Contents

1 CHAPTER I — TENSORS AND EXTERIOR CALCULUS ON MANIFOLDS

This chapter first introduces a view of vectors as objects that can be discussed without explicit reference to a basis (or cobasis), as is implicit in the elementary notation $\mathbf{u}$, or $\vec{u}$. An alternative description in terms of components in a so-called **dual** basis will be introduced and its meaning explored. Our powerful geometric approach will then allow a conceptually simple generalisation of vectors to the even more important tensors. For example, it is difficult to understand electromagnetism if one insists on thinking of the electric and magnetic fields as two vector fields connected by Maxwell equations, instead of the six non-zero components of the Faraday tensor, $\mathbf{F}$[†]. The need to describe how vectors and p-forms change in time and space will lead to the **exterior derivative**, of which gradient, divergence and curl are but special cases. We will also see that **differential** p-forms in fact are the only objects that can be meaningfully integrated. The concise language of p-forms can illuminate many other areas of physics, such as mechanics, thermodynamics, general relativity and quantum field theory.

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 a **linear**, uniquely invertible, commutative and associative map, $\mathcal{V} \times \mathcal{V} \longmapsto \mathcal{V}$, called addition, and another, $s \cdot \mathcal{V} \longmapsto \mathcal{V}$, called s-multiplication (multiplication by a number $s \in \mathbb{F}$), are defined. One single element must be the identity under addition. Thus, any two elements $\mathbf{u}$ and $\mathbf{v}$ of $\mathcal{V}$ satisfy:

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

$\forall\, a,\ b \in \mathbb{F}$; in what follows, $\mathbb{F} = \mathbb{R}$. We will call elements of a vector space **vectors**.

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, s-multiplication by $\lambda \in \mathbb{R}$ defined as multiplying each entry by λ, and a "null" vector under addition specified, is perhaps the best-known vector space.

Definition 1.2. If there exists a finite subset, $\{\mathbf{e}_\alpha \in \mathcal{V}\}$, such that any $\mathbf{v} \in \mathcal{V}$ can be written as a unique linear combination[‡]:

$$\mathbf{v} \;=\; \sum_{\alpha}^{n<\infty} v^\alpha\,\mathbf{e}_\alpha \;\equiv\; v^\alpha\,\mathbf{e}_\alpha \qquad \text{summation over repeated indices implied!} \tag{1.1}$$

then that set is said to **span** the vector space $\mathcal{V}$.

If, furthermore, this set is **linearly independent**, in the sense that $\mathbf{v} = 0 \Longrightarrow v^\alpha = 0$, then it is a **basis** of $\mathcal{V}$. The number n of vectors in a basis defines the dimension of $\mathcal{V}$, and we often write $\mathcal{V}^n$.

The (real, and unique!) coefficients v^α are called the **contravariant components** of the vector $\mathbf{v}$ in this basis. This one-to-one *correspondence* between $\mathcal{V}^n$ and $\mathbb{R}^n$ 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, the index on $\mathbf{e}_\alpha$ identifies the *vector*, not a component of the vector.

Example 1.2. 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. Adding components in this basis, and only in this basis, defines vector addition.

Example 1.3. The polynomials $p_n(x) = \sum^n a_i x^i$, with $(a_i, x) \in \mathbb{R}$, of degree $\leq n$, form the vector space $P_{\leq n}$. One possible basis is the set $\{1,\, x, \ldots,\, x^n\}$, which obviously spans P_n and whose elements are linearly independent, since $\sum^n c_i x^i = 0$ forces $c_i = 0$.

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

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

1.1.2 The space dual to a vector space

Let $\mathcal{V}$ and $\mathcal{W}$ be two vector spaces over the same field, which here we take to be $\mathbb{R}$. We shall be interested in the set of all **linear mappings**, $\mathrm{Hom}\,(\mathcal{V},\mathcal{W}) \equiv \mathcal{L}(\mathcal{V},\mathcal{W}) \coloneqq \{\mathbf{T} : \mathcal{V} \to \mathcal{W}\}$, such that, $\forall\, \mathbf{T}_i \in \mathcal{L}(\mathcal{V},\mathcal{W})$:

$$(a\,\mathbf{T}_i + \mathbf{T}_j)(\mathcal{V}) \;=\; a\,\mathbf{T}_i(\mathcal{V}) \;+\; \mathbf{T}_j(\mathcal{V}) \quad \boxed{\in \mathcal{W}} \qquad\qquad a\,\mathbf{T}_i + \mathbf{T}_j \in \mathcal{L}(\mathcal{V},\mathcal{W})$$

One can define linear mappings on $\mathcal{L}\,(\mathcal{V},\mathcal{W})$, ie., one can **compose** linear mappings into a linear map.

An important subset of the set of linear mappings is $\mathcal{L}(\mathcal{V},\,\mathbb{R})$ that contains all linear, real-valued functions on a vector space. We say that 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 called **covectors**, or **linear functionals** (in linear algebra), or **1-forms**. One example would be definite integrals on the vector space of polynomials.

There comes following important definition:

Definition 1.3. If $\{\mathbf{e}_\alpha\}$ is a basis of a space $\mathcal{V}^n$, its *unique* **dual basis** (**cobasis**) in $\mathcal{V}^*$, $\{\boldsymbol{\omega}^\alpha\}$, satisfies:

$$\boldsymbol{\omega}^\alpha(\mathbf{e}_\beta) \;\coloneqq\; \delta^\alpha{}_\beta \qquad (\alpha,\,\beta) = 1,\,\ldots,\,n \tag{1.2}$$

where $\delta^\alpha{}_\beta$ is the Kronecker delta. As an example, let $\mathbf{e}_\beta = x^\beta$ $(1 \le \beta \le n)$ be a basis in the space of polynomials of degree n. Then $\boldsymbol{\omega}^\alpha = (1/\alpha!)\partial^\alpha\big|_{x=0}$ is the corresponding cobasis of the dual space.

We write a covector as $\boldsymbol{\sigma} = \sigma_\alpha\boldsymbol{\omega}^\alpha$, where σ_α are the **covariant components** of $\boldsymbol{\sigma}$ in this dual basis.

From this we derive the action of an element $\boldsymbol{\omega}^\alpha$ of the cobasis of $\mathcal{V}^*$ on a vector $\mathbf{v} \in \mathcal{V}$:

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

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), $\boldsymbol{\sigma} = \sigma_\beta\boldsymbol{\omega}^\beta$, with σ_β the **covariant** components of $\boldsymbol{\sigma}$? Well,

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

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. Also, in $\boldsymbol{\sigma} = \sigma_\beta\boldsymbol{\omega}^\beta$ as well as 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 to be discussed later. Fluency in both is highly recommended.

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:

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

Therefore, we can also think of $\boldsymbol{\sigma}$ as 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 one-row matrix with components σ_α by a one-column matrix with components v^β:

$$\begin{pmatrix} v^1 & v^2 & \ldots & v^n \end{pmatrix}^{\mathrm{T}} \underset{\boldsymbol{\sigma}}{\longmapsto} \begin{pmatrix} \sigma_1 & \sigma_2 & \ldots & \sigma_n \end{pmatrix} \begin{pmatrix} v^1 & v^2 & \ldots & v^n \end{pmatrix}^{\mathrm{T}} \;=\; \sigma_\alpha\,v^\alpha \tag{1.3}$$

Since $\mathcal{L}(\mathcal{V},\,\mathbb{R})$, or $\mathcal{V}^*$, is a vector space, it has its own dual space, $\mathcal{L}(\mathcal{V}^*,\,\mathbb{R})$, or $\mathcal{V}^{**}$. We realise that nothing prevents us (*only with finite-dimensional spaces*) from viewing 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 $\mathbf{v}(\boldsymbol{\sigma}) = v^\alpha\sigma_\alpha$, exactly as in eq. (1.3) above.

These considerations suggest that just like a vector, we can view a 1-form (covector), as a kind of machine[‡] but with a *vector* as input and a number as output. The following tables summarise these results:

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

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

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

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

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 unique connection between $\boldsymbol{\sigma}$ and a vector in $\mathcal{V}$. So, tempting as it is to identify $\sigma_\alpha v^\alpha$ with the scalar product of two vectors, let us resist that urge.

For a given n-dimensional vector $\mathbf{v}$, there exists a unique set of parallel (n-1)-dimensional hyperplanes that can provide a pictorial representation of 1-forms. This is easiest when $n = 2$. Then $a = \sigma_1 v^1 + \sigma_2 v^2$ determines a perpendicular to $\mathbf{v}$ with equation $\sigma_2 = a/v^2 - \sigma_1 v^1/v^2$. The lines generated by different a all have slope $-v^1/v^2$.

1.2 Where Do Vectors Live?

The obvious answer has to be: in a vector space! Therefore, we should learn how to identify (or construct) such vector spaces. This is anything but trivial: by no means can all spaces be equipped with a vector-space structure.

We start by defining functions on a set of points called a **manifold**, M, with each point surrounded by an open neighbourhood and completely specified by *smooth*, real coordinate functions, or parameters, x^ν $(1 \leq \nu \leq n)$, n being the **dimension** of M^n (see Appendix A for a more rigorous treatment of the topics in this section).

Now the naïve notion of a vector as a straight arrow from one point to another, while acceptable in $\mathbb{R}^n$, cannot be extended to arbitrary manifolds, on which straightness has no meaning (try a straight arrow *on* a sphere). Manifolds are not vector spaces; so where do vectors at a point in M^n actually live? And could a vector involve only *that* point, independent of coordinate charts, instead of the two points involved in the arrow picture?

Through each point $\mathcal{P} \in M$ run an infinite set of **curves**, each a subset, $\Gamma_\lambda \in M$, described by a single parameter λ. Then, at any given $\mathcal{P}(\lambda_0)$ traversed by Γ_λ, we define a **velocity**, $\mathbf{v}_{(\Gamma_\lambda,\mathcal{P})}$, that maps smooth, real-valued functions at $\mathcal{P}$ to a real number:

$$\mathbf{v}_{(\Gamma,\mathcal{P})}(f) \coloneqq \mathrm{d}_\lambda f\big|_{\lambda_0} \qquad \big(\mathrm{d}_\lambda \coloneqq \mathrm{d}/\mathrm{d}\lambda\big) \tag{1.4}$$

Now consider our curve Γ as a one-dim region of a n-dim manifold M^n, described by n coordinate functions $x^\nu(\lambda)$. From the chain rule the velocity itself becomes:

$$\mathbf{v}_{(\Gamma_\lambda,\mathcal{P})} = \sum_\nu \mathrm{d}_\lambda x^\nu(\lambda)\big|_{\lambda_0} (\boldsymbol{\partial}_\nu)_\mathcal{P} \tag{1.5}$$

Of particular interest is the **coordinate curve**, Γ^ν_α, generated by varying one of the coordinate functions, say x^α, while holding all the others constant. We find that, up to a scale factor, $\boldsymbol{\partial}_\alpha$ is simply the velocity at $\mathcal{P}$ for Γ_α.

We assert that *velocities* are in fact vectors, in the sense that when multiplied by a scalar, they form another velocity for the same curve at the same point, and that the result of adding the velocities for two curves at some point $\mathcal{P}$ yields the velocity for some other curve at $\mathcal{P}$. The infinite set of all the (tangent) velocity vectors at $\mathcal{P}$ forms a vector space, $\mathcal{T}_\mathcal{P}$, the **tangent space** to the manifold M^n at $\mathcal{P} \in M^n$.

The n vectors $\boldsymbol{\partial}_\nu$ in eq. (1.5) form a basis for the tangent space, called a **coordinate basis**. To justify this statement, let $\boldsymbol{\partial}_{x^\mu}$ act on $f = x^\nu$, the coordinates on M; then $a^\mu \boldsymbol{\partial}_{x^\mu} x^\nu\big|_\mathcal{P} = a^\mu \delta_\mu{}^\nu = a^\nu$, where $a^\nu \in \mathbb{R}$ (see the paragraph after eq. (A.2) for the meaning of $\boldsymbol{\partial}_{x^\mu} x^\nu$). If $a^\mu \boldsymbol{\partial}_{x^\mu} x^\nu\big|_\mathcal{P} = 0$, then the coefficients a^μ all vanish, and the $\boldsymbol{\partial}_\nu$ are linearly independent. Since from eq. (1.5) they span the tangent space, they do form a basis.

We have found a home for *all* vectors at $\mathcal{P}$: they live in the tangent space of the manifold at $\mathcal{P}$. The components of $\boldsymbol{\partial}_\alpha$ at $\mathcal{P}$ are usually calculated by embedding M^n in $\mathbb{R}^N$ $(N > n)$, and writing the position[†] of $\mathcal{P}$ in $\mathbb{R}^N$ as N functions y^ν of the n coordinates on M. Then one simply computes $\partial_\alpha(y^1, \ldots, y^N)$ and evaluates the result at $\mathcal{P}$.

[†]Notice that I am not calling the position a vector, which it is not, because there is no notion of addition for the coordinates of points in a manifold.

Thus, in a coordinate basis, a vector $\mathbf{u}$ is written: $\mathbf{u} = u^\alpha \boldsymbol{\partial}_\alpha$. If its components u^α are differentiable functions of the coordinates, we say that $\mathbf{u}$ is a vector **field**.

Example 1.4. On S^2 (embedded in $\mathbb{R}^3$), a point $\mathcal{P}$ is mapped into the spherical coordinates (θ, ϕ), with $\theta \neq (0, \pi)$; $\mathcal{P}$ is described in $\mathbb{R}^3$ coordinates $(\sin\theta\,\cos\phi, \sin\theta\,\sin\phi, \cos\theta)$. Freezing say, θ generates a great circle on the sphere, of radius $\sin\theta$, at "colatitude" θ, and $\boldsymbol{\partial_\phi}$ has $\mathbb{R}^3$ 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 $\mathcal{P} \in S^2$ parametrised by (θ, ϕ), this is a vector tangent to the circle at colatitude θ.

