given value of j, m can take $2j+1$ values, from $-j$ to j.

It is worth stressing that these results were found without any knowledge of an explicit form for the eigenfunctions of $\mathbf{J}^2$, or indeed $\mathbf{J}$. And they apply to all self-adjoint operators which satisfy the canonical commutation relations (6.26).

With the help of eq. (6.27), we can now exhibit the full action of J_- on a normalised eigenfunction f_{jm} of $\mathbf{J}^2$ and J_z. Let $J_- f_{jm} = c_-\, f_{j,m-1}$. Then, using the rules for taking adjoints, and with $(f,\, g)$ the inner product of f and g:

$$(f_{jm},\, J_+ J_-\, f_{jm}) \;=\; (J_- f_{jm},\, J_-\, f_{jm}) \;=\; (c_- f_{j,m-1},\, c_- f_{j,m-1}) \;=\; |c_-|^2\,(f_{j,m-1},\, f_{j,m-1}) \;=\; |c_-|^2$$

But since $J_\pm\, J_\mp = \mathbf{J}^2 - J_z^2 \,\pm\, J_z$, we also have that:

$$(f_{lm},\, J_+ J_-\, f_{jm}) \;=\; (f_{jm},\, (\mathbf{J}^2 - J_z^2 + J_z)\, f_{jm}) \;=\; j(j+1) \,-\, m^2 \,+\, m$$

Comparing yields c_- up to an unimportant exponential phase factor which we put equal to 1. We find the coefficient in $J_+ f_{jm} = c_+\, f_{j,m+1}$ in a strictly analogous way. The results for both ladder operators are:

$$J_\pm f_{jm} \;=\; \sqrt{j(j+1) \,-\, m(m \,\pm 1)}\, f_{j,m\pm1} \tag{6.28}$$

6.4.3 Eigenfunctions of $\mathbf{L}^2$ and $\mathbf{L}_z$

To find the common eigenfunctions for $\mathbf{J}$ and $\mathbf{J}^2$ operators, we must know their form. Here, we will be interested in the $\mathbf{L}$ operator whose form we do know and which makes up the angular part of the Laplacian in spherical coordinates.

The eigenfunctions of L_z are readily obtained by solving the differential equation:

$$L_z f(\theta,\phi) \;=\; -\mathrm{i}\,\partial_\phi f(\theta,\phi) \;=\; m\, f(\theta,\phi)$$

With a separation ansatz: $f(\theta,\phi) = F(\theta)G(\phi)$, the solution for G is:

$$G(\phi) = \mathrm{e}^{\mathrm{i}m\phi} \tag{6.29}$$

Now we require that G (and f) be single-valued, that is, $G(\phi+2\pi) = G(\phi)$. Thus:

$$\mathrm{e}^{\mathrm{i}m(\phi+2\pi)} \;=\; \mathrm{e}^{\mathrm{i}m\phi} \quad\Longrightarrow \mathrm{e}^{2\mathrm{i}m\pi} \;=\; \cos 2m\pi \;+\; \mathrm{i}\,\sin 2m\pi \;=\; 1$$

which constrains m to be any *integer*. Therefore, $l \coloneqq m_{\max}$ must also be an integer. Thus, we find that the particular form $\mathbf{L} = -\mathrm{i}\mathbf{x}\times\boldsymbol{\nabla}$ rules out the possibility of half-integer values of j allowed for a self-adjoint $\mathbf{J}$ that satisfies the canonical commutation relations (6.26).

The θ dependence of the eigenfunctions must be derived from the eigenvalue equation for $\mathbf{L}^2$. Call $f(\theta,\phi) = Y_l^m(\theta,\phi) = F(\theta)G(\phi)$; these must satisfy:

$$-\left[\frac{1}{\sin\theta}\,\partial_\theta\left(\sin\theta\,\partial_\theta\right) \;+\; \frac{1}{\sin^2\theta}\,\partial_\phi^2\right] Y_l^m(\theta,\phi) \;=\; l(l+1)\,Y_l^m(\theta,\phi)$$

Inserting $Y_l^m(\theta,\phi) = F(\theta)\mathrm{e}^{\mathrm{i}m\phi}$ into this equation leaves:

$$-\left[\frac{1}{\sin\theta}\,\mathrm{d}_\theta\left(\sin\theta\,\mathrm{d}_\theta\right) \;-\; \frac{m^2}{\sin^2\theta}\right] F(\theta) \;=\; l(l+1)\,F(\theta)$$

Instead of solving this equation by brute force, we use a clever technique involving the ladder operators $L_\pm$:

$$L_\pm \;=\; \pm\,\mathrm{e}^{\mathrm{i}\phi}\left(\partial_\theta \;\pm\; \mathrm{i}\,\cot\theta\,\partial_\phi\right)$$

Now, when $m = l$, we have:

$$L_+ Y_l^l \;=\; \mathrm{e}^{\mathrm{i}\phi}\left(\partial_\theta \;+\; \mathrm{i}\,\cot\theta\,\partial_\phi\right) Y_l^l(\theta,\phi) \;=\; 0$$

Inserting $Y_l^l = F(\theta)\mathrm{e}^{\mathrm{i}l\phi}$, this reduces to the much simpler

$$\mathrm{d}_\theta F(\theta) \;-\; l\,\cot\theta\,F(\theta) \;=\; 0$$

whose solution is $F(\theta) = (\sin\theta)^l$. Therefore, $Y_l^l = (\sin\theta)^l\mathrm{e}^{\mathrm{i}l\phi}$. Applying L_- the requisite number of times generates the other Y_l^m ($0 < m < l$): $Y_l^m \propto L_-^{l-m}Y_l^l$. When normalised, these are the **spherical harmonics** already found in eq. (5.50):

$$Y_l^m(\theta,\phi) \;=\; \frac{(-1)^m}{2^l l!}\sqrt{\frac{2l+1}{4\pi}\,\frac{(l-m)!}{(l+m)!}}\,(1-x^2)^{m/2}\left[\mathrm{d}_x^{l+m}(x^2-1)^l\right]\mathrm{e}^{\mathrm{i}m\phi} \qquad x=\cos\theta \tag{6.30}$$

6.4.4 General Solution of a Spherically-Symmetric, 2nd-order, Homogeneous, Linear Equation

Suppose we are presented with the equation $\left[\nabla^2 + \gamma(\mathbf{x})\right]\Psi(\mathbf{x}) = 0$. Work in spherical coordinates, and make the ansatz: $\Psi(\mathbf{x}) = R(r)F(\theta,\phi)$. Using the form for ∇^2 derived earlier, eq. (6.23), we have:

$$\begin{aligned}\nabla^2\Psi \,+\, \gamma(\mathbf{x})\Psi \;&=\; -\,\frac{\mathbf{L}^2\Psi}{r^2} \,+\, \frac{1}{r}\left[\partial_r\Psi \,+\, \partial_r(r\,\partial_r\Psi)\right] \,+\, \gamma(\mathbf{x})\,\Psi \\ &= -\,R(r)\frac{\mathbf{L}^2F(\theta,\phi)}{r^2} \,+\, \frac{F(\theta,\phi)}{r}\left[\mathrm{d}_rR(r) \,+\, \mathrm{d}_r(r\,\mathrm{d}_rR(r))\right] \,+\, \gamma(\mathbf{x})\,R(r)F(\theta,\phi)\end{aligned}$$

Dividing the second line by $R(r)F(\theta,\phi)$ and multiplying by r^2, we see that the equation is separable provided $\gamma(\mathbf{x}) = \gamma(r)$:

$$\mathbf{L}^2F(\theta,\phi) \;=\; \lambda\,F(\theta,\phi) \qquad\qquad \mathrm{d}_rR(r) \,+\, \mathrm{d}_r(r\,\mathrm{d}_rR(r)) \,+\, r\,\gamma(r)\,R(r) \;=\; \lambda\,\frac{R(r)}{r}$$

The first equation is the eigenvalue equation for $\mathbf{L}^2$, whose eigenvalues are $\lambda = l(l+1)$ ($l \geq 0 \in \mathbb{Z}$) with the spherical harmonics $Y_l^m(\theta,\phi)$ as eigenfunctions.

The radial equation can thus be written: $\dfrac{1}{r^2}\,\mathrm{d}_r\big(r^2\,\mathrm{d}_rR_l(r)\big) \,+\, \left(\gamma(r) \,-\, \dfrac{l(l+1)}{r^2}\right) R_l(r) \;=\; 0$

When $\gamma(r) = 0$, this is the radial part of the Laplace equation which becomes, after the change of variable $r = \mathrm{e}^x$, $\mathrm{d}_x^2R + \mathrm{d}_xR - l(l+1)R = 0$. Inserting a solution of the form e^{px} turns the equation into $p^2 + p - l(l+1) = 0$, that is, $p = l$ or $p = -(l+1)$, which leads to $R = A\mathrm{e}^{lx} \,+\, B\mathrm{e}^{-(l+1)x} = Ar^l + Br^{-(l+1)}$. Therefore, the general solution to the Laplace equation in spherical coordinates is:

$$\Psi(r,\theta,\phi) \;=\; \sum_{l=0}^{\infty}\sum_{m=-l}^{l}\left(A_{lm}\,r^l \,+\, \frac{B_{lm}}{r^{l+1}}\right) Y_l^m(\theta,\phi) \tag{6.31}$$

The coefficients A_{lm} and B_{lm} are determined from boundary or matching conditions. In regions either containing the origin, or extending all the way to infinity, $B_{lm} = 0$ or $A_{lm} = 0$, respectively. Clearly, if this solution is to be regular, and if it holds *everywhere*, it must vanish. In other words, if the Laplace equation is valid everywhere, it has no non-vanishing regular solution. For a non-trivial solution, there must be a region of space where there exists an inhomogeneous term acting as a *source*.