The vector, $\boldsymbol{\partial}_\theta$, tangent to a meridian, has 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 $\mathbb{R}^2$-plane tangent to S^2 at $\mathcal{P}$.

Also, notice that $\boldsymbol{\partial}_\phi$ is not normalised to 1. In general, coordinate bases and cobases are not normalised. But $\boldsymbol{\partial}_{\hat\phi} \coloneqq \frac{1}{\sin\theta}\boldsymbol{\partial_\phi}$ is normalised; it is an element of the *non-coordinate* basis $\{\boldsymbol{\partial}_{\hat\theta},\, \boldsymbol{\partial}_{\hat\phi}\}$.

1.2.1 Directional derivatives as vectors

Notice that eq. (1.5) looks like the **directional derivative** of f in the direction of $\mathbf{v}$, which in basic calculus is written $\partial_{\mathbf{v}} f \coloneqq (\mathbf{v}\cdot\boldsymbol{\nabla})f$. Then the velocity vector has components $v^\nu = \mathrm{d}_\lambda x^\nu$. *This motivates us to identify any tangent vector* $\mathbf{t}$ *at that single point* $\mathcal{P}$ *with the directional derivative at* $\mathcal{P}$ *in the direction of* $\mathbf{t}$. Thus:

Definition 1.4. Given a differentiable function f on a manifold M^n parametrised in a local coordinate system by $(x^1, \ldots, x^n)$, then the action of a vector $\mathbf{t} \in \mathcal{T}_\mathcal{P}$ on f at a point $\mathcal{P}$ is defined as:

$$\mathbf{t}(f)\Big|_\mathcal{P} \;\coloneqq\; \partial_\mathbf{t} f \;=\; (t^\nu\,\boldsymbol{\partial}_\nu) f\,\Big|_\mathcal{P} \tag{1.6}$$

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

Definition 1.5. Let $x^\mu \in \mathbb{R}^n$ be the coordinate functions at $\mathcal{P} \in M^n$, and f a real-valued differentiable 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 from eq. (1.6):

$$[\mathbf{d}f](\mathbf{t}) \;\coloneqq\; \mathbf{t}(f) \;=\; \partial_\mathbf{t} f \tag{1.7}$$

To find the components of $\mathbf{d}f$ in the cobasis dual to the coordinate basis $\{\boldsymbol{\partial}_\mu\}$ at $\mathcal{P} \in M$, recall that the action of a 1-form on a coordinate-basis vector $\boldsymbol{\partial}_\nu$ outputs the corresponding component of the 1-form in that cobasis: $[\mathbf{d}f]_\nu = \mathbf{d}f(\boldsymbol{\partial}_\nu)$, which, from eq. (1.7), is the ordinary derivative of f in the direction of the basis vector $\boldsymbol{\partial}_\nu$, so $\partial_\nu f$. Now, taking $f = x^\mu$, the same argument immediately leads to: $[\mathbf{d}x^\mu](\boldsymbol{\partial}_\nu) = \partial_\nu x^\mu = \delta^\mu{}_\nu$, which is the defining equation (1.2) for a cobasis, with $\mathbf{e}_\mu = \boldsymbol{\partial}_\mu$ and $\boldsymbol{\omega}^\mu = \mathbf{d}x^\mu$. Then, choosing $\{\boldsymbol{\partial}_\mu\}$ as basis for $\mathcal{T}_\mathcal{P}$, we conclude that $\{\mathbf{d}x^\mu\}$ *is the basis, dual to* $\{\boldsymbol{\partial}_\mu\}$*, of the* **cotangent space**, $\mathcal{T}^*_\mathcal{P}$, *dual to* $\mathcal{T}_\mathcal{P}$. When written in a coordinate cobasis, $\boldsymbol{\sigma} = \sigma_\alpha \mathbf{d}x^\alpha$ is called a **differential form**.

If we think of f as a 0-form, the differential of f is the **gradient** 1-form $\mathbf{d}f$:

$$\mathbf{d}f \;=\; \partial_\mu f\,\mathbf{d}x^\mu \tag{1.8}$$

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 basis vector is inserted in its input slot, as in eq. (1.7), can it output a number, the change of f in the direction of the basis 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 is not valid on an arbitrary manifold, since $\mathbf{d}x^\mu$, like all 1-forms at a point, lives in the cotangent space, not the manifold.

1.2.3 Transformations on bases, cobases, and components

Let (U_1, x) and (U_2, y) be two overlapping charts (see definition A.1) on a manifold, with x and y their coordinates. At a point $\mathcal{P} \in U_1 \cap U_2$, the transformation on the components of a vector $\mathbf{v}$ is shown in Appendix B to be:

$$v^\nu_y \;=\; \partial_{x^\mu} y^\nu\big|_{x_\mathcal{P}}\, v^\mu_x \tag{1.9}$$

Remarkably, this transformation of components is *linear* and *homogeneous*, even though coordinate transformations beween (U_1, x) and (U_2, y) are often non-linear. Coordinates are not vector components! Thus, in coordinate bases, the coefficients in the transformation law are the entries, $\partial_{x^\mu} y^\nu$, of the **Jacobian matrix** evaluated at $\mathcal{P}$. With this one quickly shows (EXERCISE), using the chain rule, that $\mathbf{v}$ is unchanged by the transformation.

In general bases, transformations L on vector components must be assumed homogeneous and linear:

$$v^{\alpha'} \;=\; v^\mu\, L^{\alpha'}{}_\mu \;=\; L^{\alpha'}{}_\mu\, v^\mu \tag{1.10}$$

where the prime refers to the y coordinates in (1.9). This is the more traditional definition of vector components still in use in physics. For the vector itself not to change under a change of coordinates, components of a basis vector, $\mathbf{e}_\mu$, written as a row-matrix $\mathbf{e}_\mu$ in a "background" reference grid, must transform as:

$$\mathbf{e}_{\alpha'} \;=\; (L^{-1})^\mu{}_{\alpha'}\, \mathbf{e}_\mu \;=\; \mathbf{e}_\mu\, (L^{-1})^\mu{}_{\alpha'} \tag{1.11}$$

Note that the last expression is *not* matrix multiplication, because the subscript of a basis vector is a *label for the vector*, not for a component of this vector. It should be viewed as a linear combination of basis vectors.

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 $\mathbf{e}_\mu = \mathbf{e}_{\nu'} L^{\nu'}{}_\mu$, where the $L^{\nu'}{}_\mu$ are the coefficients of the matrix L representing the same transformation L as above. Let $\{\boldsymbol{\omega}^\alpha\}$ and $\{\boldsymbol{\omega}^{\alpha'}\}$ be their two respective cobases in $\mathcal{V}^*$. Then, writing $\boldsymbol{\omega}^\alpha = M^\alpha{}_{\beta'} \boldsymbol{\omega}^{\beta'}$ where the $M^\alpha{}_{\beta'}$ are the matrix coefficients of the corresponding transformation M between the cobases, it can be shown (EXERCISE) that M is the inverse of L, ie. $M^\alpha{}_{\nu'}\, L^{\nu'}{}_\beta = \delta^\alpha{}_\beta$ in index notation and $M = L^{-1}$ in matrix notation. This means that: $\boldsymbol{\omega}^{\alpha'} = L^{\alpha'}{}_\beta\, \boldsymbol{\omega}^\beta$ (again, this is not matrix multiplication). Cobases transform with the same matrix as vector components!

The transformation law of the components σ_α of a 1-form $\boldsymbol{\sigma}$ immediately follows (EXERCISE). Indeed, $\sigma_\alpha \boldsymbol{\omega}^\alpha = \sigma_{\beta'}\, \boldsymbol{\omega}^{\beta'}$ yields:

$$\sigma_{\alpha'} \;=\; \sigma_\mu\, (L^{-1})^\mu{}_{\alpha'} \tag{1.12}$$

The following table summarises all the possible transformations, both in general and in coordinate bases:

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

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 $(\mathsf{L}^{-1})^\alpha{}_{\beta'}$.

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

Also, in a coordinate basis, we cannot call $\partial_\mu f$ the components of the gradient *vector*, $\boldsymbol{\nabla} f$. They do not transform as vector components, as can be seen by calculating $\partial_{\mu'} f$ in terms of $\partial_\nu f$ using the chain rule (EXERCISE).

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 $\mathcal{V}$ and its dual $\mathcal{V}^*$, 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, $\{T : \mathcal{V}^* \times \ldots \times \mathcal{V}^* \times \mathcal{V} \times \ldots \times \mathcal{V} \longmapsto \mathbb{R}\}$, with $r + s = \dim \mathcal{V}$:

Definition 1.6. General tensors $\mathbf{T} \in \mathcal{T}^r_s$ of **type** (order, valence) (r, s) are multilinear mappings of r covectors and s vectors to a real number:

$$\begin{aligned}\mathbf{T}(\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_r, \mathbf{u}_1, \ldots, \mathbf{u}_s) &= \sigma_{\mu_1} \ldots \sigma_{\mu_r}\, u^{\nu_1} \ldots u^{\nu_s}\, \mathbf{T}(\mathbf{d}x^{\mu_1}, \ldots, \mathbf{d}x^{\mu_r}, \boldsymbol{\partial}_{\nu_1}, \ldots, \boldsymbol{\partial}_{\nu_s}) \\ &= T^{\mu_1 \ldots \mu_r}{}_{\nu_1 \ldots \nu_s}\, \sigma_{\mu_1} \ldots \sigma_{\mu_r}\, u^{\nu_1} \ldots u^{\nu_s} \end{aligned} \tag{1.13}$$

Tensors of type $(r, 0)$ are often said to be **contravariant**, and tensors of type $(0, s)$, **covariant**.

$T^{\mu_1 \ldots \mu_r}{}_{\nu_1 \ldots \nu_s}$ are the **mixed** components of $\mathbf{T}$ *in a basis and cobasis*, chosen here to be coordinate ones. When these components are are real-valued differentiable functions of coordinates on the manifold, we speak of $\mathbf{T}$ as a **tensor field**, eg., the coordinate vector field $\boldsymbol{\partial}_\mu$, the gravitational and electric fields at a point.

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 inputs, one for each slot.

1.3.1 The outer product of tensors

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

Definition 1.7. The **Kronecker (outer) 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 tensors $\boldsymbol{\sigma} \otimes \boldsymbol{\tau}$:

$$[\boldsymbol{\sigma} \otimes \boldsymbol{\tau}](\mathbf{u},\, \mathbf{v}) \;=\; \boldsymbol{\sigma}(\mathbf{u})\, \boldsymbol{\tau}(\mathbf{v}) \tag{1.14}$$

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 tensors $\mathbf{u} \otimes \mathbf{v}$ of rank 2:

$$[\mathbf{u} \otimes \mathbf{v}](\boldsymbol{\sigma},\, \boldsymbol{\tau}) \;=\; \mathbf{u}(\boldsymbol{\sigma})\, \mathbf{v}(\boldsymbol{\tau}) \tag{1.15}$$

There are outer-product spaces $\mathcal{V}_1 \otimes \mathcal{V}_2^*$ with 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 elements $\boldsymbol{\sigma} \otimes \mathbf{v}(\mathbf{u}, \boldsymbol{\tau}) = \boldsymbol{\sigma}(\mathbf{u})\mathbf{v}(\boldsymbol{\tau})$. An outer product of tensors is said to be **decomposable**.

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

Example 1.5. Let $P_{\leq n}$ be the vector space whose elements are polynomials of degree $\leq n$ over $[-1, 1]$ (see example 1.3). Then we can construct a map $\mathbf{F} : P_{\leq n} \times P_{\leq n} \longmapsto \mathbb{R}$ defined by $\int_{-1}^{1} p(x)\, q(x)\, \mathrm{d}x$, where $(p, q) \in P_{\leq n}$. This *symmetric, bilinear* map is a $(0, 2)$ tensor with two vectors as inputs and a number as output.

Now take $\mathcal{V}_1 = \mathcal{V}_2 = \mathcal{V}$. If $\{\boldsymbol{\partial}_\mu\}$ is a coordinate basis for $\mathcal{V}$, then $\{\boldsymbol{\partial}_\mu \otimes \boldsymbol{\partial}_\nu\}$ is a coordinate basis for $\mathcal{V} \otimes \mathcal{V}$. Similarly, if $\{\mathbf{d}x^\alpha\}$ is a coordinate basis 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 $(2, 0)$ tensor $\mathbf{T}$ lives in $\mathcal{V} \times \mathcal{V}$, and *any* covariant $(0, 2)$ tensor $\mathbf{F}$ lives in $\mathcal{V}^* \times \mathcal{V}^*$:

$$\mathbf{T} \;=\; T^{\mu\nu} \boldsymbol{\partial}_\mu \otimes \boldsymbol{\partial}_\nu, \qquad\qquad \mathbf{F} \;=\; F_{\alpha\beta}\, \mathbf{d}x^\alpha \otimes \mathbf{d}x^\beta \tag{1.16}$$

Therefore, the action of $\mathbf{T}$ on pairs of one-forms, and of $\mathbf{F}$ on pairs of vectors, outputs numbers:

$$\begin{aligned}\mathbf{T}(\boldsymbol{\sigma},\, \boldsymbol{\tau}) &= T^{\mu\nu} \boldsymbol{\partial}_\mu \otimes \boldsymbol{\partial}_\nu\, (\boldsymbol{\sigma},\, \boldsymbol{\tau}) \;=\; T^{\mu\nu} \boldsymbol{\partial}_\mu(\boldsymbol{\sigma})\, \boldsymbol{\partial}_\nu(\boldsymbol{\tau}) \;=\; T^{\mu\nu} \sigma_\mu \tau_\nu \\ \mathbf{F}(\mathbf{u},\, \mathbf{v}) &= F_{\alpha\beta}\, \mathbf{d}x^\alpha \otimes \mathbf{d}x^\beta(\mathbf{u},\, \mathbf{v}) \;=\; F_{\alpha\beta}\, \mathbf{d}x^\alpha(\mathbf{u})\, \mathbf{d}x^\beta(\mathbf{v}) \;=\; F_{\alpha\beta}\, u^\alpha v^\beta \end{aligned} \tag{1.17}$$

We can also input a single vector (1-form), so long as we specify into which of the two input slots it should be inserted. For instance, we could write $\mathbf{F}(\mathbf{u},\ )$, or $\mathbf{F}(\ ,\mathbf{u})$:

$$\mathbf{F}(\mathbf{u},\ ) \ = \ \big(F_{\alpha\beta}\,\mathbf{d}x^\alpha(\mathbf{u})\big)\,\mathbf{d}x^\beta \ = \ \big(F_{\alpha\beta}\,u^\alpha\big)\,\mathbf{d}x^\beta \ = \ \sigma_\beta\,\mathbf{d}x^\beta \ = \ \boldsymbol{\sigma}$$
$$\mathbf{F}(\ ,\mathbf{u}) \ = \ F_{\alpha\beta}\,\mathbf{d}x^\alpha\mathbf{d}x^\beta(\mathbf{u}) \ = \ \big(F_{\alpha\beta}\,u^\beta\big)\,\mathbf{d}x^\alpha \ = \ \tau_\alpha\,\mathbf{d}x^\alpha \ = \ \boldsymbol{\tau}$$

Unless the $F_{\alpha\beta}$ happen to be symmetric in their indices, the two resulting 1-forms $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ are not the same! EXERCISE: what does one get when one lets a $(1,1)$ tensor $\mathbf{T}$ act on a single vector, $\mathbf{u}$, or a single covector, $\boldsymbol{\sigma}$? Can this outcome happen with a tensor of any other type?

Generalising, $\boldsymbol{\partial}_{\mu_1}\otimes\cdots\otimes\boldsymbol{\partial}_{\mu_r}\otimes\mathbf{d}x^{\nu_1}\otimes\cdots\otimes\mathbf{d}x^{\nu_s}$ forms a basis for the tangent space of (r,s) tensors:

$$\mathbf{T} \ = \ T^{\mu_1\ldots\mu_r}{}_{\nu_1\ldots\nu_s}\,\boldsymbol{\partial}_{\mu_1}\otimes\cdots\otimes\boldsymbol{\partial}_{\mu_r}\otimes\mathbf{d}x^{\nu_1}\otimes\cdots\otimes\mathbf{d}x^{\nu_s} \tag{1.18}$$

To obtain a number, all input slots must be filled. Inserting m input vectors in a (r,s) tensor outputs a $(r,s{-}m)$ tensor; inserting q input 1-forms outputs a $(r{-}q,s)$ tensor.

It is important to remember that interchanging vectors or 1-forms in input slots may result in different output.

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.

> **Definition 1.8.** A tensor that remains unchanged under transposition of two of its input slots *of the same type* is said to be **symmetric** in these slots. Its components are unchanged under permutation of indices corresponding to those slots. If it switches sign under transposition, we say that it is **antisymmetric** in these slots, and the components corresponding to these slots also switch sign. Inserting the *same* 1-form (in any two contravariant slots) or vector (in any two covariant slots) outputs zero.
>
> Symmetry and antisymmetry are basis-independent properties.
>
> **Example 1.6.** Take an antisymmetric $(0,2)$ tensor: $\mathbf{F} = F_{[\mu\nu]}\,\mathbf{d}x^\mu\otimes\mathbf{d}x^\nu$, where square brackets mean that indices are antisymmetric. Then $\mathbf{F}(\mathbf{u},\,\mathbf{u}) = F_{\mu\nu}u^\mu\,u^\nu = F_{\nu\mu}u^\nu\,u^\mu = -F_{\mu\nu}u^\mu\,u^\nu \ = \ 0$.

Among important tensors are those whose components are **completely symmetric** in all their covariant (or contravariant) indices, and those which are completely antisymmetric (**skew-symmetric**, **alternating**) in all their covariant (or 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 three dimensions, many physically relevant tensors are symmetric (moment of inertia, electrical polarisation, multipole moment, Maxwell stress tensor). Antisymmetric 3-d rank-2 tensors are not usual, although later I will argue that the 3-dim magnetic field is more naturally described by an antisymmetric $(0,2)$ tensor than by a vector.

In four dimensions, we also have symmetric tensors, such as the important energy-momentum tensor, and the famous antisymmetric $(0,2)$ Faraday electromagnetic field tensor $\mathbf{F}$.

It can be useful to symmetrise or skew-symmetrise a general $(r,0)$ or $(0,s)$ tensor. To symmetrise the components of a $(0,s)$ tensor $\mathbf{T}$, construct:

$$T_{(\mu_1\ldots\mu_s)} \ = \ \frac{1}{s!}\sum_\pi T_{\pi(\mu_1\ldots\mu_s)} \tag{1.19}$$

with round brackets around symmetric indices, and where the sum runs over all permutations π of $\mu_1\ldots\mu_s$. Contravariant components are symmetrised in the same way.

To antisymmetrise the components of a $(0,s)$ or $(r,0)$ tensor $\mathbf{T}$, construct, e.g.:

$$T_{[\mu_1\ldots\mu_s]} \ = \ \frac{1}{s!}\sum_\pi \mathrm{sgn}(\pi)\,T_{\pi(\mu_1\ldots\mu_s)} \ = \ \frac{1}{s!}\,\delta^{\nu_1\ldots\nu_s}_{\mu_1\ldots\mu_s}\,T_{\nu_1\ldots\nu_s} \tag{1.20}$$

with square brackets around antisymmetric indices, and the **general permutation symbol**, $\delta^{j_1\dots j_s}_{i_1\dots i_s}$, defined as:

$$\delta^{j_1\dots j_s}_{i_1\dots i_s} := \begin{cases} +1 & j_1\dots j_s \text{ an even permutation of } i_1\ \dots\ i_s \\ -1 & j_1\dots j_s \text{ an odd permutation of } i_1\ \dots\ i_s \\ 0 & j_1\dots j_s \text{ not a permutation of } i_1\ \dots\ i_s \\ 0 & j_k\ =\ j_l \text{ or } i_k\ =\ i_l \text{ for some } k,\ l \end{cases} \tag{1.21}$$

where even/odd means an even/odd number of transpositions (switches) of two indices. The permutation symbol is seen to be antisymmetric in its upper and lower indices.

$s!$ 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 $T_{\mu_1\dots\mu_s}$ should already be symmetric or skew-symmetric. A simple example is that of a $(2,0)$ tensor:

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

A $(0,2)$ tensor can also be similarly decomposed.

EXERCISE: Symmetrise and antisymmetrise $\mathbf{F}(\boldsymbol{\sigma},\boldsymbol{\tau},\boldsymbol{\theta})$. If $\mathbf{F}=F^{\mu\nu\lambda}\mathbf{e}_\mu\otimes\mathbf{e}_\nu\otimes\mathbf{e}_\lambda$, write $\mathbf{F}_{\rm s}$ and $\mathbf{F}_{\rm a}$. How many components do $\mathbf{F}_{\rm s}$ and $\mathbf{F}_{\rm a}$ have when $\mathbf{F}$ is defined over a 3-dim space? a 4-dim space? Is it possible to reconstruct $F^{\mu\nu\lambda}$ from $F^{(\mu\nu\lambda)}$ and $F^{[\mu\nu\lambda]}$?

1.3.3 Transformations on tensors

Using the transformations in the table of section 1.2.3, 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:

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

We obtain:

$$T^{\alpha'_1\dots\alpha'_r}{}_{\beta'_1\dots\beta'_s}\ =\ T^{\mu_1\dots\mu_r}{}_{\nu_1\dots\nu_s}\,L^{\alpha'_1}{}_{\mu_1}\ \dots\ L^{\alpha'_r}{}_{\mu_r}\,(L^{-1})^{\nu_1}{}_{\beta'_1}\ \dots\ (L^{-1})^{\nu_s}{}_{\beta'_s} \tag{1.22}$$

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

$$S^{\alpha'\beta'}\ =\ S^{\mu\nu}\,L^{\alpha'}{}_{\mu}\,L^{\beta'}{}_{\nu}\ =\ L^{\alpha'}{}_{\mu}\,S^{\mu\nu}\,(L^{\rm T})_{\nu}{}^{\beta'}\qquad\Longleftrightarrow\qquad\mathsf{S}'\ =\ \mathsf{L}\,\mathsf{S}\,\mathsf{L}^{\rm T}$$

where the equation on the right is in matrix form. and $\mathsf{L}^{\rm T}$ is the transpose of L. EXERCISE: Find the matrix form for the transformation of a $(1,1)$ tensor. Tensors of rank 2 $\big((2,\,0),\,(0,\,2),\,(1,\,1)\big)$ on n-dim spaces $\mathcal{V}$ and $\mathcal{V}^*$can be represented by $n\times n$ matrices. EXERCISE: Does the determinant of the matrix representation of any of the three abovementioned rank-2 tensors transform as a scalar?

An immediate consequence of eq. (1.22) is that a tensor that is zero in a basis will remain zero in any other basis. Thus, *any equation made of tensors (or components) that is valid in one basis must hold in any other basis.*

In the older view of tensors defined by transformations, an object may have tensor character 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.

The transformation rules can always be used to establish whether an object is a tensor. For instance, on a space of dimension n, the Kronecker delta, with components $\delta^\nu{}_\mu$, is represented by the $n\times n$ identity matrix. It is a mixed rank-2 tensor. Indeed, from the transformation law, eq. (1.22):

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

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!

Fortunately, often we can avoid using the transformation law (1.22) if we build tensors from other objects known to be tensors. The following section presents some important examples.

1.4 Two More Ways to Construct Tensors

1.4.1 Contracted tensors

Definition 1.9. The **contraction** of a mixed-type tensor is a linear map $\mathcal{T}^r_s \to \mathcal{T}^{r-1}_{s-1}$, $(r \geq 1,\, s \geq 1)$. More precisely, going back to eq. (1.13), insert a basis and its cobasis into *only two* input slots:

$$\mathbf{T}(\ldots, \mathbf{d}x^\gamma, \ldots, \boldsymbol{\partial}_\gamma, \ldots) \;=\; T^{\cdots\,\gamma\,\cdots}{}_{\cdots\,\gamma\,\cdots}\, \ldots\otimes\boldsymbol{\partial}_{\mu_{i-1}}\otimes\boldsymbol{\partial}_{\mu_{i+1}}\otimes\ldots\ldots\otimes\mathbf{d}x^{\nu_{j-1}}\otimes\mathbf{d}x^{\nu_{j+1}}\otimes\ldots \quad (1.23)$$

In terms of components, one just makes a contravariant index μ the same as a covariant index, ν, by multiplying the component by $\delta^\nu{}_\mu$, thus forcing a summation over these indices.

For instance, the contraction of $\mathbf{T} = T^\alpha{}_\beta \boldsymbol{\partial}_\alpha \otimes \mathbf{d}x^\beta$ is a scalar (an invariant!), 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, there has been no unique link between tensors of type $(r, 0)$, $(0, r)$, or $(r-q, q)$. To set up such a link, new structure is needed, in the form of an inner product:

Definition 1.10. The **inner product** of two vectors, $\mathbf{u}$ and $\mathbf{v}$, in a general basis, $\{\mathbf{e}_\mu\}$, is defined as:

$$<\mathbf{u},\, \mathbf{v}> \;=\; u^\mu\, v^\nu \;<\mathbf{e}_\mu,\, \mathbf{e}_\nu> \;:=\; u^\mu\, v^\nu\, \mathbf{e}_\mu(\mathbf{e}_\nu) \quad (1.24)$$

where $<\mathbf{e}_\mu,\, \mathbf{e}_\nu> = <\mathbf{e}_\nu,\, \mathbf{e}_\mu>$ and , as before, the action of $\mathbf{e}_\mu$ on $\mathbf{e}_\nu$ is just the matrix product of the row representation of $\mathbf{e}_\mu$ and the column representation of $\mathbf{e}_\nu$. Thus, the inner product provides a symmetric bilinear map, $\mathfrak{g} \,:\, \mathcal{V} \times \mathcal{V} \longmapsto \mathbb{R}$, with $\mathfrak{g}(\mathbf{e}_\mu,\, \mathbf{e}_\nu) = g_{\mu\nu}$ the components of the symmetric $(0,\, 2)$ (covariant) tensor $\mathfrak{g}$. In a coordinate basis, $\mathfrak{g} = g_{\mu\nu}\mathbf{d}x^\mu \otimes \mathbf{d}x^\nu$.

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

The bilinear map on the space $P_{\leq n}$ of polynomials defined in example 1.5 is one possible inner product. But the basis $\{1, x, \ldots, x^n\}$ is not orthogonal with respect to it. Do you know a basis that is?

We can insert a single vector in, say, the second slot of $\mathfrak{g}$, with as result a covector:

$$\mathfrak{g}(\quad,\mathbf{u}) \;=\; g_{\mu\nu}\, \mathbf{d}x^\mu\, \mathbf{d}x^\nu(\mathbf{u}) \;=\; (g_{\mu\nu}u^\nu)\, \mathbf{d}x^\mu \;\in\; \mathcal{V}^*$$

As a map from $\mathcal{V}$ to its dual space, $\mathfrak{g}$, because it is symmettric, establishes a *unique* correspondence between $\mathbf{u}$ and a 1-form with components $g_{\mu\nu}u^\nu$!