Note, however, that the general solution holds at any point where there is no source. The effect of sources is encoded in the coefficients A_{lm} and B_{lm}.

When $\gamma(r) = k^2 > 0$, we get the radial part of the Helmholtz equation in spherical coordinates:

$$\mathrm{d}_r^2R_l(r) \,+\, \frac{2}{r}\,\mathrm{d}_rR_l(r) \,+\, \left(k^2 \,-\, \frac{l(l+1)}{r^2}\right) R_l(r) \;=\; 0$$

which the substitutions $R_l = u_l/\sqrt{r}$ and $x = kr$ readily transform into:

$$\mathrm{d}_x^2u_l(x) \,+\, \frac{1}{x}\,\mathrm{d}_xu_l(x) \,+\, \left(1 \,-\, \frac{(l+1/2)^2}{x^2}\right) u_l(x) \;=\; 0$$

that is, the Bessel equation (4.31) with $n = l + 1/2$. The solutions are the **spherical Bessel functions** of the first and second (Neumann) kind, usually written as (see also Jackson's *Classical Electrodynamics*, section 9.6):

$$j_l(x) \;=\; \sqrt{\frac{\pi}{2x}}\,J_{l+1/2}(x) \;=\; (-x)^l\left(\frac{1}{x}\frac{\mathrm{d}}{\mathrm{d}x}\right)^l\left(\frac{\sin x}{x}\right) \qquad \sim \begin{cases} x^l & x \ll (1,l) \\ \dfrac{1}{x}\sin(x - l\pi/2) & x \gg l \end{cases} \tag{6.32}$$

$$n_l(x) \;=\; \sqrt{\frac{\pi}{2x}}\,N_{l+1/2}(x) \;=\; -\,(-x)^l\left(\frac{1}{x}\frac{\mathrm{d}}{\mathrm{d}x}\right)^l\left(\frac{\cos x}{x}\right) \qquad \sim \begin{cases} -\dfrac{1}{x^{l+1}} & x \ll (1,l) \\ -\dfrac{1}{x}\cos(x - l\pi/2) & x \gg l \end{cases} \tag{6.33}$$

The n_l diverge at the origin and thus are excluded from any solution regular at the origin.

(Spherical) Bessel functions of the third kind, aka **Hankel** functions of the first and second kind, sometimes come in handy: $h_l^{(1,2)}(x) = j_l(x) \pm \mathrm{i}\, n_l(x)$. One can express the general solution of the Helmholtz equation in terms of the j_l and n_l, or in terms of the $h_l^{(1,2)}$.

6.5 Second 3-dim Green Identity, or Green's Theorem

Before discussing the all-important subject of boundary conditions, we derive a result that will prove very useful in the study of 3-dim *elliptic* problems. We assume that $L[f] = \partial^i\big(\alpha(\mathbf{x})\partial_i f\big) + \gamma(\mathbf{x}) f$ (Sturm-Liouville form). If it isn't, it is possible to bring it into that form, *provided that* the solutions of $L[g] = 0$ go to zero at infinity or else that $\alpha \to 0$ sufficiently fast.

Write the divergence theorem for $\boldsymbol{\nabla} \cdot (\alpha f \boldsymbol{\nabla} g)$ defined over a connected volume, and expand the divergence to get:

$$\int_V \Big[f\, \boldsymbol{\nabla} \cdot (\alpha\, \boldsymbol{\nabla} g) \,+\, \alpha\, \boldsymbol{\nabla} f \cdot \boldsymbol{\nabla} g \Big]\, \mathrm{d}^3x \;=\; \oint_{\partial V} \alpha\, f\, \boldsymbol{\nabla} g \cdot \mathrm{d}\mathbf{S} \tag{6.34}$$

where ∂V is the boundary of the volume V of integration. This is **Green's first identity in three dimensions**; when α is a constant, it reduces to the more familiar form:

$$\int_V \Big[f\, \nabla^2 g \,+\, \boldsymbol{\nabla} f \cdot \boldsymbol{\nabla} g \Big]\, \mathrm{d}^3x \;=\; \oint_{\partial V} f\, \boldsymbol{\nabla} g \cdot \mathrm{d}\mathbf{S} \tag{6.35}$$

Interchanging f and g in the first identity (6.34) and subtracting, we easily find, after adding and subtracting $\gamma f g$ in the volume itnegral, the **second Green identity in three dimensions**:

$$\int_V \big(f\, L[g] \,-\, g\, L[f] \big) \mathrm{d}^3x \;=\; \oint_{\partial V} \alpha\, \big(f\, \boldsymbol{\nabla} g \,-\, g\, \boldsymbol{\nabla} f \big) \cdot \mathrm{d}\mathbf{S} \tag{6.36}$$

With α a constant, this becomes the well-known Green theorem:

$$\int_V \big(f\, \nabla^2 g \,-\, g\, \nabla^2 f \big) \mathrm{d}^3x \;=\; \oint_{\partial V} \big(f\, \boldsymbol{\nabla} g \,-\, g\, \boldsymbol{\nabla} f \big) \cdot \mathrm{d}\mathbf{S} \tag{6.37}$$

Note that in the case of conpact regions without boundaries (sphere, torus), the right-hand side vanishes.

6.5.1 Uniqueness and existence of solutions for the inhomogeneous Laplace equation with B.C.

The inhomogeneous Laplace (aka Poisson) equation is of the form $\nabla^2 \Psi(\mathbf{x}) = F(\mathbf{x})$, where the right-hand side is called a **source** term. As justified below, we also specify B.C. for *either* Ψ *or* $\hat{\mathbf{n}} \cdot \boldsymbol{\nabla} \Psi$ on a closed boundary.

Now, with $f = g = \Psi_3$ and α constant, Green's first identity—eq. (6.35)— becomes:

$$\int_V [\Psi_3 \nabla^2 \Psi_3 + (\boldsymbol{\nabla} \Psi_3)^2]\, \mathrm{d}^3x \;=\; \oint_{\partial V} \Psi_3\, \partial_n \Psi_3\, \mathrm{d}S$$

where we have introduced the **normal derivative** $\partial_n \Psi_3$, ie. the component of $\boldsymbol{\nabla} \Psi_3$ normal *outward* to ∂V.

Suppose there exist two solutions, Ψ_1 and Ψ_2, of $\nabla^2 \Psi(\mathbf{x}) = F(\mathbf{x})$ that satisfy the same conditions on the surface. Define $\Psi_3 := \Psi_2 - \Psi_1$. Then $\nabla^2 \Psi_3 = 0$ inside the volume. The surface integral is zero because either $\Psi_3 = 0$ or $\partial \Psi_3 / \partial n = 0$ on the surface, and $\int (\boldsymbol{\nabla} \Psi_3)^2 \mathrm{d}^3x = 0$ everywhere. Also, Ψ_3 being twice differentiable at all points in the volume, $\boldsymbol{\nabla} \Psi_3$ is continuous and therefore zero everywhere inside the volume, so that Ψ_3 is a constant. It follows immediately that if $\Psi_3 = 0$ on the boundary, $\Psi_3 = 0$ everywhere; on the other hand, when $\partial \Psi_3 / \partial n = 0$ on the boundary, Ψ_3 can be a non-zero constant inside.

We conclude that $\Psi_1 = \Psi_2$ inside the volume (up to a possible additive constant), and that the solution, *if it exists*, is uniquely determined. The importance of this result cannot be overemphasised: any function that satisfies the inhomogeneous Laplace (aka Poisson) equation and the B.C. is *the* solution, no matter how it was found!

Moreover, we see that we cannot arbitrarily specify *both* Ψ and $\partial\Psi/\partial n$ on the boundary since one suffices to determine the unique solution.

The BC determine the solution, but only if it exists. Further conditions must be met for this to happen. Indeed, with $\phi(\mathbf{x}) \coloneqq \partial_n \Psi\big|_{\mathbf{x}\in\partial V}$, integrate $\nabla^2\Psi(\mathbf{x}) = F(\mathbf{x})$ over (connected!) V; the divergence theorem immediately yields a condition linking ϕ to the source F:

$$\int_V F(\mathbf{x})\,\mathrm{d}^3x \;=\; \int_{\partial V} \phi(\mathbf{x})\,\mathrm{d}S \tag{6.38}$$

Another condition for the existence of a solution is that the enclosing boundary be "reasonably" smooth (eg. no pointy parts), otherwise it may prove impossible even to impose B.C.