We think of these components as the *covariant* components, u_ν, of $\mathbf{u}$. Now, $\mathfrak{g}$ must be invertible (ie. $\det \mathbf{g} \neq 0$), so that $\mathfrak{g}^{-1}$ takes a 1-form to a vector, 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:

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

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 other words, $\mathfrak{g}$ may be used to convert any contravariant index of a given tensor into a covariant one ("**lowering the index**"), while $\mathfrak{g}^{-1}$ may be used to convert any covariant index of a given tensor into a contravariant one ("**raising the index**"). Thus, we say that the inner product sets up an isomorphism between a vector space and its dual. Because of this connection, we also have: $\boldsymbol{\partial}_\mu = g_{\mu\nu}\mathbf{d}x^\nu$. A given tensor can have all-contravariant, all-covariant, or mixed components! In particular:

$$g^\mu{}_\nu \;=\; \mathfrak{g}(\mathbf{d}x^\mu,\, \boldsymbol{\partial}_\nu) \;\equiv <\mathbf{d}x^\mu,\, \boldsymbol{\partial}_\nu> \;=\; \mathbf{d}x^\mu(\boldsymbol{\partial}_\nu) \;=\; \delta^\mu{}_\nu \quad (1.25)$$

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 $g^{\mu\lambda}g_{\lambda\nu} = g^\mu{}_\nu = \delta^\mu{}_\nu$, which justifies our earlier assertion that $g^{\mu\nu}$ are the components of $\mathfrak{g}^{-1}$. On a n-dim space, $g^\mu{}_\mu = \delta^\mu{}_\mu = n$.

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 $\mathbf{I}$.

A final word of caution: we always wrote our matrices as $L^\mu{}_\nu$, with the left 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 a basis in $\mathcal{V}^n$ to its *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* a tensor component, but $g_{\mu\nu}$ is.

1.4.3 The metric

The inner product plays another extremely important rôle: it allows us to define distances on the manifold $\mathcal{M}^n$:

Definition 1.11. As a **metric tensor** (metric for short), $\mathfrak{g}$ tells us how to calculate lengths in a vector space tangent to a point on a manifold, as well as distances on the manifold itself. The name is often extended (abusively) to its components $g_{\mu\nu}$. Thus,

$$\Delta s^2 \;=\; g_{\mu\nu}\,\Delta x^\mu\,\Delta x^\nu \tag{1.26}$$

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, one identifies the bilinear form $\mathbf{d}\boldsymbol{s}^\mathbf{2}$ with $\mathfrak{g} = g_{\mu\nu}\,\mathbf{d}x^\mu \otimes \mathbf{d}x^\nu$ which then represents the interval Δs^2 for a $\Delta\mathbf{x}$ to be specified: $\Delta s^2 = <\Delta\mathbf{x}, \Delta\mathbf{x}>$, identical to the standard eq. (1.26).

Example 1.7. Consider the positions $\mathbf{x}_1$ and $\mathbf{x}_2$ in $\mathbb{R}^3$ in Cartesian coordinates:

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

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 1.8. In $\mathbb{R}^4$, let $\mathbf{x}_i$ $(i = 1,\ 2)$ be two positions with (ct_i, x_i, y_i, z_i) as contravariant and $(-ct_i, x_i, y_i, z_i)$ as covariant components. Then take the *indefinite* $\boldsymbol{\eta} \equiv \mathfrak{g}$ with matrix representation: $\eta_{\mu\nu} = \mathrm{diag}\,(-1,\ 1,\ 1,\ 1)$. Then:

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

is the spacetime distance between two events in Special Relativity, with c the speed of light. Since $\boldsymbol{\eta}$ is indefinite, vectors such that $\mathfrak{g}(\mathbf{x}, \mathbf{x}) = 0$ exist. And, just as $\partial_\nu x^\mu = \delta^\mu{}_\nu$, we must write: $\partial_\nu x_\mu = \eta_{\mu\nu}$.

On general manifolds, the distance between two points $\lambda = a$ and $\lambda = b$ on a curve parametrised by λ is given by $\int_a^b \sqrt{\mathfrak{g}(\mathbf{v},\mathbf{v})}\mathrm{d}\lambda = \int_a^b \sqrt{g_{\mu\nu}\mathrm{d}_\lambda x^\mu \mathrm{d}_\lambda x^\nu}\mathrm{d}\lambda$, where $\mathbf{v}$ is the velocity vector and x^μ are the coordinates describing the curve on the manifold. The metric, or line element, is said to define the geometry of a manifold. Two manifolds of the same dimension can have different geometries, eg. $\mathbb{R}^4$ with a positive-definite ($\Delta s^2 > 0$) metric is not the metric of 4-dim "flat" spacetime of special Relativity.

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 covariant and contravariant *components* can be different, even in the same coordinates.

Definition 1.12. 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 $\{\mathbf{e}_{\hat{\mu}}\}$, extending the usual definition of orthonormality which admits only $+1$ and 0, and useful in the case of indefinite metrics.

Let n_+ (n_-) denote the number of diagonal elements $\mathfrak{g}(\mathbf{e}_{\hat{\mu}}, \mathbf{e}_{\hat{\mu}})$ 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 so is the signature.

The sign of the signature of an indefinite metric is arbitrary and is set by convention, which can be a source of confusion. Example 1.8 sets $s=+2$, a nice choice when spatial indices are often raised/lowered. In more general spacetimes, $s=-2$ is often used (but not always… see Misner, Thorne and Wheeler's *Gravitation*). Thus, beware!

Definition 1.13. 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**, or **Lorentzian** (aka **Minkowski** when $n=4$). Example 1.8 has a Minkowski metric in four-dimensional space.

Thanks to the metric, we recover the *vector* gradient of a function defined in calculus. You may have noticed that throughout our discussion of manifolds and tangent spaces, no mention was made of an inner product, because none was needed—until now. A metric $\mathfrak{g}$ pairs the 1-form $\mathbf{d}f$ with a vector, $\boldsymbol{\nabla} f$; indeed, from eq. (1.24):

$$\mathfrak{g}(\boldsymbol{\nabla} f,\, \mathbf{v}) \;=\; (g_{\mu\nu}\partial^\mu f)\, v^\nu \;=\; (\partial_\nu f)\, v^\nu \;=\; [\mathbf{d}f](\mathbf{v}) \tag{1.27}$$

where $\mathbf{v}$ is an arbitrary vector, and the components of $\boldsymbol{\nabla} f$ in a coordinate basis are given by: $\partial^\mu = g^{\mu\nu}\,\partial_\nu f$. Only in a Euclidean metric with a standard basis are the components of $\boldsymbol{\nabla} f$ the same as those of $\mathbf{d}f$.

Example 1.9. In Minkowski spacetime with coordinates (ct, x^1, x^2, x^3) and metric $\eta_{\mu\nu}=\mathrm{diag}(-1,1,1,1)$:

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

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

$$g' = \left|\frac{\partial x}{\partial x'}\right|^2 g \tag{1.28}$$

where $|\partial x/\partial x'|$ is the **Jacobian** of the transformation matrix *from* x to x' coordinates. Then g is not invariant!

Definition 1.14. 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.

This might seem no more than an exotic property until we consider the n-dim volume element in an integral. As we know from calculus, $\mathrm{d}^n x' = |\partial x'/\partial x|\mathrm{d}^n x$ (note the position of the prime!), so is not invariant. Then $\int f(\mathbf{x})\mathrm{d}^n x$ is not invariant and seems to have kept a memory of the integration variables. Instead, transform $\sqrt{|g|}\,\mathrm{d}^n x$:

$$\sqrt{|g'|}\,\mathrm{d}^n x' \;=\; \left|\frac{\partial x}{\partial x'}\right| \sqrt{|g|}\, \left|\frac{\partial x'}{\partial x}\right| \mathrm{d}^n x \;=\; \sqrt{|g|}\,\mathrm{d}^n x$$

which is seen to be a scalar! Integrals written as $\int \sqrt{|g|} f(\mathbf{x})\mathrm{d}^n x$ *are* invariant. 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, section 1.5.2 will introduce a deeper definition of the volume element.

1.5 Exterior Algebra

1.5.1 The exterior product

Definition 1.15. The **exterior** product of two 1-forms is the antisymmetrised outer product 2-form:

$$\boldsymbol{\sigma}\wedge\boldsymbol{\tau} \;:=\; \boldsymbol{\sigma}\otimes\boldsymbol{\tau} \,-\, \boldsymbol{\tau}\otimes\boldsymbol{\sigma}$$

In general, the exterior (or wedge) product can be used to construct a **simple** (or decomposable) skew-symmetric covariant $(0,p)$ tensor out of p 1-forms:

$$\boldsymbol{\sigma}^1\wedge\ldots\wedge\boldsymbol{\sigma}^{\mathrm{p}} \;=\; \delta^{1\,\ldots\,p}_{\mu_1\,\ldots\,\mu_p}\,\boldsymbol{\sigma}^{\mu_1}\otimes\cdots\otimes\boldsymbol{\sigma}^{\mu_p} \;=\; \epsilon_{\mu_1\,\ldots\,\mu_p}\,\boldsymbol{\sigma}^{\mu_1}\otimes\cdots\otimes\boldsymbol{\sigma}^{\mu_p} \tag{1.29}$$

where $\epsilon_{\mu_1\,\ldots\,\mu_p}$ is the Levi-Civita symbol introduced in Appendix C.

Applied to a cobasis element $\mathbf{d}x^{\rho_1} \otimes \cdots \otimes \mathbf{d}x^{\rho_p}$, where $1 \le \rho_1 < \ldots < \rho_p \le n$, this becomes: $\mathbf{d}x^{\rho_1} \wedge \cdots \wedge \mathbf{d}x^{\rho_p} = \delta^{\rho_1 \ldots \rho_p}_{\mu_1 \ldots \mu_p}\, \mathbf{d}x^{\mu_1} \otimes \cdots \otimes \mathbf{d}x^{\mu_p}$, so that:

$$[\mathbf{d}x^{\rho_1} \wedge \cdots \wedge \mathbf{d}x^{\rho_p}](\boldsymbol{\partial}_{\nu_1}, \ldots, \boldsymbol{\partial}_{\nu_p}) \;=\; \delta^{\rho_1 \ldots \rho_p}_{\mu_1 \ldots \mu_p}\, \delta^{\mu_1}{}_{\nu_1} \cdots \delta^{\mu_p}{}_{\nu_p} \;=\; \delta^{\rho_1 \ldots \rho_p}_{\nu_1 \ldots \nu_p} \qquad (1 \le \rho_1 < \ldots < \rho_p \le n) \tag{1.30}$$

$(0,p)$ skew-symmetric tensors (usuallly called **p-forms**) live in the cotangent space, denoted by $\boldsymbol{\Omega}^p(M^n)$, at a point in the manifold M^n; they act on vectors in $\mathbb{R}^n$.

Thus, from $\{\mathbf{d}x^\rho\}$ $(1 \le \rho \le n)$ a basis of $\boldsymbol{\Omega}^p(\mathbb{R}^n)$ can be constructed which contains $n!/(p!(n-p)!)$ independent, non-zero elements. In particular, a n-form on a n-dimensional space (aka **top form**) is a one-component object, 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. Skew-symmetry forces the maximum rank of a non-trivial p-form in n dimensions to be n (why?).

Recall from the table in section 1.2.3 that cobasis 1-forms transform from x coordinates to u coordinates as: $\mathbf{d}x^\alpha = (\partial x^\alpha/\partial u^\beta)\mathbf{d}u^\beta$. Then the cobasis top form transforms as: $\mathbf{d}x^1 \wedge \ldots \wedge \mathbf{d}x^n = |\partial x/\partial u|\, \mathbf{d}u^1 \wedge \ldots \wedge \mathbf{d}u^n$, where $|\partial x/\partial u|$ is the Jacobian of the transformation. EXERCISE: With $x = r\cos\theta$ and $y = r\sin\theta$ in $\mathbb{R}^2$ with the origin removed (why?), find the cobasis $\mathbf{d}x \wedge \mathbf{d}y$ in (r,θ) coordinates.

The exterior product of a basis of $\boldsymbol{\Omega}^p$ and a basis of $\boldsymbol{\Omega}^q$ is a basis, $\mathbf{d}x^{\rho_1} \wedge \ldots \wedge \mathbf{d}x^{\rho_p} \wedge \mathbf{d}x^{\rho_{p+1}} \wedge \ldots \wedge \mathbf{d}x^{\rho_{p+q}}$, of $\boldsymbol{\Omega}^{p+q}$, again with indices in increasing order, and $p+q \le n$.

Then we construct a $(p+q)$-form out of the antisymmetrised outer product of $\boldsymbol{\sigma} \in \boldsymbol{\Omega}^p$ and $\boldsymbol{\tau} \in \boldsymbol{\Omega}^q$:

$$[\boldsymbol{\sigma} \wedge \boldsymbol{\tau}](\mathbf{u}_{\rho_1}, \ldots, \mathbf{u}_{\rho_{p+q}}) = \delta^{\mu_1 \ldots \mu_p \nu_1 \ldots \nu_q}_{\rho_1 \ldots \rho_{p+q}}\, \boldsymbol{\sigma}(\mathbf{u}_{\mu_1} \ldots \mathbf{u}_{\mu_p})\, \boldsymbol{\tau}(\mathbf{u}_{\nu_1} \ldots \mathbf{u}_{\nu_q}) \qquad \mu_1 < \mu_2 \ldots < \mu_p, \quad \nu_1 < \ldots < \nu_q \tag{1.31}$$

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

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

An important property of the exterior product is its **graded** (dependent on the degree of forms) commutativity:

$$\boldsymbol{\sigma} \wedge \boldsymbol{\tau} \;=\; (-1)^{pq}\, \boldsymbol{\tau} \wedge \boldsymbol{\sigma} \tag{1.32}$$

where $\boldsymbol{\sigma}$ is a p form and $\boldsymbol{\tau}$ is a q form This follows directly from eq. (1.31) 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 $\delta^{\mu_1 \ldots \mu_p \nu_1 \ldots \nu_q}_{\rho_1 \ldots \rho_{p+q}}$. Thus, the exterior product commutes except when both forms have odd rank.

Eq. (1.31) is often easier to use than it might appear. Here are three examples:

Example 1.10. 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, $\boldsymbol{\tau}$:

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

Example 1.11. 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 1.12. 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, $\boldsymbol{\Omega}^2(\mathcal{V})$. Therefore, the *most general* (not necessarily simple!) 2-form can be written as:

$$\boldsymbol{\tau} \;=\; \tau_{12}\, \mathbf{d}x^1 \wedge \mathbf{d}x^2 \;+\; \tau_{23}\, \mathbf{d}x^2 \wedge \mathbf{d}x^3 \;+\; \tau_{31}\, \mathbf{d}x^3 \wedge \mathbf{d}x^1 \;=\; \frac{1}{2}\, \tau_{\mu\nu}\, \mathbf{d}x^\mu \wedge \mathbf{d}x^\nu \tag{1.33}$$

The summation on the right of the second equality is now unrestricted.

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

$$(\sigma^1{}_1\sigma^2{}_2 - \sigma^1{}_2\sigma^2{}_1)\,\mathbf{d}x^1 \wedge \mathbf{d}x^2 + (\sigma^1{}_3\sigma^2{}_1 - \sigma^1{}_1\sigma^2{}_3)\,\mathbf{d}x^3 \wedge \mathbf{d}x^1 + (\sigma^1{}_2\sigma^2{}_3 - \sigma^1{}_3\sigma^2{}_2)\,\mathbf{d}x^2 \wedge \mathbf{d}x^3 \tag{1.34}$$

Notice that, 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 $\boldsymbol{\sigma}^2$.

In four dimensions, a basis for $\boldsymbol{\Omega}^2$ contains 6 elements. EXERCISE: What are the components of the exterior product of two 1-forms in three and four dimensions? (Hint: the components must look like: $\sigma^1_\mu\sigma^2_\nu - \sigma^2_\mu\sigma^1_\nu$.)

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, with eq. (1.29):

$$\begin{aligned}\boldsymbol{\sigma}^1 \wedge \ldots \wedge \boldsymbol{\sigma}^{\mathrm{p}} &= \delta^{1\,\ldots\,p}_{\mu_1\,\ldots\,\mu_p}\,\boldsymbol{\sigma}^{\mu_1} \otimes \cdots \otimes \boldsymbol{\sigma}^{\mu_p} \qquad \text{(unrestricted sum over } \mu_i\text{)} \\ &= \left[\epsilon_{\mu_1\,\ldots\,\mu_p}\,\sigma^{\mu_1}{}_{\nu_1} \cdots \sigma^{\mu_p}{}_{\nu_p}\right] \mathbf{d}x^{\nu_1} \otimes \ldots \otimes \mathbf{d}x^{\nu_p} \qquad \text{(unrestricted sums over } \mu_i \text{ and } \nu_i\text{)}\end{aligned} \tag{1.35}$$

where definition C.1 has been used, and 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 $\boldsymbol{\sigma}^{\mathrm{i}}$, we may notice, referring back to section C.0.1, that the expression inside the square brackets in the second line is 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. (1.36), each term in the sum over the ν_i 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 the indices $\nu_1 < \ldots < \nu_p$ on these components must be the same as the ones on $\mathbf{d}x^{\nu_1} \wedge \ldots \wedge \mathbf{d}x^{\nu_p}$ in that term. Also, the ν_i indices in the square bracket have been antisymmetrised at the same time, automatically antisymmetrising the tensor-product basis elements $\mathbf{d}x^{\nu_1} \otimes \ldots \otimes \mathbf{d}x^{\nu_p}$. Therefore:

$$\boldsymbol{\sigma}^1 \wedge \ldots \wedge \boldsymbol{\sigma}^{\mathrm{p}} = \left[\epsilon_{\mu_1\,\ldots\,\mu_p}\,\sigma^{\mu_1}{}_{\nu_1} \cdots \sigma^{\mu_p}{}_{\nu_p}\right] \mathbf{d}x^{\nu_1} \wedge \ldots \wedge \mathbf{d}x^{\nu_p} \qquad (\nu_1 < \cdots < \nu_p) \tag{1.36}$$

With eq. (1.29), the output (a number!) resulting from inputting $\mathbf{u}_1, \ldots, \mathbf{u}_p$ into $\boldsymbol{\sigma}^1 \wedge \ldots \wedge \boldsymbol{\sigma}^p$ is:

$$\boldsymbol{\sigma}^1 \wedge \cdots \wedge \boldsymbol{\sigma}^{\mathrm{p}}(\mathbf{u}_1, \ldots, \mathbf{u}_{\mathrm{p}}) = \delta^{1\,\ldots\,p}_{\mu_1 \ldots \mu_p}\,\boldsymbol{\sigma}^{\mu_1} \otimes \cdots \otimes \boldsymbol{\sigma}^{\mu_p}(\mathbf{u}_1, \ldots, \mathbf{u}_{\mathrm{p}}) = \epsilon_{\mu_1\,\ldots\,\mu_p}\,\boldsymbol{\sigma}^{\mu_1}(\mathbf{u}_1) \cdots \boldsymbol{\sigma}^{\mu_p}(\mathbf{u}_p) = \det\left[\boldsymbol{\sigma}^i(\mathbf{u}_j)\right] \tag{1.37}$$

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

Example 1.13. 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:

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

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.

Example 1.14. 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. (1.21), follows from eq. (1.30):

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

Then eq. (1.37) becomes:

$$\delta^{\nu_1\ldots\nu_n}_{\mu_1\ldots\mu_n} = \begin{vmatrix} \delta^{\nu_1}{}_{\mu_1} & \cdots & \delta^{\nu_1}{}_{\mu_n} \\ \vdots & & \vdots \\ \delta^{\nu_n}{}_{\mu_1} & \cdots & \delta^{\nu_n}{}_{\mu_n} \end{vmatrix} \tag{1.38}$$

There is also an easy test for the linear independence of p 1-forms: if $\boldsymbol{\sigma}^1 \wedge \cdots \wedge \boldsymbol{\sigma}^p \neq 0$, they are linearly independent. If they were not, one of them at least could be written as a linear combination of the others and the wedge product would vanish. Conversely, if the p 1-forms are linearly independent,

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

Definition 1.16. Two bases are said to have the same (opposite) **orientation** if the determinant of the matrix of the transformation between them is positive (negative). Therefore, bases fall into two classes, or orientations. Orienting a manifold then means *arbitrarily* specifying one orientation to be positive (**right-handed**), and the other negative (**left-handed**). Manifolds on which transport of a basis around a closed loop reverses orientation are **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 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 similarly reached from $\mathbf{e}_x \wedge \mathbf{e}_y \wedge \mathbf{e}_z$: it is an element of a left-handed basis.

Definition 1.17. 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 1.13 above, the simple n-form $\mathbf{d}x^1 \wedge \cdots \wedge \mathbf{d}x^n$, 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 $\mathbf{v}_1, \ldots \mathbf{v}_n$. With $p = n$ in eq. (1.37), this is readily computed as the determinant of all the vector components. There is also a sign involved, with $+$ corresponding to the orientation of the vectors being the same as that of the basis. We then say that this volume is oriented. Because it changes sign under interchange of any two basis vectors, we recognise it as a pseudoform.

Definition 1.18. In general coordinates u^i on a n-dim manifold, we define the **volume pseudoform**:

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

where the x^i form an *orthonormal* basis, usually Cartesian, and we have used eq. (1.28) with $|g| = 1$ for orthonormal bases.

1.5.3 The Hodge dual of a p-form

To a vector $\mathbf{v}$ on a n-dim metric-endowed space corresponds a pseudoform $\boldsymbol{\sigma}$ of rank n-1:

$$\boldsymbol{\sigma} \;=\; v^\nu\, \epsilon_{\nu\mu_1\cdots\mu_{n-1}}\, \mathbf{d}u^{\mu_1} \wedge \ldots \wedge \mathbf{d}u^{\mu_{n-1}} \qquad (\mu_1 < \ldots < \mu_{n-1}) \tag{1.39}$$

which, like $\mathbf{v}$, has n (independent!) components. **Notation alert**: Starting from this equation, $\epsilon_{\nu\mu_1\cdots\mu_{n-1}}$ is to be understood as the a component of the Levi-Civita pseudo-tensor constructed in Appendix C as: $\sqrt{|g|}\,[\mu_1 \ldots \mu_n]$, with $[\mu_1 \ldots \mu_n]$ the Levi-Civita symbol, which on its own is not the component of a tensor.

In 3-dim $\mathbb{R}^3$ this is the pseudo-2-form:

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

Also, there must be a mapping between the 1-form dual to $\mathbf{v}$ and the (n-1)-pseudoform. Generalising to $\boldsymbol{\Omega}^p$:

Definition 1.19. Let $\mathcal{V}^n$ be endowed with a metric and a coordinate 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 $\star\boldsymbol{\sigma} = (\star\boldsymbol{\sigma})_{|\nu_1\ldots\nu_{n-p}|}\mathbf{d}u^{\nu_1} \wedge \ldots \wedge \mathbf{d}u^{\nu_{n-p}}$, where we introduce the compact notation: $|\mu_1 \ldots \mu_p| \equiv \mu_1 < \ldots < \mu_p$, and with components:

$$(\star\boldsymbol{\sigma})_{\nu_1\ldots\nu_{n-p}} \;=\; \frac{1}{p!}\, \sigma_{\mu_1\ldots\mu_p}\, \epsilon^{\mu_1\ldots\mu_p}{}_{\nu_1\cdots\nu_{n-p}} \qquad \left(=\; \sigma^{|\mu_1\ldots\mu_p|}\, \epsilon_{\mu_1\ldots\mu_p\nu_1\ldots\nu_{n-p}}\right) \tag{1.40}$$

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

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

One immediate application of eq. (1.40) is that the n-dim volume form is the Hodge dual of the 0-form 1:

$$\star 1 \;=\; \epsilon_{|\mu_1\cdots\mu_n|}\,\mathbf{d}u^{\mu_1}\wedge\cdots\wedge\mathbf{d}u^{\mu_n} \;=\; \sqrt{|g|}\,\mathbf{d}u^1\wedge\cdots\wedge\mathbf{d}u^n$$

A very important consequence of the fact that $\star\star\boldsymbol{\sigma} \;=\; \pm\boldsymbol{\sigma}$ 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, a 4-form with components $\sigma_{\mu\nu\lambda\rho}$ 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 to its dual pseudo-1-form whose *four* components 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:

$$\begin{aligned}(\star T)_\lambda &= \epsilon_{|\mu\nu|\lambda}\,T^{|\mu\nu|} \qquad \text{in } \mathbb{R}^3\\ (\star T)_{\lambda\rho} &= \epsilon_{|\mu\nu|\lambda\rho}\,T^{|\mu\nu|} \qquad \text{in } \mathbb{R}^4\end{aligned}$$

With an orthonormal metric, it is not hard to work out that in the first line $\star T_1 = T^{23}$, $\star T_2 = T^{31}$, and $\star T_3 = T^{12}$, so that the 1-form dual to $\mathbf{T}$ contains only the three independent components of $\mathbf{T}$.

In the 3-dim Euclidean space of p-forms of example 1.12, $\{\mathbf{d}x^1, \mathbf{d}x^2, \mathbf{d}x^3\}$ 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 for $\boldsymbol{\Omega}^1$ and $\boldsymbol{\Omega}^2$, respectively. The two bases are each other's Hodge dual. In fact, we can Hodge-dualise the (co)basis: $\star(\mathbf{d}u^{\mu_1}\wedge\cdots\wedge\mathbf{d}u^{\mu_p}) \;=\; \epsilon^{\mu_1\cdots\mu_p}{}_{|\mu_{p+1}\cdots\mu_n|}\,\mathbf{d}u^{\mu_{p+1}}\wedge\cdots\wedge\mathbf{d}u^{\mu_n}$, (or divide by $(n-p)!$ if the summations are unrestricted), in which case the components are not changed—they are just re-allocated to basis elements of $\boldsymbol{\Omega}^{n-p}$. There are corresponding expressions for Hodge-dualising coordinate bases or the components of contravariant tensors, as illustrated by the above example.

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:

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

If $\boldsymbol{\sigma}$ corresponds to the vector $\mathbf{u}$ and $\boldsymbol{\tau}$ to $\mathbf{v}$ via the metric, this says that: $\star(\mathbf{u}\wedge\mathbf{v}) = \mathbf{u}\times\mathbf{v}$, or, with eq. (1.40), $(\mathbf{u}\times\mathbf{v})^\mu \;=\; \frac{1}{2}g^{\mu\rho}\epsilon_{\nu\lambda\rho}(u^\nu v^\lambda - u^\lambda v^\nu) = g^{\mu\rho}\epsilon_{\rho\nu\lambda}u^\nu v^\lambda$. So when calculating a vector product, one is implicitly taking a Hodge dual, 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:

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

1.6 Exterior Calculus

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-field components can be calculated in the same way.

For general (r,s) tensors, however, because of point-dependent bases, defining differentiation requires extra structure, called a **connection**, or **covariant derivative**. It turns out that in four dimensions there is an infnite number of ways to construct such a connection. A few, however, have gained favour as "natural". Here we only discuss a particular type of differentiation that acts only on p-forms and offers a neat unification of the ideas of gradient, divergence and curl in vector calculus, *without the need for a connection*.

1.6.1 Exterior derivative

We introduce the **exterior derivative operator**[†], $\mathbf{d}$, which acts on p-forms $\boldsymbol{\sigma} = \sigma_{|\mu_1\ldots\mu_p|}\mathbf{d}x^{\mu_1}\wedge\ldots\wedge\mathbf{d}x^{\mu_p}$, 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) If $\boldsymbol{\sigma}$ is a 0-form, ie. just a function, then $\mathbf{d}\boldsymbol{\sigma}$ is the 1-form gradient of that function, introduced in section 1.2.2.