6.6 3-dim Boundary Value (Elliptic) Problems with Green's Functions

Introduce Green functions that satisfy $L\big[G(\mathbf{x},\mathbf{x}')\big] = \delta(\mathbf{x}-\mathbf{x}')$ in compact regions with closed boundaries or in non-compact regions[†] (again, some authors multiply the right-hand side by $\pm 4\pi$). If we are a little careful, we will find that for some important cases this kind of problem can admit unique Green functions.

6.6.1 Dirichlet and Neumann Boundary Conditions

Suppose that $\Psi(\mathbf{x})$ satisfies[†] $L\,\Psi(\mathbf{x}) = F(\mathbf{x})$; then take $f = \Psi$ and $g = G$ in Green's second identity (eq. (6.36)):

$$\int_V \big(\Psi\,L[G] - G\,L[\Psi]\big)\,\mathrm{d}^3x' \;=\; \oint_{\partial V} \alpha\,(\Psi\,\partial_{n'}G - G\,\partial_{n'}\Psi)\,\mathrm{d}S'$$

where, as in example 6.5.1, $\boldsymbol{\nabla} f\cdot\mathrm{d}\mathbf{S} = \partial_n f\,\mathrm{d}S$. We obtain:

$$\int_V \big[\Psi(\mathbf{x}')\,\delta(\mathbf{x}-\mathbf{x}') \;-\; F(\mathbf{x}')\,G(\mathbf{x},\mathbf{x}')\big]\,\mathrm{d}^3x' \;=\; \oint_{\partial V} \alpha\,(\Psi\,\partial_{n'}G \;-\; G\,\partial_{n'}\Psi)\,\mathrm{d}S'$$

With $\mathbf{x}$ inside the volume, re-arranging then yields:

$$\Psi(\mathbf{x}) \;=\; \int_V F(\mathbf{x}')\,G(\mathbf{x},\mathbf{x}')\,\mathrm{d}^3x' \;+\; \oint_{\partial V} \alpha\,(\Psi\,\partial_{n'}G - G\,\partial_{n'}\Psi)\,\mathrm{d}S' \tag{6.39}$$

where the normal derivatives in the integrand on the right-hand side are to be evaluated *on* ∂V, the boundary of the arbitrary volume. This expression for Ψ cannot be considered a solution yet; it is still "just" an *identity*.

Again, note that Ψ and $\partial\Psi/\partial n$ are *in general not independent on the boundary*. We are not free to specify them both arbitrarily at the same point on ∂V as such values will in general be inconsistent.

Specifying Ψ on the boundary gives **Dirichlet B.C.**, whereas specifying $\partial\Psi/\partial n$ gives **Neumann B.C.**

How do we get a solution for Ψ then? *In principle*, this is simple. We use the fact that the Green functions we find by solving $L\big[G(\mathbf{x},\mathbf{x}')\big] = \delta(\mathbf{x}-\mathbf{x}')$ are not unique; indeed, we can add to them any function G_1 that satisfies $L\,G_1(\mathbf{x},\mathbf{x}') = 0$. "All" we have to do then is find a G_1 that eliminates one of the two surface integrals.

Suppose we wish to specify Ψ freely on the boundary (Dirichlet problem). Then we should ensure that $G_{\mathrm{D}}(\mathbf{x},\mathbf{x}') = 0\;\forall\,\mathbf{x}'\in\partial V$. The *solution* for Ψ would then be:

$$\Psi(\mathbf{x}) \;=\; \int_V F(\mathbf{x}')\,G_{\mathrm{D}}(\mathbf{x},\mathbf{x}')\,\mathrm{d}^3x' \;+\; \oint_{\partial V} \alpha\,\Psi(\mathbf{x}')\,\partial_{n'}G_{\mathrm{D}}(\mathbf{x},\mathbf{x}')\,\mathrm{d}S' \tag{6.40}$$

The solution is now uniquely determined by the B.C. on Ψ via G_{D}. Note that the total surface ∂V enclosing the volume may be disjoint, as occurs for instance with the volume between two concentric spheres.

[†] In compact domains without boundaries (see remark in last section), this defining equation is inconsistent and must be modified (EXERCISE: Can you see why, and how?)

[†] Although we call it "inhomogeneous", nothing in what we will do here prevents $F(\mathbf{x})$ from depending on $\Psi(\mathbf{x})$

If we have managed to find $G_{\rm D}$ for a particular type of boundary, the source-free solution ($F(\mathbf{x}') = 0$) is just the surface integral; on the other hand, if it happens that $\Psi = 0$ on ∂V, only the volume integral contributes. Many boundary-value problems in electrostatics, for which the boundary conditions are reasonably simple, can be solved this way.

With Green's second identity, it is also straightforward to prove (exercise) that $G_{\rm D}(\mathbf{x}, \mathbf{x}')$ is symmetric in its arguments.

Similar considerations apply to Neumann boundary conditions, ie. when $\partial\Psi/\partial n$ rather than Ψ is known on the boundary. But we must be a little careful about the boundary conditions on $\partial_n G_{\rm N}$: we cannot simply put this equal to 0 in eq. (6.39). Indeed, take for instance $L = \nabla^2$; then, from the divergence theorem and the defining equation $L[G_{\rm N}] = \delta(\mathbf{x} - \mathbf{x}')$:

$$\int \boldsymbol{\nabla} \cdot \boldsymbol{\nabla} G_{\rm N} \, \mathrm{d}^3 x \;=\; \oint_{\partial V} \partial_n G_{\rm N} \, \mathrm{d}S \;=\; 1$$

A consistent boundary condition is $\partial_n G_{\rm N}\big|_{\partial V} = 1/S$, and we obtain:

$$\Psi(\mathbf{x}) \;= < \Psi >_{\partial V} \; + \int_V F(\mathbf{x}') \, G_{\rm N}(\mathbf{x}, \mathbf{x}') \, \mathrm{d}^3 x' \;-\; \oint_{\partial V} G_{\rm N}(\mathbf{x}, \mathbf{x}') \, \partial_{n'} \Psi(\mathbf{x}') \, \mathrm{d}S' \tag{6.41}$$

Up to the a priori unknown average of Ψ over the surface, $< \Psi >_{\partial V}$, a constant, this is the solution to that Neumann problem. Often (but not always!) the volume is bounded by two surfaces, one closed and finite and the other at infinity, in which case the normal derivative of $G_{\rm N}$ can be set to zero on the entire boundary, and the average of Ψ over ∂V (the first term) vanishes.

6.6.2 Green's function for the 3-d Elliptic Helmholtz operator without boundary conditions

We proceed to find a Green function for the operator $\boldsymbol{\nabla}^2 + \lambda$, with λ a constant. Using eq. (5.13), the Fourier transform of $(\nabla^2 + \lambda)\,\Psi(\mathbf{x}) = F(\mathbf{x})$ is $(-k^2 + \lambda)\psi(\mathbf{k}) = F(\mathbf{k})$. We must distinguish between two possibilities:

1. $\lambda = -\kappa^2 \leq 0$, $\kappa \geq 0$

   Then, similarly to what happens in one dimension (example 6.3), an "inhomogeneous" solution is:

   $$\Psi(\mathbf{x}) \;=\; -\frac{1}{(2\pi)^{3/2}} \int \frac{F(\mathbf{k})}{k^2 \,+\, \kappa^2} \, \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{x}} \, \mathrm{d}^3 k \;=\; -\frac{1}{(2\pi)^3} \iint \mathrm{d}^3 x' \, \mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{x}'} \, \frac{F(\mathbf{x}')}{k^2 \,+\, \kappa^2} \, \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{x}} \, \mathrm{d}^3 k$$

   Compare with the Green-function form of the inhomogeneous solution, $\int_V F(\mathbf{x}') G(\mathbf{x}, \mathbf{x}') \, \mathrm{d}^3 x'$ (EXERCISE):

   $$G(\mathbf{x}, \mathbf{x}') \;=\; -\frac{1}{(2\pi)^3} \int \frac{\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}')}}{k^2 \,+\, \kappa^2} \, \mathrm{d}^3 k \;=\; \frac{\mathrm{i}}{(2\pi)^2 \, |\mathbf{x} - \mathbf{x}'|} \int_{-\infty}^{\infty} \frac{k \, \mathrm{e}^{\mathrm{i}k|\mathbf{x}-\mathbf{x}'|}}{k^2 \,+\, \kappa^2} \, \mathrm{d}k$$

   This integral is easily evaluated as part of a contour integral around a semi-circle at infinity in the upper complex k half-plane. As before, the contribution at infinity vanishes, and the residue due to the pole at $k = \mathrm{i}\kappa$ is $\mathrm{e}^{-\kappa|\mathbf{x}-\mathbf{x}'|}/2$. The Residue theorem then yields the (sometimes called fundamental, or singular) solution:

   $$G(\mathbf{x}, \mathbf{x}') \;=\; -\frac{1}{4\pi} \, \frac{\mathrm{e}^{-\kappa|\mathbf{x}-\mathbf{x}'|}}{|\mathbf{x} - \mathbf{x}'|} \tag{6.42}$$

   This is very well-behaved at infinity. For $\lambda = 0$ ($\kappa = 0$), we obtain a Green function for the Laplacian operator.

   With $\kappa = 0$ and $F(\mathbf{x}) = -4\pi\rho(\mathbf{x})$ (Gaussian units!), for instance, an inhomogeneous solution is the generalised Coulomb Law for the electrostatic potential of a localised charge density $\rho(\mathbf{x})$, or one that vanishes at infinity faster than $1/|\mathbf{x} - \mathbf{x}'|$.

2. $\lambda = \kappa^2 \geq 0$

In order to invert the algebraic equation for $\psi(\mathbf{k})$, we write $\lambda = (q \pm \mathrm{i}\epsilon)^2$ ($\epsilon \geq 0$). Then we arrive at:

$$G_q^{(\pm)}(\mathbf{x},\mathbf{x}') \;=\; -\frac{1}{(2\pi)^3}\lim_{\epsilon\to 0}\int \frac{\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}')}}{k^2 \,-\, (q\pm\mathrm{i}\epsilon)^2}\,\mathrm{d}^3k \;=\; -\frac{1}{4\pi}\,\frac{\mathrm{e}^{\pm\mathrm{i}q|\mathbf{x}-\mathbf{x}'|}}{|\mathbf{x}-\mathbf{x}'|} \tag{6.43}$$

For details of the calculation, see pp. BF415–416.

You should check that these Green functions satisfy $(\nabla^2+\lambda)G(\mathbf{x},\mathbf{x}') = \delta(\mathbf{x}-\mathbf{x}')$. But do note that they are *not* the general solution of this equation, since any function that satisfies the homogeneous equation can be added to them!

If the volume integral extends over all space, the surface integral in the Dirichlet solution for the case $\lambda < 0$ certainly vanishes at infinity for fairly weak conditions on $\Psi(\mathbf{x})$, because of the exponential factor in Green's function. When $\lambda \geq 0$, the surface integral also vanishes provided $\Psi(\mathbf{x}) \to 0$ faster than $1/|\mathbf{x}-\mathbf{x}'|^2$, (since $\mathrm{d}S \sim |\mathbf{x}-\mathbf{x}'|^2$), and we are left with just the inhomogeneous integral:

$$\Psi_q^{(\pm)}(\mathbf{x}) \;=\; -\frac{1}{4\pi}\int_V \frac{F(\mathbf{x}')\,\mathrm{e}^{\pm\mathrm{i}q|\mathbf{x}-\mathbf{x}'|}}{|\mathbf{x}-\mathbf{x}'|}\,\mathrm{d}^3x' \tag{6.44}$$

If, however, $\Psi(\mathbf{x})$ does not vanish fast enough at infinity, it is more convenient to write it in terms of the solution of the homogeneous equation $(\nabla^2+q^2)\Psi(\mathbf{x}) = 0$, plus the volume integral:

$$\Psi_q^{(\pm)}(\mathbf{x}) \;=\; A\,\mathrm{e}^{\mathrm{i}\mathbf{q}\cdot\mathbf{x}} \;-\; \frac{1}{4\pi}\int_V \frac{F(\mathbf{x}')\,\mathrm{e}^{\pm\mathrm{i}q|\mathbf{x}-\mathbf{x}'|}}{|\mathbf{x}-\mathbf{x}'|}\,\mathrm{d}^3x' \tag{6.45}$$

Note that these expressions for Green's functions assume no boundary surfaces (except at infinity)!

6.6.3 Dirichlet Green function for the Laplacian

When there are no boundary conditions for Ψ on *finite* surfaces, the volume integral $\int F(\mathbf{x}')G(\mathbf{x},\mathbf{x}')\mathrm{d}^3x'$ can be taken as the solution to $L[\Psi] = F$. For instance, in the case of a point-source located at $\mathbf{y}$: $F(\mathbf{x}') = -4\pi q\delta(\mathbf{y}-\mathbf{x}')$, with q some constant, we see that $\Psi(\mathbf{x}) = -4\pi q G(\mathbf{x},\mathbf{y}) = q/|\mathbf{x}-\mathbf{y}|$ in the case of $L = \nabla^2$.

When there are finite boundaries, however, as in a Dirichlet problem, we know that we have to ensure that $G_\mathrm{D}(\mathbf{x},\mathbf{x}') = 0$ when either $\mathbf{x}$ or $\mathbf{x}'$ is a point on the surface that encloses the volume in which our solution is valid. Obviously, with the Green function given in eq. (6.42), which vanishes only on a boundary at infinity, this is impossible. It is time to exercise our freedom to add to G a function that satisfies the homogeneous equation $L[G] = 0$ and contains free parameters that can be set so as force the combined Green function to vanish on the boundary. In the case of the Laplacian, we take:

$$G_\mathrm{D}(\mathbf{x},\mathbf{x}') \;=\; -\frac{1}{4\pi}\left(\frac{1}{|\mathbf{x}-\mathbf{x}'|} \;+\; \frac{g}{|\mathbf{x}-\mathbf{x}''|}\right)$$

where g and $\mathbf{x}''$ will ensure that the second term satisfies the Laplace equation $\forall\,\mathbf{x}$ *inside* the volume where we are looking for a solution, as well as vanishing on the boundary.

Example 6.4. Solution of the Dirichlet problem on a sphere for the Laplacian

Consider a sphere of radius a centered on the origin. We want: $G_\mathrm{D}(a\hat{\mathbf{n}},\mathbf{x}') = G_\mathrm{D}(\mathbf{x},a\hat{\mathbf{n}}') = 0$ Symmetry dictates that $\mathbf{x}''$ and $\mathbf{x}'$ be collinear, which means that, at $|\mathbf{x}| = r = a$, we can write:

$$G_\mathrm{D}(a\hat{\mathbf{n}},\mathbf{x}') \;=\; -\frac{1}{4\pi}\left(\frac{1}{a\,|\hat{\mathbf{n}}-\frac{r'}{a}\hat{\mathbf{n}}'|} \;+\; \frac{g}{r''\,|\frac{a}{r''}\hat{\mathbf{n}} \,-\, \hat{\mathbf{n}}''|}\right)$$

where $r\hat{\mathbf{n}} = \mathbf{x}$, etc. By inspection, we see that if $G_{\rm D}(a\hat{\mathbf{n}}, \mathbf{x}')$ is to vanish for $\hat{\mathbf{n}}$ in an arbitrary direction, we must have: $1/a = -g/r''$ and $r'/a = a/r''$, as well as: $\hat{\mathbf{n}}\cdot\hat{\mathbf{n}}' = \hat{\mathbf{n}}\cdot\hat{\mathbf{n}}''$ ($\hat{\mathbf{n}}'$ and $\hat{\mathbf{n}}''$ collinear). Then:

$$g \;=\; -\,a/r', \qquad\qquad r'\,r'' \;=\; a^2 \tag{6.46}$$

Thus, $\mathbf{x}''$ lies *outside* the sphere if $\mathbf{x}'$ is inside, and *vice-versa*. We replace $a\hat{\mathbf{n}}$ by $r\hat{\mathbf{n}} = \mathbf{x}$ to obtain:

$$G_{\rm D}(\mathbf{x},\mathbf{x}') \;=\; -\,\frac{1}{4\pi}\,\left[\frac{1}{|\mathbf{x}-\mathbf{x}'|} \;-\; \frac{1}{|(r'/a)\mathbf{x} - (a/r')\mathbf{x}'|}\right]$$

If this is evaluated in spherical coordinates centered on the sphere, the angle γ between $\mathbf{x}$ amd $\mathbf{x}'$ is, from spherical trigonometry: $\cos\gamma \;=\; \cos\theta\,\cos\theta' + \sin\theta\,\sin\theta'\,\cos(\phi-\phi')$, and Green's function becomes:

$$G_{\rm D}(\mathbf{x},\mathbf{x}') \;=\; -\,\frac{1}{4\pi}\,\left[\frac{1}{\sqrt{r^2+r'^2-2rr'\cos\gamma}} \;-\; \frac{1}{\sqrt{r^2r'^2/a^2+a^2-2rr'\cos\gamma}}\right] \tag{6.47}$$