(b) Without loss of generality, working in some coordinate system:

$$\mathbf{d}\boldsymbol{\sigma} \coloneqq \partial_{\mu_0}\sigma_{|\mu_1\ldots\mu_p|}\,\mathbf{d}x^{\mu_0}\wedge\mathbf{d}x^{\mu_1}\wedge\ldots\wedge\mathbf{d}x^{\mu_p} = \left(\mathbf{d}\sigma_{|\mu_1\ldots\mu_p|}\right)\wedge\mathbf{d}x^{\mu_1}\wedge\ldots\wedge\mathbf{d}x^{\mu_p}$$

Each component of $\mathbf{d}\boldsymbol{\sigma}$ with indices in increasing order is a sum of $p+1$ terms:

$$(\mathbf{d}\boldsymbol{\sigma})_{\mu_1\ldots\mu_{p+1}} = \delta^{\nu_0\nu_1\ldots\nu_p}_{\mu_1\ldots\mu_{p+1}}\,\partial_{\nu_0}\sigma_{|\nu_1\ldots\nu_p|} = \partial_{\mu_1}\sigma_{|\mu_2\ldots\mu_{p+1}|} - \partial_{\mu_2}\sigma_{|\mu_1\mu_3\ldots\mu_{p+1}|} + \partial_{\mu_3}\sigma_{|\mu_1\mu_2\ldots\mu_{p+1}|} - \ldots$$

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

(d) $\mathbf{d}(\boldsymbol{\sigma}\wedge\boldsymbol{\tau}) = \mathbf{d}\boldsymbol{\sigma}\wedge\boldsymbol{\tau} + (-1)^p\,\boldsymbol{\sigma}\wedge\mathbf{d}\boldsymbol{\tau}$ (graded Leibniz rule).

(e) $\mathbf{d}^2\boldsymbol{\sigma} = 0$ identically.

We shall not prove the graded Leibniz rule (you can do it as an EXERCISE), but the nilpotency of $\mathbf{d}$ deserves some proof:

$$\mathbf{d}^2\boldsymbol{\sigma} = \mathbf{d}\left[\partial_\nu\sigma_{|\mu_1\ldots\mu_p|}\,\mathbf{d}x^\nu\wedge\mathbf{d}x^{\mu_1}\wedge\ldots\wedge\mathbf{d}x^{\mu_p}\right] = \left[\partial_\rho\partial_\nu\sigma_{|\mu_1\ldots\mu_p|}\right]\mathbf{d}x^\rho\wedge\mathbf{d}x^\nu\wedge\mathbf{d}x^{\mu_1}\wedge\ldots\wedge\mathbf{d}x^{\mu_p} = 0$$

because of the symmetry of the partial derivatives in ρ and ν.

Example 1.18. 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:

$$\mathbf{d}f = \partial_u f\,\mathbf{d}u + \partial_v f\,\mathbf{d}v + \partial_w f\,\mathbf{d}w \tag{1.41}$$

This is valid only for a *coordinate* basis. 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. But if we insist on a basis whose elements 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:

$$\mathbf{d}f = \partial_r f\,\mathbf{d}\hat{r} + \frac{1}{r}\,\partial_\theta f\,\mathbf{d}\hat{\theta} + \frac{1}{r\sin\theta}\,\partial_\phi f\,\mathbf{d}\hat{\phi} \tag{1.42}$$

Example 1.19. The exterior derivative of a 1-form $\boldsymbol{\sigma}$ is the 2-form:

$$\begin{aligned}\boldsymbol{\theta} = \mathbf{d}\boldsymbol{\sigma} &= \partial_\mu\sigma_\nu\,\mathbf{d}x^\mu\wedge\mathbf{d}x^\nu \\ &= (\partial_\mu\sigma_\nu - \partial_\nu\sigma_\mu)\,\mathbf{d}x^\mu\wedge\mathbf{d}x^\nu \qquad (\mu<\nu)\end{aligned} \tag{1.43}$$

The components of $\boldsymbol{\theta}$ are: $\theta_{\mu\nu} = \partial_\mu\sigma_\nu - \partial_\nu\sigma_\mu$. (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 1.20. 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 $\mathbb{R}^3$? With the equivalent expression: $\mathbf{d}\boldsymbol{\sigma} = \mathbf{d}\sigma_{_\nu} \wedge \mathbf{d}x^\nu$, we obtain:

$$\begin{aligned} \mathbf{d}\boldsymbol{\sigma} &= \big(\partial_v\sigma_u\,\mathbf{d}v + \partial_w\sigma_u\,\mathbf{d}w\big) \wedge \mathbf{d}u + \big(\partial_u\sigma_v\,\mathbf{d}u + \partial_w\sigma_v\,\mathbf{d}w\big) \wedge \mathbf{d}v + \big(\partial_u\sigma_w\,\mathbf{d}u + \partial_v\sigma_w\,\mathbf{d}v\big) \wedge \mathbf{d}w \\ &= (\partial_v\sigma_w - \partial_w\sigma_v)\,\mathbf{d}v \wedge \mathbf{d}w + (\partial_w\sigma_u - \partial_u\sigma_w)\,\mathbf{d}w \wedge \mathbf{d}u + (\partial_u\sigma_v - \partial_v\sigma_u)\,\mathbf{d}u \wedge \mathbf{d}v \end{aligned} \tag{1.44}$$

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

$$\star\mathbf{d}\boldsymbol{\sigma} = \sqrt{|g|}\,\Big[(\partial_v\sigma_w - \partial_w\sigma_v)\,\mathbf{d}u + (\partial_w\sigma_u - \partial_u\sigma_w)\,\mathbf{d}v + (\partial_u\sigma_v - \partial_v\sigma_u)\,\mathbf{d}w\Big] \tag{1.45}$$

By analogy with tensor algebra results, we can recover the *contravariant* 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.

As we know all too well, the *vector* components of the curl of a *vector* in curvilinear coordinates can be quite complicated; this is largely due to our insisting on working with objects which are less natural. Exterior derivatives do not involve raising indices with a metric, and so are more natural.

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:

$$\begin{aligned} \mathbf{d}\boldsymbol{\tau} &= (\partial_u\tau_{_{vw}}\,\mathbf{d}u) \wedge \mathbf{d}v \wedge \mathbf{d}w + (\partial_v\tau_{_{wu}}\,\mathbf{d}v) \wedge \mathbf{d}w \wedge \mathbf{d}u + (\partial_w\tau_{_{uv}}\mathbf{d}w) \wedge \mathbf{d}u \wedge \mathbf{d}v \\ &= (\partial_u\tau_{_{vw}} + \partial_v\tau_{_{wu}} + \partial_w\tau_{_{uv}})\,\mathbf{d}u \wedge \mathbf{d}v \wedge \mathbf{d}w \end{aligned} \tag{1.46}$$

Now, in three-dimensions $\boldsymbol{\tau}$ can be viewed as the Hodge dual, $\boldsymbol{\tau} = \star\boldsymbol{\sigma}$, 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 $\star(\mathbf{d}u \wedge \mathbf{d}v \wedge \mathbf{d}w) = \epsilon^{123} = (-1)^{n_-}/\sqrt{|g|}$ from section C.0.2, gives:

$$(-1)^{n_-}\,\star\mathbf{d}\star\boldsymbol{\sigma} = \frac{1}{\sqrt{|g|}}\,\partial_\mu(\sqrt{|g|}\,\sigma^\mu) \tag{1.47}$$

Definition 1.20. 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 holds in any coordinates in a metric-endowed space.

The operator $\star\mathbf{d}\star$ 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, $\boldsymbol{\delta}\,\boldsymbol{\sigma} = (-1)^{n(p+1)+1}\star\mathbf{d}\star\boldsymbol{\sigma}$, and $\boldsymbol{\delta}\,\boldsymbol{\sigma} = (-1)^{n(p+1)}\star\mathbf{d}\star\boldsymbol{\sigma}$ 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, $\boldsymbol{\delta}^2 = \star\mathbf{d}\star\star\mathbf{d}\star = \pm\star\mathbf{d}^2\star = 0$.

Definition 1.21. We define the **divergence** of any p-form: $\mathbf{div}\,\boldsymbol{\sigma} := -\boldsymbol{\delta}\boldsymbol{\sigma} = (-1)^{n(p+1)+n_-}\star\mathbf{d}\star\boldsymbol{\sigma}$. This ensures consistency between eq. (1.47) and the conversion between $\star\mathbf{d}\star$ and $\boldsymbol{\delta}$. We extend eq. (1.47) to the divergence of any p-form $\boldsymbol{\sigma}$ on a n-dim space:

$$(\mathbf{div}\,\boldsymbol{\sigma})_{\mu_1\cdots\mu_{p-1}} := \frac{1}{\sqrt{|g|}}\,\partial_\nu\big(\sqrt{|g|}\,\sigma^\nu{}_{\mu_1\cdots\mu_{p-1}}\big) = \frac{1}{\sqrt{|g|}}\,\partial_\nu\big(\sqrt{|g|}\,g^{\nu\rho}\sigma_{\rho\mu_1\cdots\mu_{p-1}}\big) \tag{1.48}$$

From eq. (1.47) follows the definition of the 3-dim **Laplacian** of a scalar function f in coordinates u^i:

$$\nabla^2 f = \frac{1}{\sqrt{|g|}}\,\partial_i\big(\sqrt{|g|}\partial^i f\big) = \frac{1}{\sqrt{|g|}}\,\partial_i\big(\sqrt{|g|}\,g^{ij}\partial_j f\big) \tag{1.49}$$

1.6.2 Laplace-de Rham operator, harmonic forms, and the Hodge decomposition

Definition 1.22. The **Laplace-de Rham operator** is defined as $\boldsymbol{\Delta} = \boldsymbol{\delta}\,\mathbf{d} + \mathbf{d}\,\boldsymbol{\delta} = (\mathbf{d} + \boldsymbol{\delta})^2$. It is a mapping $\boldsymbol{\Delta} : \boldsymbol{\Omega}^p \longmapsto \boldsymbol{\Omega}^p$. When acting on a scalar function, $\boldsymbol{\Delta} = \boldsymbol{\delta}\,\mathbf{d}$; then we also speak of the **Laplace-Beltrami** operator.

It is not hard to show that it reduces to the negative of the Laplacian operator of vector analysis, ie. $\boldsymbol{\Delta} = \boldsymbol{\delta}\,\mathbf{d} = -\star\mathbf{d}\star\mathbf{d} = -\partial_i\partial^i = -\nabla^2$, 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.

Example 1.22. Let the Laplace de Rham operator act on a 1-form $\boldsymbol{\sigma}$ in Euclidean $\mathbb{R}^3$. That is, take $\boldsymbol{\Delta\sigma} = \star\mathbf{d}\star\mathbf{d}\boldsymbol{\sigma} - \mathbf{d}\star\mathbf{d}\star\boldsymbol{\sigma}$ using the conversion formula between δ and $\star\mathbf{d}\star$. Using eq. (1.45), 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-de Rham operator is related to the d'Alembertian operator $\Box := \partial_\mu\partial^\mu$: $\Delta = -\Box$. This *defines* the d'Alembertian of any p-form in Minkowski space.

Definition 1.23. A p-form $\boldsymbol{\sigma}$ is said to be **harmonic** if $\boldsymbol{\Delta\sigma} = 0$. This generalises the notion of functions being called harmonic when they satisfy the Laplace equation.

Definition 1.24. 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, since $\mathbf{d}^2 = 0$, 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, Poincaré's lemma (not proved) states that only in a **simply connected** submanifold, in which all closed curves can be shrunk to a point (no doughnuts!), does $\mathbf{d}\boldsymbol{\sigma} = 0$ entail the existence in that submanifold of a non-unique (p - 1)-form whose exterior derivative is $\boldsymbol{\sigma}$.

We quote without proof an important result of Hodge: On finite-volume (compact) manifolds without boundaries, such as S^n, or on a torus, $\boldsymbol{\Delta\sigma} = 0$ if, and only if, $\mathbf{d}\boldsymbol{\sigma} = 0$ and $\mathbf{d}\star\boldsymbol{\sigma} = 0$ (or $\boldsymbol{\delta}\,\boldsymbol{\sigma} = 0$). Harmonic forms are both closed and co-closed! 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.

Definition 1.25. Assuming a compact manifold without boundaries or, failing that, compact support (sufficiently fast fall-off at infinity), the unique **Hodge decomposition** writes a p-form $\boldsymbol{\sigma}$ as a sum of exact (closed), co-closed, and harmonic p-forms:

$$\boldsymbol{\sigma} \;=\; \mathbf{d}\boldsymbol{\alpha} \;+\; \boldsymbol{\delta}\,\boldsymbol{\beta} \;+\; \text{harmonic } p\text{-form} \tag{1.50}$$

where $\boldsymbol{\alpha}$ is a (p-1)-form and $\boldsymbol{\beta}$ is a (p+1)-form, *both non-unique*. $\mathbf{d}\boldsymbol{\alpha}$, $\boldsymbol{\delta}\,\boldsymbol{\beta}$ and the harmonic p-form in the decomposition live in orthogonal subspaces of $\boldsymbol{\Omega}^p$.

Example 1.23. Let $\mathbf{A}$ be a vector field with compact support on Euclidean $\mathbb{R}^3$. Then its 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 $\star\boldsymbol{\beta}$ is a pseudo-1-form in $\mathbb{R}^3$, $\boldsymbol{\delta}\,\boldsymbol{\beta} = \star\mathbf{d}\star\boldsymbol{\beta}$ is a 1-form. Then, from eq. (1.45) this term corresponds to the curl of a pseudovector. Therefore, we obtain *in terms of vectors*:

$$\mathbf{A} \;=\; \boldsymbol{\nabla}\phi \;+\; \boldsymbol{\nabla}\times\mathbf{M} \;+\; \mathbf{H} \tag{1.51}$$

where ϕ is a scalar field, $\mathbf{M}$ a pseudovector 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, and we have the **Helmholtz decomposition** for a vector field with compact support.

The curl of $\boldsymbol{\nabla}\phi$ vanishes *identically*, and is often called the **longitudinal** projection of $\mathbf{A}$; the divergence of $\boldsymbol{\nabla} \times \mathbf{M}$ vanishes *identically*, and we can call it the **transverse** projection of $\mathbf{A}$. Thus, $\boldsymbol{\nabla} \cdot \mathbf{A}$ 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 fields. If $\boldsymbol{\nabla} \cdot \mathbf{A} = 0$ *everywhere*, we can conclude that $\mathbf{A}$ is purely transverse, since then ϕ in eq. (1.51) satisfies the Laplace equation *everywhere*, so must vanish if it has compact support.