In this form it is most easy to see that $G_{\rm D}(a\hat{\mathbf{n}},\mathbf{x}') = G_{\rm D}(\mathbf{x},a\hat{\mathbf{n}}') = 0$, as desired. The Dirichlet Green function we have found is valid for *any* sphere since it does not care about which particular B.C. is specified for $\Psi(\mathbf{x})$ on the sphere. When $\Psi(r'=a) = 0$, the surface integral in eq. (6.40) vanishes; the volume integral remains the same since that is independent of the B.C. for Ψ. If $\Psi(r'=a) \neq 0$, we must evaluate $\partial_{n'}G_{\rm D}$. In spherical coordinates, this is:

$$\frac{\partial G_{\rm D}}{\partial n'} \;=\; \pm\,\frac{\partial G_{\rm D}}{\partial r'}\Big|_{r'=a} \;=\; \mp\,\frac{1}{4\pi}\,\frac{r^2-a^2}{a\,(r^2+a^2-2ar\cos\gamma)^{3/2}}$$

depending on whether $\mathrm{d}\mathbf{S}' \;=\; a^2\mathrm{d}\Omega'\hat{\mathbf{n}}'$, the normal to the surface which always points *out* of the volume, is in the direction of $\mathbf{x}'$ or in the opposite direction. For instance, the general solution of the inhomogeneous Laplace equation with B.C. specified on the surface $r = a$ for Ψ is:

$$\begin{aligned}\Psi(\mathbf{x}) \;=\; &\frac{1}{4\pi}\int F(\mathbf{x}')\,\left[\frac{1}{\sqrt{r^2r'^2/a^2+a^2-2rr'\cos\gamma}} \;-\; \frac{1}{\sqrt{r^2+r'^2-2rr'\cos\gamma}}\right]\,\mathrm{d}^3x' \\ &\mp\,\frac{1}{4\pi}\oint \Psi(r'=a)\,\frac{r^2-a^2}{a\,(r^2+a^2-2ar\cos\gamma)^{3/2}}\,\mathrm{d}S'\end{aligned} \tag{6.48}$$

where the $(-)$ sign refers to the solution for $r < a$ and the $(+)$ sign applies to $r > a$. In the latter case, there is an implicit assumption that the integrand, $\Psi\partial_{n'}G$, of the surface integral vanishes *at infinity* faster than $1/r'^2$. When $F(\mathbf{x}) = 0$ everywhere inside the volume where the solution is valid, we are left with $\nabla^2\Psi = 0$, and:

$$\Psi(\mathbf{x}) \;=\; \mp\,\frac{1}{4\pi}\oint \Psi(a,\theta',\phi')\,\frac{a(r^2-a^2)}{(r^2+a^2-2ar\,\cos\gamma)^{3/2}}\,\mathrm{d}\Omega' \tag{6.49}$$

Clearly, if $\Psi(a,\theta',\phi') \neq 0$ and $r > a$, $F(\mathbf{x}) \neq 0$ somewhere in the region $r < a$, and vice-versa.

6.6.4 An important expansion for Green's Functions in Spherical Coordinates

The angular dependence in the Green functions such as derived above is quite complicated and may well not yield a solution in closed form when integrated, so it is often sensible to use a expansion appropriate to the coordinate system selected for the problem. Indeed, let us do this for the Laplacian in spherical coordinates.

In spherical coordinates, Green functions for the Laplacian operator all satisfy:

$$\begin{aligned}\nabla^2_{\mathbf{x}} G(\mathbf{x},\mathbf{x}') &= \delta(\mathbf{x}-\mathbf{x}') \\ &= \frac{1}{r^2}\delta(r-r')\sum_{l=0}^{\infty}\sum_{m=-l}^{l} Y^*_{lm}(\theta',\phi')Y_{lm}(\theta,\phi)\end{aligned} \tag{6.50}$$

where the completeness relation (5.53) for spherical harmonics has been invoked.

We shall look for an expansion over separable terms of the form:

$$G(\mathbf{x},\mathbf{x}') = \sum_{l=0}^{\infty}\sum_{m=-l}^{l} g_l(r,r')\, Y^*_{lm}(\theta',\phi')Y_{lm}(\theta,\phi)$$

Inserting into eq. (6.50), we immediately find that $g_l(r,r')$ must satisfy the radial equation:

$$r^2\,\nabla^2_r g_l(r,r') = \mathrm{d}_r\big[r^2\,\mathrm{d}_r g_l(r,r')\big] - l(l+1)\,g_l(r,r') = \delta(r-r')$$

We now find ourselves in the familiar territory of 1-dim Green-function problems and Sturm-Liouville operators. For instance, we can connect with eq. (6.13) for a 1-dim Dirichlet Green function. We take $t_0 = a$ and $t_1 = b$, corresponding to two concentric spheres of radius a and b,

We have $\alpha(r') = r'^2$ and, with $f_1 = r^l$ and $f_2 = r^{-(l+1)}$, $W(r') = -(2l+1)/r'^2$. Also, let $r_< \equiv \min(r,r')$ and $r_> \equiv \max(r,r')$. It takes only a straightforward computation to arrive at (EXERCISE):

$$G_{\mathrm{D}}(\mathbf{x},\mathbf{x}') = \sum_{l=0}^{\infty}\sum_{m=-l}^{l}\frac{Y^*_{lm}(\theta',\phi')Y_{lm}(\theta,\phi)}{(2l+1)\,[1-(a/b)^{2l+1}]}\left(r_<^l - \frac{a^{2l+1}}{r_<^{l+1}}\right)\left(\frac{r_>^l}{b^{2l+1}} - \frac{1}{r_>^{l+1}}\right) \tag{6.51}$$

Inspection of the last two factors shows that this expression vanishes at $r = a$ and $r = b$ (and when $r' = a$ or $r' = b$), as it should. We did not have to require this since it is built in the derivation of the 1-dim Dirichlet Green function. Two important cases:

$$G_{\mathrm{D}}(\mathbf{x},\mathbf{x}') = \sum_{l=0}^{\infty}\sum_{m=-l}^{l}\frac{Y^*_{lm}(\theta',\phi')Y_{lm}(\theta,\phi)}{(2l+1)}\,r_<^l\left(\frac{r_>^l}{b^{2l+1}} - \frac{1}{r_>^{l+1}}\right) \qquad (a=0) \tag{6.52}$$

$$G_{\mathrm{D}}(\mathbf{x},\mathbf{x}') = \sum_{l=0}^{\infty}\sum_{m=-l}^{l}\frac{Y^*_{lm}(\theta',\phi')Y_{lm}(\theta,\phi)}{(2l+1)}\frac{1}{r_>^{l+1}}\left(\frac{a^{2l+1}}{r_<^{l+1}} - r_<^l\right) \qquad (b\to\infty) \tag{6.53}$$

The first expression gives the Green function inside a sphere of radius b; the second one, outside a sphere of radius a and all the way to infinity. When there are no boundary surfaces, we obtain over all space:

$$G(\mathbf{x},\mathbf{x}') = -\sum_{l=0}^{\infty}\sum_{m=-l}^{l}\frac{1}{2l+1}\frac{r_<^l}{r_>^{l+1}}\,Y^*_{lm}(\theta',\phi')Y_{lm}(\theta,\phi) \tag{6.54}$$

Comparing with eq. (6.42) with $\kappa = 0$ also yields a useful expansion of the ubiquitous distance factor $1/|\mathbf{x}-\mathbf{x}'|$.

When $0 \le r \le b$ (interior case) we can immediately rewrite (EXERCISE) the surface integral in eq. (6.40) as:

$$\sum_{l=0}^{\infty}\sum_{m=-l}^{l}\left[\int \Psi(b,\theta',\phi')\,Y^*_{lm}(\theta',\phi')\,\mathrm{d}\Omega'\right]\left(\frac{r}{b}\right)^l Y_{lm}(\theta,\phi)$$

where $\Psi(b,\theta',\phi')$ is specified on the surface $r=b$. The normal derivative of the Green function on the surface, $\partial G/\partial n' = \partial G/\partial r'\big|_{r'=b}$, has been evaluated for $r_< = r$ and $r_> = r'$ since $r < r' = b$. Also, the surface element on a sphere of radius b is $dS' = b^2 \mathrm{d}\Omega'$. This expression is still rather complicated, but it simplifies considerably if $\Psi(b,\theta',\phi')$ exhibits a symmetry (eg. azimuthal). Also, if one can write $\Psi(b,\theta',\phi')$ as a linear combination of spherical harmonics, the angular integration becomes trivial due to the orthonormality of the harmonics, and only a few terms in the sums might contribute.

6.6.5 An Elliptic Problem with a Twist: the Time-independent Schrödinger Equation

The time-independent Schrödinger equation (TISE) for a potential $V(\mathbf{x})$ takes the following suggestive form:

$$(\nabla^2 + k^2)\,\psi(\mathbf{x}) = \frac{2m}{\hbar^2}\,V(\mathbf{x})\,\psi(\mathbf{x}) \tag{6.55}$$

where $k^2 = 2mE/\hbar^2$. Although the right-hand side is not inhomogeneous, this in no way invalidates our previous results.

For bound states ($E<0$) of an attractive potential, $k^2 = \lambda < 0$, and we have the integral equation:

$$\psi(\mathbf{x}) = -\frac{m}{2\pi\hbar^2}\int \frac{\mathrm{e}^{-\kappa|\mathbf{x}-\mathbf{x}'|}}{|\mathbf{x}-\mathbf{x}'|}\,V(\mathbf{x}')\,\psi(\mathbf{x}')\,\mathrm{d}^3x'$$

with $\kappa^2 = -2mE/\hbar^2$. A somewhat simpler integral expression may be derived by writing the left-hand side of eq. (6.55) as its Fourier representation, and viewing $V(x)\psi(x)$ as the Fourier transform of the convolution $[V*\psi](\mathbf{k})$ (see section 5.5.4):

$$-\int (k^2+\kappa^2)\psi(\mathbf{k})\,\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{x}}\,\mathrm{d}^3k = \frac{2m}{\hbar^2}\int\left[\frac{1}{(2\pi)^{3/2}}\int V(\mathbf{k}-\mathbf{k}')\psi(\mathbf{k}')\,\mathrm{d}^3k'\right]\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{x}}\,\mathrm{d}^3k$$

Extracting $\psi(\mathbf{k})$, there comes:

$$\psi(\mathbf{k}) = -\frac{2m}{(2\pi)^{3/2}\hbar^2}\int \frac{V(\mathbf{k}-\mathbf{k}')\psi(\mathbf{k}')}{k^2+\kappa^2}\,\mathrm{d}^3k'$$

See p. BF414 for more details and an application to the Yukawa potential.

For unbound states ($E>0$), $k^2 = \lambda > 0$, and we can immediately write the **Lippmann-Schwinger equation**:

$$\psi_q^{(\pm)}(\mathbf{x}) = \frac{A}{(2\pi)^{3/2}}\,\mathrm{e}^{\mathrm{i}\mathbf{q}\cdot\mathbf{x}} - \frac{m}{2\pi\hbar^2}\int \frac{\mathrm{e}^{\pm\mathrm{i}q|\mathbf{x}-\mathbf{x}'|}}{|\mathbf{x}-\mathbf{x}'|}\,V(\mathbf{x}')\,\psi_q^{(\pm)}(\mathbf{x}')\,\mathrm{d}^3x' \tag{6.56}$$

with $q = \sqrt{2mE/\hbar^2}$.

The asymptotic form of the Lippmann-Schwinger equation is of particular interest. When $r >> r'$, we can expand $|\mathbf{x}-\mathbf{x}'| = \sqrt{r^2 - 2\mathbf{x}\cdot\mathbf{x}' + r'^2} \approx r - \hat{\mathbf{n}}\cdot\mathbf{x}'$, with $\hat{\mathbf{n}} = \mathbf{x}/r$. Inserting into the integral equation yields:

$$\begin{aligned}\psi_{\mathbf{q}}^{(\pm)}(\mathbf{x}) &\underset{r\to\infty}{=} \frac{A}{(2\pi)^{3/2}}\,\mathrm{e}^{\mathrm{i}\mathbf{q}\cdot\mathbf{x}} - \frac{m}{2\pi\hbar^2}\,\frac{\mathrm{e}^{\pm\mathrm{i}qr}}{r}\int \mathrm{e}^{\mp\mathrm{i}q\hat{\mathbf{n}}\cdot\mathbf{x}'}\,V(\mathbf{x}')\,\psi^{(\pm)}(\mathbf{x}')\,\mathrm{d}^3x' \\ &= \frac{A}{(2\pi)^{3/2}}\left[\mathrm{e}^{\mathrm{i}\mathbf{q}\cdot\mathbf{x}} + f_\pm(\mathbf{q})\,\frac{\mathrm{e}^{\pm\mathrm{i}qr}}{r}\right]\end{aligned}$$

This expression represents the spatial dependence of a superposition of a plane wave and a **scattered** spherical wave propagagating inward or outward from the origin. The function $f_\pm(\mathbf{q})$ is called the **scattering amplitude**; it also obeys an integral equation, eq. BF7.75, and its square modulus is directly related to experimental data.

6.7 A Hyperbolic Problem: the d'Alembertian Operator

With the Fourier integral representation (note the different normalisation and sign in the exponentials!):

$$\begin{aligned}\Psi(\mathbf{x},t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\Psi(\mathbf{x},\omega)\,\mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{d}\omega \\ \Psi(\mathbf{x},\omega) &= \int_{-\infty}^{\infty}\Psi(\mathbf{x},t)\,\mathrm{e}^{\mathrm{i}\omega t}\,\mathrm{d}t\end{aligned} \tag{6.57}$$

we can transform a typical inhomogeneous wave equation:

$$\Box\Psi(\mathbf{x},t) = \nabla^2\Psi(\mathbf{x},t) - \frac{1}{c^2}\,\partial_t^2\Psi(\mathbf{x},t) = F(\mathbf{x},t)$$

where $F(\mathbf{x},t)$ is a known source, to its Helmholtz form:

$$(\nabla^2 + k^2)\Psi(\mathbf{x},\omega) = F(\mathbf{x},\omega) \tag{6.58}$$

where $k^2 > 0$ can be taken as a short form for $(\omega/c)^2$.

Just as for the Laplacian operator, there exist Green functions for $\nabla^2 + k^2$; we have found them a little earlier in eq. (6.43):

$$G^{(\pm)}(R) = -\frac{1}{4\pi}\,\frac{\mathrm{e}^{\pm\mathrm{i}kR}}{R} \tag{6.59}$$

where $R = |\mathbf{x}-\mathbf{x}'|$.

Now we are ready to derive the full Green functions for the d'Alembertian operator, which satisfy:

$$\Box_{\mathbf{x}}G(\mathbf{x},t;\mathbf{x}',t') = \delta(\mathbf{x}-\mathbf{x}')\,\delta(t-t') \tag{6.60}$$

or, in the frequency domain:

$$(\nabla_x^2 + k^2)G(\mathbf{x},\mathbf{x}',\omega,t') = \delta(\mathbf{x}-\mathbf{x}')\,\mathrm{e}^{\mathrm{i}\omega t'}$$

Assume separable solutions of the form $G(\mathbf{x},\mathbf{x}')\mathrm{e}^{\mathrm{i}\omega t'}$; inserting into this equation, we get from (6.58) the solutions $G^{\pm}(\mathbf{x},\mathbf{x}',\omega,t') = -\mathrm{e}^{\mathrm{i}(\pm kR+\omega t')}/4\pi R$. Then, transforming back to the time domain and using the representation (5.16) for the δ-function yields the Green functions:

$$G^{(\pm)}(\mathbf{x},t;\mathbf{x}',t') = -\frac{1}{8\pi^2 R}\int_{-\infty}^{\infty}\mathrm{e}^{\mathrm{i}\omega[\pm R/c+(t'-t)]}\,\mathrm{d}\omega = -\frac{1}{4\pi}\,\frac{\delta\big(t' - [t\mp R/c]\big)}{R} \tag{6.61}$$

In higher dimensions, Green functions can contain δ-functions and so may not be actual *functions*!

Using eq. (6.60), we also recognise that:

$$\Box_{\mathbf{x}}\int_{\text{allspace}}\mathrm{d}^3x'\int_{-\infty}^{\infty}G^{(\pm)}(\mathbf{x},t;\mathbf{x}',t')\,F(\mathbf{x}',t')\,\mathrm{d}t' = \int\mathrm{d}^3x'\int_{-\infty}^{\infty}F(\mathbf{x}',t')\,\Box_{\mathbf{x}}G^{(\pm)}(\mathbf{x},t;\mathbf{x}',t')\mathrm{d}t' = F(\mathbf{x},t)$$

has the generic form $\Box\Psi(\mathbf{x},t) = F(\mathbf{x},t)$, which shows that the general solution of a wave equation with sources can be written either as the **retarded** solution:

$$\begin{aligned}\Psi(\mathbf{x},t) &= \Psi_{\text{in}}(\mathbf{x},t) + \int\int_{-\infty}^{\infty}G^{(+)}(\mathbf{x},t;\mathbf{x}',t')\,F(\mathbf{x}',t')\,\mathrm{d}^3x'\,\mathrm{d}t' \\ &= \Psi_{\text{in}}(\mathbf{x},t) - \frac{1}{4\pi}\int\frac{F(\mathbf{x}',t_{\text{ret}})}{|\mathbf{x}-\mathbf{x}'|_{\text{ret}}}\,\mathrm{d}^3x'\end{aligned} \tag{6.62}$$

or, equivalently, as the **advanced** solution:

$$\begin{aligned}\Psi(\mathbf{x},t) &= \Psi_{\text{out}}(\mathbf{x},t) + \int\int_{-\infty}^{\infty}G^{(-)}(\mathbf{x},t;\mathbf{x}',t')\,F(\mathbf{x}',t')\,\mathrm{d}^3x'\,\mathrm{d}t' \\ &= \Psi_{\text{out}}(\mathbf{x},t) - \frac{1}{4\pi}\int\frac{F(\mathbf{x}',t_{\text{adv}})}{|\mathbf{x}-\mathbf{x}'|_{\text{adv}}}\,\mathrm{d}^3x'\end{aligned} \tag{6.63}$$

where the suffixes *ret* and *adv* stand for the fact that t' must be evaluated at the **retarded time** $t_{\rm ret} = t - R/c$, or the **advanced time** $t_{\rm adv} = t + R/c$. This ensures the proper causal behaviour of the solutions, in the sense that, eg., the solution at time t is only influenced by the behaviour of the source point $\mathbf{x}'$ at time $t - R/c$. $\Psi_{\rm in}$ and $\Psi_{\rm out}$ are possible plane-wave solutions of the *homogeneous* wave equation for Ψ. Most often they can be taken to be zero.

6.8 Initial Value Problem with Constraints

The Initial Value Problem (IVP) consists in finding which data must be specified at a given time for the time evolution of variables to be uniquely determined by their equations of "motion".