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

Let $\mathbf{F} \in \mathbf{\Omega}^2$ on Minkowski (pseudo-Euclidean) $\mathbb{R}^4$. Demand that $\mathbf{F}$ be exact and with compact support. 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 $\mathbf{dF} = 0$. It is clear from Poincaré's lemma that $\mathbf{dF} = 0$ knows nothing about $\mathbf{A}$: 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-3-form $\mathbf{d} \star \mathbf{F}$, as a "source" $\boldsymbol{\mathcal{J}}$, with compact support and Hodge dual 1-form $\mathbf{J} = \star\boldsymbol{\mathcal{J}}$. Then we have the inhomogeneous equation:

$$\mathbf{d} \star \mathbf{F} = 4\pi\, \boldsymbol{\mathcal{J}} \tag{1.52}$$

If we take the exterior derivative of the equation, the left-hand side vanishes *identically*, and the right-hand side becomes: $\mathbf{d}\,\boldsymbol{\mathcal{J}} = 0$. This is better known as: $\mathbf{div}\,\mathbf{J} = 0$, a conservation law for the source.

What we have constructed is Maxwell's theory, with $\mathbf{F}$ the Faraday 2-form, $\mathbf{A}$ the electromagnetic potential 1-form, and $\mathbf{J}$ the 4-current. Our treatment assumes a mostly positive metric, as in MTW or Griffiths' *Introduction to Electrodynamics*. With a mostly negative metric, there is a minus sign on the right-hand side of eq. (1.52).

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

$$\star\mathbf{d} \star \mathbf{F} = 4\pi \star\boldsymbol{\mathcal{J}} = 4\pi\,\mathbf{J} \qquad \Longleftrightarrow \qquad \mathbf{div}\,\mathbf{F} = -4\pi\mathbf{J} \tag{1.53}$$

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

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

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. (1.52) 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. (1.40)), whereas its 3-form version does not. The price, of course, is that the 3-form version has 64 explicit components, although still only four independent ones.

It is worth noting that, although $\mathbf{dF} = 0$ and the source equation (1.52) completely determine $\mathbf{F}$, $\mathbf{A}$ is determined only up to an additive term $\mathbf{d}f$, where f is an arbitrary differentiable function.

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. This is the same as $\mathbf{d} \star (\star\mathbf{F}) = 0$ whose Hodge dual is $\boldsymbol{\delta} \star \mathbf{F} = 0$. In other words, the divergence of $\star\,\mathbf{F}$ vanishes, only four equations. Actually, this is a general, easily shown property (EXERCISE): whenever the exterior derivative of a p-form in some manifold vanishes, so does the codifferential of its dual, and vice-versa.

[†]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.

Another great advantage of writing Maxwell's equations as $\mathbf{dF} = 0$ and $\mathbf{d} \star \mathbf{F} = 4\pi \boldsymbol{\mathcal{J}}$ is that, provided the source is smoothly varying, they are formally the same in curved spacetime! Only when divergences are written in index notation are covariant derivatives involving a connection needed. Even in index notation, the first equation does not involve the connection; it does not even require a metric.

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

1.7 Integrals of Differential (Pseudo)Forms

As we figure out the meaning of $\int \boldsymbol{\sigma}^p$, where we use the notation $\boldsymbol{\sigma}^p$ to show explicitly the rank of a p-form, we shall discover that pretty much any integral in n-dim calculus is the integral of some (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, $\int \star f$, 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 f, of course, is a pseudo-n-form whose single *independent* component is f. Then:

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

where u are general coordinates and $\mathbf{d}^n x$ is the volume pseudo-n-form in Cartesian coordinates. Then we define:

Definition 1.26.

$$\int_V f(\mathbf{x})\,\mathbf{d}x^1 \wedge \cdots \wedge \mathbf{d}x^n \;:=\; \int_V f(\mathbf{x})\,\mathrm{d}x^1 \cdots \mathrm{d}x^n \;=\; \int_V f(\mathbf{x})\,\mathrm{d}^n x \tag{1.54}$$

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

We know that a curve C can be parametrised in terms of some real parameter $t \in [a, b]$. Then, if $\boldsymbol{\alpha}$ is a 1-form field on $\mathbb{R}^3$, eq. (1.54) 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 described in $\mathbb{R}^3$ 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 1.25. Integral of a 2-form over a surface

Let S be some surface described in $\mathbb{R}^3$ with three coordinate functions $u^i(x^1, x^2)$. The surface is parametrised with $(x^1,\, x^2) \in \mathbb{R}^2$, with two basis vectors $\boldsymbol{\partial}_{x^i} \equiv \boldsymbol{\partial}_i$ along the x^i direction, for which some orientation has been defined as positive. What meaning can we give to the integral of a 2-form field $\boldsymbol{\beta}$ over S? From the chain rule and eq. (1.54) 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 of S in parameter space. The two coordinate vectors (see section A.3), $\partial_1 \mathbf{u}$ and $\partial_2 \mathbf{u}$, are tangent to S at every point, and are usually linearly independent, so form a basis for the space tangent to the surface at a point, with no metric required as yet.

The Hodge dual of $\boldsymbol{\beta}$, 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. (1.37), 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 (1.18) can be identified with the volume pseudo-form $\mathbf{d}^3u$. Then our integral can be written:

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

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 S, we can choose the unit $\hat{\mathbf{n}}$ normal to the surface so that $\hat{\mathbf{n}}$ 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 function $\mathbf{d}^3u(\mathbf{B}, \partial_1\mathbf{u}, \partial_2\mathbf{u})$ is the normal component of $\mathbf{B}$ multiplied by the surface of the parallelogram defined by the coordinate vectors (see example 1.13) . Defining the surface element $\mathrm{d}S \equiv |\partial_1\mathbf{u} \times \partial_2\mathbf{u}|$, there comes:

$$\int_S \boldsymbol{\beta} \;=\; \int B_\mathrm{n}\, \mathrm{d}S \;=\; \int \mathbf{B}\cdot\mathrm{d}\mathbf{S} \tag{1.55}$$

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-Cartan 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:

$$\int_V \mathbf{d}\boldsymbol{\omega} \;=\; \int_{\partial V} \boldsymbol{\omega} \tag{1.56}$$

A technicality is that both V and ∂V must have compatible orientations. But no metric is required. 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 1.26. At the end of example 1.24 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. (1.44) and example 1.25, $\int_S \mathbf{d}\boldsymbol{\omega} = \int_S \boldsymbol{\nabla} \times \mathbf{A} \cdot \mathrm{d}\mathbf{S}$, and we recover the well-known Kelvin-Stokes formula.

Finally, when $\boldsymbol{\omega}$ is a pseudo-2-form in Euclidean $\mathbb{R}^3$ and S a surface enclosing a volume V, we recover the divergence theorem: $\int_V \boldsymbol{\nabla} \cdot \mathbf{B}\, \mathrm{d}V = \oint_S \mathbf{B} \cdot \mathrm{d}\mathbf{S}$, from examples 1.21 and 1.25.

Note that a metric is required for the translation from the Stokes-Cartan theorem to the divergence and Kelvin-Stokes theorems in vector calculus.

1.8 Maxwell Differential Forms in 3 + 1 Dimensions

With $\mathbf{F}$ the Faraday 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:

$$\mathbf{F} \;=\; F_{|\mu\nu|}\mathrm{d}x^\mu \wedge \mathbf{d}x^\nu \;\coloneqq\; \boldsymbol{\mathcal{E}} \wedge \mathbf{d}t \,+\, \boldsymbol{\mathcal{B}} \tag{1.57}$$

where:

$$\boldsymbol{\mathcal{E}} \;\coloneqq\; F_{10}\mathbf{d}x^1 \,+\, F_{20}\mathbf{d}x^2 \,+\, F_{30}\mathbf{d}x^3 \qquad \boldsymbol{\mathcal{B}} \;\coloneqq\; F_{12}\mathbf{d}x^1 \wedge \mathbf{d}x^2 \,+\, F_{31}\mathbf{d}x^3 \wedge \mathbf{d}x^1 \,+\, F_{23}\mathbf{d}x^2 \wedge \mathbf{d}x^3 \tag{1.58}$$

Now, formally, $\mathbf{d} = \vec{\mathbf{d}} + \mathbf{d}t \wedge \partial_t$. where $\vec{\mathbf{d}}$ denotes the *3-dim* exterior derivative. Then Maxwell's $\mathbf{d}\mathbf{F} = 0$ becomes:

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

The plus sign in the round brackets is the result of applying the commutation formula eq. (1.32) to the 1-form $\mathbf{d}t$ and the 2-form $\boldsymbol{\mathcal{B}}$. In three dimensions, then, the homogeneous Maxwell equation gives rise to:

$$\vec{\mathbf{d}}\boldsymbol{\mathcal{B}} \;=\; 0 \qquad\qquad \vec{\mathbf{d}}\boldsymbol{\mathcal{E}} \,+\, \partial_t\boldsymbol{\mathcal{B}} \;=\; 0 \tag{1.59}$$

Eq. (1.59) is *metric-independent*, and will thus hold in any spacetime in a coordinate basis.

The Hodge duals of eq. (1.59) can be written as:

$$\mathbf{div} \star\boldsymbol{\mathcal{B}} \;=\; 0 \qquad\qquad \star\vec{\mathbf{d}}\boldsymbol{\mathcal{E}} \,+\, \partial_t\star\boldsymbol{\mathcal{B}} \;=\; 0$$

If we identify the contravariant components of the pseudo-1-form $\star\boldsymbol{\mathcal{B}}$ with the usual components of the magnetic-field pseudo-vector, and use eq. (1.45), we see that these are equivalent to the homogeneous Maxwell equations in their vector-calculus form: $\boldsymbol{\nabla} \cdot \mathbf{B} = 0$ and $\boldsymbol{\nabla} \times \mathbf{E} + \partial_t \mathbf{B} = 0$.

We 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. and 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 Maxwell equation requires much more care, and is treated in Appendix D.

Appendices

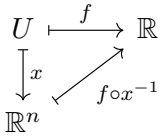
A Manifolds, Curves, and Tangent Spaces

A.1 Manifolds and coordinates

Definition A.1. A **differentiable manifold**, or just **manifold**, M is a set of elements ("points"), all of which have an **open ball** (or **neighbourhood**) around them in M, such that M can be entirely covered by a union of possibly overlapping open (without boundary) subsets U_i, each mapped in a one-to-one way to an *open* subset of $\mathbb{R}^n$ by a non-unique, **differentiable coordinate map**: $x : U_i \longmapsto \mathbb{R}^n$. Each (U_i, x) forms a **coordinate chart** (**local coordinate system**), and an **atlas** is any collection of charts that covers the whole M. Also, on any non-empty overlap $U_i \cap U_j \subset M$, only so-called **compatible** charts (U_i, x) and (U_j, y), ie., those for which the (often non-linear!) **coordinate transformation** $y \circ x^{-1} : \mathbb{R}^n \underset{x^{-1}}{\longmapsto} U_i \cap U_j \underset{y}{\longmapsto} \mathbb{R}^n$, and its inverse, between them are differentiable, are allowed.

We should also define what we mean when we say that a function f, (e.g., a coordinate function) on U is differentiable.

As illustrated in the diagram on the right, f is a map from U to $\mathbb{R}$. But in calculus we only know how to differentiate maps from $\mathbb{R}^n$ to $\mathbb{R}$. This suggests that we make a detour via $\mathbb{R}^n$. The combined map, or **chart representative**, $f \circ x^{-1}$ goes form $\mathbb{R}^n$ to $\mathbb{R}$ and its derivative can be computed if it exists. Therefore, when we say that f is differentiable. we mean that $f \circ x^{-1}$ is differentiable.



The minimum number n of parameters—each a map $x^i : U \longmapsto \mathbb{R}$ $(1 \leq k \leq n)$—that uniquely specify every point in U is its **dimension**.

Example A.1.

- $\mathbb{R}^n$ can be promoted to a manifold; it can be covered with just one coordinate chart, Cartesian (**standard**) coordinates. Other charts are possible, eg. polar coordinates .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 $\Phi : \mathcal{V} \longmapsto \mathbb{R}^n$. Conversely, however, a manifold is *not* equipped with a vector-space structure! Even in $\mathbb{R}^n$, adding the curvilinear coordinates of two points as if they were vector components does not make sense.
- 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; but the *open* unit ball, $\sum x_i^2 < 1$, is a manifold. So is the **unit n-sphere**, $S^n \coloneqq \left\{x \in \mathbb{R}^{n+1} : \sum_{}^{n+1} x_i^2 = 1\right\}$, embedded in $\mathbb{R}^{n+1}$.

  The unit circle in the plane $\mathbb{R}^2$, S^1, is an archetypal example of (closed) curves in $\mathbb{R}^2$, and the 2-dim sphere, S^2 in $\mathbb{R}^3$, of (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 $x = \pm 1$ 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 repeat with two submanifolds in correspondence with $x > 0$ and $x < 0$, and an atlas with four charts has been constructed.

  S^1 also has a local coordinate, θ, related to x by: $\theta = \tan^{-1}(y/x) = \tan^{-1}\left(\sqrt{1/x^2 - 1}\right)$. To avoid any point being mapped to more than one value, θ must map to $[0,\, 2\pi)$ in $\mathbb{R}$.

  An atlas can also be constructed for S^2 out of patches similar to those for S^1 for each of the Cartesian $\pm x > 0$, $\pm y > 0$, and $\pm z > 0$. Each point in each patch unambiguously maps to $\mathbb{R}^2$.

On S^2 we could also use spherical coordinates θ and ϕ that map to he region of $\mathbb{R}^2$: $\theta \in (0,\, \pi)$, $\phi \in [0,\, 2\pi)$, with the poles removed since ϕ is undetermined there. More patches are needed to cover S^2.