By **initial data**, one means the state of the system of variables and their first-order derivatives on a three-dimensional spacelike hypersurface; usually, this means at some time t_0 everywhere in space. The IVP together with the evolution equations constitute the **Cauchy Problem** of the theory. If the IVP can be solved, the dynamical behaviour of the system can be uniquely predicted from its initial data.

Most often, the equations of "motion" take the form of a set of wave equations, each of the form $\Box f = F$. If they always told the whole story, the Cauchy problem would be solved by specifying the value of f and its first-order time derivatives at $t = t_0$. Things are not so simple, however, when there are inherent, built-in **constraints** on the initial data. Those constraint equations must be discovered and solved. Also, we must find which initial data we are allowed to specify freely.

We study in some depth a very important example: Maxwell's theory. In linear, unpolarised and unmagnetised media, Maxwell's equations are:

$$\begin{aligned} \boldsymbol{\nabla}\cdot\mathbf{E} &= 4\pi k_{\rm e}\,\rho & \boldsymbol{\nabla}\times\mathbf{B} - \frac{k_{\rm m}}{k_{\rm e}}\,\partial_t\mathbf{E} &= 4\pi k_{\rm m}\,\mathbf{J} \\ \boldsymbol{\nabla}\cdot\mathbf{B} &= 0 & \boldsymbol{\nabla}\times\mathbf{E} + \partial_t\mathbf{B} &= 0 \end{aligned} \tag{6.64}$$

where $k_{\rm e}$ and $k_{\rm m}$ are constants that depend on the system of units, and $k_{\rm e}/k_{\rm m} = c^2$, with c the speed of light. The source terms ρ and $\mathbf{J}$ are not independent; indeed, one derives from eq. (6.64) a continuity equation:

$$\partial_t\rho = \frac{1}{4\pi k_{\rm e}}\boldsymbol{\nabla}\cdot\partial_t\mathbf{E} = -\boldsymbol{\nabla}\cdot\mathbf{J} \tag{6.65}$$

The two homogeneous equations are equivalent to:

$$\mathbf{E} = -\partial_t\mathbf{A} - \boldsymbol{\nabla}\Phi \qquad \mathbf{B} = \boldsymbol{\nabla}\times\mathbf{A} \tag{6.66}$$

Observe that $\mathbf{A}$ is determined only up to a term $\boldsymbol{\nabla} f$, where $f(\mathbf{x},t)$ is an arbitrary differentiable function. In fact, if we perform the **gauge transformations** $\Phi \to \Phi - \partial_t f$ and $\mathbf{A} \to \mathbf{A} + \boldsymbol{\nabla} f$, neither $\mathbf{E}$ nor $\mathbf{B}$ change! We say that Maxwell's theory is **gauge-invariant**.

The inhomogeneous Maxwell equations (6.64) become *second-order* equations for Φ and $\mathbf{A}$:

$$\begin{aligned} \nabla^2\Phi + \partial_t(\boldsymbol{\nabla}\cdot\mathbf{A}) &= -4\pi\,k_{\rm e}\rho \\ \Box\mathbf{A} - \boldsymbol{\nabla}\left(\boldsymbol{\nabla}\cdot\mathbf{A} + \frac{1}{c^2}\partial_t\Phi\right) &= -4\pi\,k_{\rm m}\,\mathbf{J} \end{aligned} \tag{6.67}$$

.

6.8.1 Second-order Cauchy problem using transverse/longitudinal projections

While eq. (6.67) are gauge-invariant, $\mathbf{A}$ and Φ themselves are not, at least at first sight. What this means is that the time-evolution of at least some of the four quantities Φ and $\mathbf{A}$ cannot be uniquely determined from their initial conditions and eq. (6.67) since we can always perform an arbitrary gauge transformation on them at some arbitrary

later time t, as often as we wish. This is a serious issue which must be understood and addressed if Φ and $\mathbf{A}$ are to be of any use at all.

One instructive approach is to note that according to the **Helmholtz theorem** any differentiable 3-dim vector field that goes to zero at infinity faster than $1/r$ may be written as the sum of two vectors:

$$\mathbf{A} \;=\; \underbrace{\boldsymbol{\nabla} u}_{\mathbf{A}_{\rm L}} \;+\; \underbrace{\boldsymbol{\nabla}\times\mathbf{w}}_{\mathbf{A}_{\rm T}}$$

The first term, $\mathbf{A}_{\rm L} = \boldsymbol{\nabla} u$, whose curl vanishes *identically*, is called the **longitudinal** part (or projection) of $\mathbf{A}$; the second, $\mathbf{A}_{\rm T} = \boldsymbol{\nabla}\times\mathbf{w}$, whose divergence vanishes *identically*, is called the **transverse** part (or projection) of $\mathbf{A}$. Since the longitudinal and transverse projections are perpendicular to each other, we can decompose (project out) Maxwell's equations for the fields $\mathbf{E}$ and $\mathbf{B}$ and for the potential $\mathbf{A}$ into longitudinal and tranverse parts.

Before we do this, however, we note that since by definition $\boldsymbol{\nabla}\cdot\mathbf{J}_{\rm T} = 0$, the continuity equation for the sources does not involve the transverse part of $\mathbf{J}$. Also, $\boldsymbol{\nabla}\cdot\mathbf{A}$ is really $\boldsymbol{\nabla}\cdot\mathbf{A}_{\rm L}$ and contains no information about $\mathbf{A}_{\rm T}$

Project the second equation (6.67). The transverse projection immediately gives:

$$\Box\,\mathbf{A}_{\rm T} \;=\; -\,4\pi\,k_{\rm m}\mathbf{J}_{\rm T} \tag{6.68}$$

where we have used the fact a gradient is a longitudinal object. The two transverse components $\mathbf{A}_{\rm T}$ satisfy a proper wave equation and correspond to physically observable quantities, in the sense that being transverse, they are *unaffected* by $\mathbf{A}\to\mathbf{A}+\boldsymbol{\nabla} f$, which can change only the longitudinal component $\mathbf{A}_{\rm L}$. *Therefore, the time evolution of the two transverse* $\mathbf{A}_{\rm T}$ *is not arbitrary and they have a well-posed Cauchy problem.*

Now, remembering that $\Box = \nabla^2 - (\partial_t^2)/c^2$, take the divergence of the longitudinal projection of (6.67):

$$\begin{aligned}\boldsymbol{\nabla}\cdot\left[\Box\mathbf{A}_{\rm L} \;-\; \boldsymbol{\nabla}\left(\boldsymbol{\nabla}\cdot\mathbf{A}_{\rm L} \;+\; \frac{1}{c^2}\partial_t\Phi\right) \;+\; 4\pi\,k_{\rm m}\,\mathbf{J}_{\rm L}\right] \;&=\; \Box(\boldsymbol{\nabla}\cdot\mathbf{A}_{\rm L}) \;-\; \nabla^2(\boldsymbol{\nabla}\cdot\mathbf{A}_{\rm L}) \;-\; \frac{\partial_t\nabla^2\Phi}{c^2} \;+\; 4\pi\,k_{\rm m}\boldsymbol{\nabla}\cdot\mathbf{J}_{\rm L}\\ &=\; -\frac{1}{c^2}\,\partial_t\left[\partial_t(\boldsymbol{\nabla}\cdot\mathbf{A}_{\rm L}) \;+\; \nabla^2\Phi \;+\; 4\pi\,k_{\rm e}\rho\right]\end{aligned}$$

where eq. (6.65) has been invoked in the second line. But the terms in the square bracket on that line are just the first of equations (6.67). Therefore, the longitudinal projection of the second Maxwell equation for the 3-vector potential contains no information about $\boldsymbol{\nabla}\cdot\mathbf{A}$ that is not in the first equation. But that is really an equation for Φ with $\boldsymbol{\nabla}\cdot\dot{\mathbf{A}}$ (more precisely, $\boldsymbol{\nabla}\cdot\dot{\mathbf{A}}_{\rm L}$) as a source together with ρ. *Therefore, Maxwell's theory cannot uniquely determine the time evolution of the divergence of the* 3*-vector potential.* Nor can it uniquely determine the time evolution of Φ, since Φ is gauge-variant. Systems whose time-evolution involves arbitrary functions are often called **singular**.

6.8.2 Choices for the divergence of A

Since the theory does not know $\boldsymbol{\nabla}\cdot\mathbf{A}$ (and its first-order time derivative for that matter), *we* have to tell it what it is by making an arbitrary choice. If we choose $\boldsymbol{\nabla}\cdot\mathbf{A}$ to vanish (**Coulomb condition**), the vector potential becomes pure transverse (hence the name "transverse gauge also given to this choice), and the equation for Φ becomes a Poisson-type equation with solution:

$$\Phi(\mathbf{x},t) \;=\; k_{\rm e}\int\frac{\rho(\mathbf{x}',t)}{|\mathbf{x}-\mathbf{x}'|}\,{\rm d}^3x' \tag{6.69}$$

This looks innocuous enough until we realise that any change in the source is *instantaneously* reflected in the scalar potential. The Coulomb condition leads to acausal behaviour, which is also a reflection of the fact that the condition is not relativistically covariant, in the sense that it is not necessarily the same in all inertial frames. But the equation for Φ is not a classical wave equation, and Φ does not really propagate as a wave, so one should not expect proper causal behaviour from it.

The problem is *seemingly* resolved just by choosing instead the Lorenz condition: $\boldsymbol{\nabla}\cdot\mathbf{A} = -\partial_t\Phi/c^2$, which turns eq. (6.67) into standard wave equations of the type $\Box(\text{potential}) = \text{source}$ with causal solution eq. (6.62). Then one can calculate the energy radiated to infinity following standard treatments (eg. chapter 10 in Griffiths or chapter 14 in Jackson) and find that the scalar potential does make a *mathematical* contribution to the energy radiated to infinity.. This, however, would not have happened if we had chosen the Coulomb condition. Therefore, we should not attach any physical significance to that contribution: it arises simply out of consistency with this particular choice of condition on $\boldsymbol{\nabla}\cdot\mathbf{A}$.

Conditions that do not continue to hold automatically in the future must be enforced by hand at all times. This could be the case for conditions such as the ones *we* have imposed. Fortunately, one can show that the Coulomb and Lorenz conditions propagate forward in time once imposed at initial time. Take for instance the wave equation for $\mathbf{A}$ in eq. (6.67). Impose $\boldsymbol{\nabla}\cdot\mathbf{A} = 0$ and $\partial_t(\boldsymbol{\nabla}\cdot\mathbf{A}) = 0$ at some time in both the equations for Φ and $\mathbf{A}$. Then take the divergence of the resulting wave equation for $\mathbf{A}$ and the time derivative of the resulting equation for Φ, and use the continuity equation to obtain:

$$\Box(\boldsymbol{\nabla}\cdot\mathbf{A}) \;=\; 0$$

showing that if we choose $\boldsymbol{\nabla}\cdot\mathbf{A} = 0$ and $\partial_t(\boldsymbol{\nabla}\cdot\mathbf{A}) = 0$ at $t = t_0$, it remains the same for all time. Similarly, it is trivial to show that imposing the Lorenz condition everywhere in space at $t = t_0$ also leads to a wave equation for $\boldsymbol{\nabla}\cdot\mathbf{A}$.

Note also that the energy radiated to infinity can be calculated solely in terms of $\mathbf{B}$ and, therefore, of $\mathbf{A}$, without Φ being involved. In fact, since $\mathbf{B} = \boldsymbol{\nabla}\times(\mathbf{A}_{\rm L} + \mathbf{A}_{\rm T}) = \boldsymbol{\nabla}\times\mathbf{A}_{\rm T}$, only the *two transverse* components of $\mathbf{A}$ contribute! These are independent of any choice, and thus entirely physical (contrary to assertions sometimes made that the electromagnetic potential is not as physical as the fields because it is not gauge-invariant—now we know that this only applies to $\mathbf{A}_{\rm L}$ and Φ).

6.8.3 First-order Cauchy problem

Now consider this same Cauchy Problem from the point of view of the fields $\mathbf{E}$ and $\mathbf{B}$. Taking the curl of the *first-order* curl equations (6.64), we arrive at:

$$\begin{aligned}\Box\mathbf{E} \;&=\; 4\pi k_{\rm e}\,\boldsymbol{\nabla}\rho \;+\; 4\pi k_{\rm m}\,\partial_t\mathbf{J}\\ \Box\mathbf{B} \;&=\; -\,4\pi\,k_{\rm m}\,\boldsymbol{\nabla}\times\mathbf{J}\end{aligned} \tag{6.70}$$

These look like wave equations for six quantities. *But only those of their solutions which also satisfy the first-order field equations (6.64), including at $t = t_0$, are acceptable.*

The two first-order divergence equations contain no time derivatives and are thus constraints on $\mathbf{E}$ and $\mathbf{B}$ at $t = t_0$. The constraint equation on $\mathbf{E}$ can be rewritten $\nabla^2 u = \rho$, a Poisson-type equation which can be solved for u *at initial time* so long as ρ falls off faster than $1/r^2$ at infinity). In the case of $\mathbf{B}$, the scalar field u satisfies a Laplace equation *everywhere* and is therefore zero. So $\mathbf{B}$ has no longitudinal component, only a transverse one just as we had found at the end of the previous section. In both cases, the longitudinal component is either zero or can be *solved* for at t_0, so cannot be freely specified.

Now look at the two first-order Maxwell field equations (6.64) which contain time derivatives. Suppose we specify $\mathbf{E}$ and $\partial_t\mathbf{E}$ at $t = t_0$, which are needed to solve the $2^{\rm nd}$-order equations, eq. (6.70). Then the two transverse components of $\mathbf{B}$ are determined by $\boldsymbol{\nabla}\times\mathbf{B} = 4\pi k_{\rm m}\mathbf{J} + \partial_t\mathbf{E}/c^2$; $\partial_t\mathbf{B}$ is determined, also at $t = t_0$, by the curl equation for $\mathbf{E}$. Therefore, once we have specified the two transverse components of $\mathbf{E}$ and their time derivatives, Maxwell's first-order equations take over and determine the others at $t = t_0$. Alternatively, we could have started with the two transverse components of $\mathbf{B}$; specifying them and their time derivatives at $t = t_0$ constrains all the other field components and time derivatives.

You can also use (exercise) the transverse/longitudinal projections of the first-order equations (6.64) to show that in source-free space, only the transverse components of $\mathbf{E}$ and $\mathbf{B}$ obey a classical wave equation.

Thus, the results of the first-order Cauchy-data analysis are fully consistent with the second-order analysis on $\mathbf{A}$: only two transverse components correspond to *independent*, physical dynamical degrees of freedom, Also, one

of the advantages of this Cauchy analysis is that it does not rely on some particular solution, but is valid for *any* electromagnetic field and potential.

Addendum: The Lorenz condition $\boldsymbol{\nabla}\cdot\mathbf{A} = -\partial_t\Phi/c^2$, which is imposed in almost all treatments of electromagnetic radiation, could lead you to believe that Φ and the three components of $\mathbf{A}$ propagate to infinity, whereas I hope to have convinced you that only the transverse components of $\mathbf{A}$ do.

The Lorenz condition relates the longitudinal component of $\mathbf{A}$ to Φ. Now I will show that $\mathbf{A}_{\rm L}$ can be made to disappear without affecting Maxwell's equations for the fields *and* the potentials.

The key observation is that one can change both $\mathbf{A}$ and Φ to new functions that still obey the Lorenz condition. Indeed, let f be some scalar function that satisfies the homogeneous wave equation $\Box f = \nabla^2 f - \frac{1}{c^2}\partial_t^2 f = 0$. Then add $\nabla^2 f$ to $\boldsymbol{\nabla}\cdot\mathbf{A}$ and $\frac{1}{c^2}\partial_t^2 f$ to $-\partial_t\Phi/c^2$ to obtain:

$$\boldsymbol{\nabla}\cdot(\mathbf{A} + \boldsymbol{\nabla} f) = -\frac{1}{c^2}\partial_t(\Phi - \partial_t f) \tag{6.71}$$

This shows that gauge-transformed potentials *still satisfy the Lorenz condition*! As noted before, it is important to keep in mind that since the transformation shifts $\mathbf{A}$ by a gradient, which is a longitudinal object, *it does not affect the transverse components of* $\mathbf{A}$.

Now, for the first time, we shall have to look at actual solutions of the wave equations for $\mathbf{A}$ and Φ. To make things as simple as possible, take plane-wave solutions $\mathbf{A} = \mathbf{A}_0 \mathrm{e}^{\mathrm{i}(kx-\omega t)}$, where the x-axis has been aligned along the direction of propagation, and $\Phi = \Phi_0 \mathrm{e}^{\mathrm{i}(kx-\omega t)}$. Then:

$$\boldsymbol{\nabla}\cdot\mathbf{A} = \partial_x A_x = \mathrm{i}kA_{0x}\mathrm{e}^{\mathrm{i}(kx-\omega t)}, \qquad \partial_t\Phi = -\mathrm{i}\omega\Phi_0\mathrm{e}^{\mathrm{i}(kx-\omega t)}$$

Inserting into the Lorenz condition with $\omega/k = c$ yields, as expected, a relation between the longitudinal component A_x and Φ: $A_{0x} = \Phi_0/c$.

Now fold in $f = f_0\mathrm{e}^{\mathrm{i}(kx-\omega t)}$ into eq. (6.71) for the gauge-transformed potentials, to get:

$$\mathrm{i}k\,(A_{0x} + \mathrm{i}k\,f_0)\mathrm{e}^{\mathrm{i}(kx-\omega t)} = \mathrm{i}\,\frac{\omega}{c^2}\,(\Phi_0 + \mathrm{i}\,\omega\,f_0)\,\mathrm{e}^{\mathrm{i}(kx-\omega t)}$$

Since f_0 is arbitrary, we can choose it to cancel A_{0x}, which at the same time gets rid of Φ_0, leaving us with the transverse components of $\mathbf{A}$ only!

Although the analysis under the Lorenz condition is quite a bit more involved than with the Coulomb condition, the conclusions are the same: only the two transverse components of $\mathbf{A}$ propagate, in the sense that they carry energy to infinity.