Notice that we have looked at S^1 and S^2 as being embedded in a higher-dimensional manifold, $\mathbb{R}^2$ and $\mathbb{R}^3$. Whitney's embedding theorems guarantee that any smooth M^n is a submanifold of $\mathbb{R}^{m>2n}$, with stronger results in restricted cases. Embedding curves and surfaces in, eg., $\mathbb{R}^3$ is great for intuition, but we are more interested in their *intrinsic* properties which should be independent of the embedding manifold.

Less technically, it is usually enough to view a manifold as a set that can be parametrised in a smooth way.

A.2 Curves, directional derivatives and vectors

Definition A.2. A **curve** $\Gamma_\lambda : \mathbb{R} \longmapsto M$ on a manifold M is a *mapping*, at least C^1 (no kinks!), of each value of a real parameter λ to a unique point $\mathcal{P}$ in M: $\Gamma(\lambda) = \mathcal{P}$. Then λ is a coordinate on Γ.

Definition A.3. Let $\Gamma_\lambda \in M$ be a curve parametrised by a coordinate λ. The **velocity** at a point $\mathcal{P}$ with coordinate λ_0 on this curve, is the linear map $\mathbf{v}_{(\Gamma_\lambda,\mathcal{P})} : C^\infty(M) \longmapsto \mathbb{R}$, defined as:

$$\mathbf{v}_{(\Gamma_\lambda,\mathcal{P})}(f) \;\coloneqq\; \mathrm{d}_\lambda(f\circ\Gamma_\lambda)\big|_{\lambda_0} \;=\; \mathrm{d}_\lambda f\big|_{\lambda_0} \qquad \big(\mathrm{d}_\lambda \coloneqq \mathrm{d}/\mathrm{d}\lambda\big) \tag{A.1}$$

where $f \in C^\infty(M)$ is a smooth function on the manifold M. The composition $f\circ\Gamma_\lambda : \mathbb{R} \underset{\Gamma}{\longmapsto} M \underset{f}{\longmapsto} \mathbb{R}$ is equivalent to $f(\lambda)$, with $\lambda \in \mathbb{R}$, and, being $\mathbb{R} \to \mathbb{R}$, it is differentiable using standard calculus.

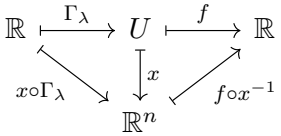
Such a curve is only one of an infinite number containing $\mathcal{P}$ each with their own velocity at $\mathcal{P}$. For instance: $\mathbf{w}_\Theta(f) = \mathrm{d}_\mu f\big|_{\mu_0}$, and $\Theta(\mu_0) = \mathcal{P}$. We say that velocities are **tangent** to the manifold at $\mathcal{P}$.

There is no notion of speed at this time, beause we do not yet know what the length of a vector is.

For some purposes we also need to view the curve Γ as embedded in a region U of a manifold M parametrised by coordinate functions denoted collectively by $x : U \longmapsto \mathbb{R}^n$, with $x\circ\Gamma$ describing what the curve "looks like" in M. Now insert the identity map $x^{-1}\circ x$ into eq. A.1:

$$\mathbf{v}_{(\Gamma_\lambda,\mathcal{P})}(f) \;=\; \mathrm{d}_\lambda\big[(f\circ x^{-1})\circ(x\circ\Gamma_\lambda)\big]_{\lambda_0}$$

As illustrated in the diagram on the right, both mappings $f\circ x^{-1}$ and $x\circ\Gamma_\lambda$ are to and from $\mathbb{R}$ and $\mathbb{R}^n$, so that their deivative can each be computed with standard methods (see below).



Now apply the multidimensional chain rule:

$$\mathbf{v}_{(\Gamma_\lambda,\mathcal{P})}(f) \;=\; \sum_\nu^n \Big[\,\mathrm{d}_\lambda(x^\nu\circ\Gamma_\lambda)\big|_{\lambda_0}\,\big[\partial_\nu(f\circ x^{-1})\big]\Big|_{x^\nu(\mathcal{P})}\,\Big] \tag{A.2}$$

where the index ν runs over the number of local coordinates that specify each point in M, and the coordinate functions $x^\nu(\lambda)$ parametrise the curve Γ_λ in M.

Equation (A.2) contains unexpected information. We already know that we can write $\mathrm{d}_\lambda(x^\nu\circ\Gamma_\lambda)\big|_0 = \mathrm{d}_\lambda x^\nu\big|_0$. And we may suspect that the second factor has something to do with the partial derivative of f. Indeed, as illustrated above, f acts on U, not $\mathbb{R}$, without any reference to a local coordinate system on U, and we would not know how to take its derivatives *with respect to some coordinates*. In order to have calculable derivatives, we take a detour via $\mathbb{R}^n$: $U \underset{x}{\longmapsto} \mathbb{R}^n \underset{f\circ x^{-1}}{\longmapsto} \mathbb{R}$. 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.

Dropping the test-function f, we rewrite eq. (A.2) for the velocity vector in more succinct form:

$$\mathbf{v}_{(\Gamma_\lambda,\mathcal{P})} \;=\; \sum_{\nu}^{n} \mathrm{d}_\lambda x^\nu(\lambda)\big|_{\lambda_0}\,(\boldsymbol{\partial}_\nu)_{\mathcal{P}} \tag{A.3}$$

In elementary calculus, this would be written as the chain rule:

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \;=\; \frac{\partial x^\nu}{\mathrm{d}\lambda}\Big|_{\lambda_0}\frac{\partial}{\partial x^\nu}$$

What meaning is assigned to $\boldsymbol{\partial}_\nu$ is discussed in the main text. Here, we construct the space where $\mathbf{v}_{(\Gamma_\lambda,\mathcal{P})}$ lives.

A.3 The tangent space at a point in a manifold

Definition A.4. A **tangent space** $\mathcal{T}_\mathcal{P}$ to a manifold M^n at a point $\mathcal{P}\in M^n$, is a set that intersects M^n only at $\mathcal{P}$ and 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. $\mathcal{T}_\mathcal{P}=\mathbb{R}^n$.

It is important to keep in mind that the tangent spaces, $\mathcal{T}_\mathcal{P}$ and $\mathcal{T}_\mathcal{Q}$ at points $\mathcal{P}$ and $\mathcal{Q}$ in M, even though they are both $\mathbb{R}^n$, are completely different and have zero intersection (no vectors in common). To relate elements of tangent spaces at close points, one needs to introduce extra structure: the **connection**, but this lies beyond the scope of this course.

Similarly, one defines a **cotangent space** $\mathcal{T}^*_\mathcal{P}$ containing all the covectors on M^n at $\mathcal{P}$.

The existence of the tangent space as a *vector* space is an assertion that should be justified. First, let us specify what is meant by addition and s-multiplication on a tangent space $\mathcal{T}_\mathcal{P}$.

Definition A.5. The addition operation on $\mathcal{T}_\mathcal{P}$ is a map, $\mathcal{T}_\mathcal{P}+\mathcal{T}_\mathcal{P}\longmapsto\mathcal{L}(C^\infty(M),\mathbb{R})$, 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, $\mathbb{R}\times\mathcal{T}_\mathcal{P}\longmapsto\mathcal{L}(C^\infty(M),\mathbb{R})$, 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, $\mu:\mathbb{R}\longmapsto\mathbb{R}$, of λ: $\mu=a\lambda+\lambda_0$, with λ now the parameter of a curve Θ_λ such that $\Theta(\lambda)=\Gamma(\mu)$ Therefore, $\Theta_\lambda(0)=\Gamma_\mu(\lambda_0)=\mathcal{P}$. As in definition A.2 we can write: $\Gamma(\mu)=\Gamma_\mu\circ\mu(\lambda)$. Insert this information into the expression for the velocity for Θ_λ at $\mathcal{P}$:

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

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, to discuss tangent spaces, we have not had to refer to coordinate charts. Unfortunately, when it comes to proving that adding two velocities in $\mathcal{T}_\mathcal{P}$ gives a velocity in $\mathcal{T}_\mathcal{P}$, we cannot add curve mappings directly since this has no meaning. Instead, assume curves Γ_μ and Θ_ν in some open subset $U\subset M$ parametrised by coordinate functions x, going through $\mathcal{P}$ at values $\mu=\lambda_1$ and $\nu=\lambda_2$. Construct a curve Σ_λ, also in U, such that:

$$(x\circ\Sigma_\lambda)(\lambda) \;=\; (x\circ\Gamma_\mu)(\lambda_1+\lambda)+(x\circ\Theta_\nu)(\lambda_2+\lambda)-(x\circ\Gamma_\mu)(\lambda_1)$$

The seemingly obvious cancellation in this expression is not allowed because the coordinate functions are not necessarily linear and do not distribute over the additions in the arguments. At $\lambda = 0$, however, the cancellation does occur, leaving $\Sigma_\lambda(0) = \Theta_\nu(\lambda_2) = \mathcal{P}$, so that Σ does run through $\mathcal{P}$ at $\lambda = 0$, as demanded by our construction.

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

$$\begin{aligned} \mathrm{d}_\lambda(x^\alpha \circ \Sigma_\lambda)\big|_{\lambda=0} &= \mathrm{d}_\lambda\Big[(x^\alpha \circ \Gamma_\mu)(\lambda_1 + \lambda) + (x^\alpha \circ \Theta_\nu)(\lambda_2 + \lambda) - (x^\alpha \circ \Gamma)(\lambda_1)\Big]\Big|_{\lambda=0} \\ &= \mathrm{d}_{\lambda_1+\lambda}(x^\alpha \circ \Gamma)\big|_{\lambda_1}\, \mathrm{d}_\lambda(\lambda_1 + \lambda)\big|_0 + \mathrm{d}_{\lambda_2+\lambda}(x^\alpha \circ \Theta)\big|_{\lambda_2}\, \mathrm{d}_\lambda(\lambda_2 + \lambda)\big|_0 \\ &= \mathrm{d}_{\lambda_1+\lambda}(x^\alpha \circ \Gamma)\big|_{\lambda_1} + \mathrm{d}_{\lambda_2+\lambda}(x^\alpha \circ \Theta)\big|_{\lambda_2} \end{aligned} \tag{A.4}$$

Now go back to our expression (A.3) for the velocity in coordinates x. The first factor on the right has been evaluated in eq. (A.4) and, running the chain of equalities in eq. (A.3) backward, there comes:

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

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! The result does not depend on whatever coordinate chart we might have used in the intermediate steps, which is entirely legitimate so long as the final results are independent of coordinates.[†]

One final point: tangent spaces exist quite independently of any embedding we may (or not) choose for a manifold. For instance, a plane tangent to a point on the sphere S^2 embedded in $\mathbb{R}^3$ should not be viewed as being in $\mathbb{R}^3$; it still exists in the absence of the embedding.

B Transformation of Vector Components Between Coordinate Systems

Let (U_1, x) and (U_2, y) be two overlapping charts (see definition A.1) 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):

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

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 $\mathbf{v}$:

$$v_y^\nu = \partial_{x^\mu} y^\nu\big|_{x_\mathcal{P}}\, v_x^\mu \tag{B.2}$$

[†]For an accessible yet rigorous discussion of manifolds and tangent spaces, see Frederic Schullers's first five lectures at the 2015 International Winter School on Gravity and Light in Linz (Austria), available on *YouTube* — especially lectures 3 and 5. Another comprehensive treatment, with many examples, is in the lecture notes for the MAT 367 course on Differential Geometry in our Mathematics Department, available at: https://www.math.toronto.edu/laithy/3672021/DiffGeomNotes_short.pdf.

C Levi-Civita Symbol and Tensor

C.0.1 The Levi-Civita symbol

Definition C.1. 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. (1.21)), as:

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

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

Linear algebra tells us that the determinant of a $n\times n$ matrix L is a product of its elements antisymmetrised with respect to rows (or columns):

$$\det \mathsf{L} \;\coloneqq\; \epsilon_{\nu_1\dots\nu_n}\, L^{\nu_1}{}_1\cdots L^{\nu_n}{}_n \tag{C.1}$$

If the Levi-Civita symbol is to be a tensor, the transformation laws on its components demand that:

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

This is the case when L is a 3-dim rotation or a Lorentz-boost matrix, under which $\epsilon_{\mu_1\dots\mu_n}$, like $\delta^\mu{}_\nu$, is invariant.

C.0.2 The Levi-Civita pseudotensor

The Levi-Civita *symbol* does not transform as a tensor. Consider, however, the volume pseudoform of definition 1.18. By inspection it is a n-form with the single *independent* component $(\mathbf{d}^n u)_{1\dots n} = \sqrt{|g|}$. Its other components are obtained by antisymmetrising with the Levi-Civita symbol, which we shall now denote by $[\mu_1\dots\mu_n]$ to avoid any confusion later. That is:

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

Thus, the right-hand side is the component of a covariant *pseudotensor*, $\boldsymbol{\epsilon}$, of rank n. Henceforth, whenever we write components $\epsilon_{\mu_1\dots\mu_n}$, they are to be understood as $\sqrt{|g|}\,[\mu_1\dots\mu_n]$, so that $\epsilon_{1\dots n} = \sqrt{|g|}$.

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

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

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

Both $\epsilon^{\nu_1\dots\nu_n}$ and $\epsilon_{\mu_1\dots\mu_n}$ being antisymmetric, we can relate the permutation symbol to the Levi-Civita pseudotensor with: $\epsilon^{\nu_1\dots\nu_n}\epsilon_{\mu_1\dots\mu_n} = a\,\delta^{\nu_1\dots\nu_n}_{\mu_1\dots\mu_n}$. To determine a, we use: $\epsilon^{1\dots n}\epsilon_{1\dots n} = (-1)^{n_-}$, and there comes:

$$\begin{aligned}
\epsilon^{\nu_1\dots\nu_n}\,\epsilon_{\mu_1\dots\mu_n} \;&=\; (-1)^{n_-}\,\delta^{\nu_1\dots\nu_n}_{\mu_1\dots\mu_n} = (-1)^{n_-}\begin{vmatrix} \delta^{\nu_1}{}_{\mu_1} & \cdots & \delta^{\nu_1}{}_{\mu_n} \\ \vdots & & \vdots \\ \delta^{\nu_n}{}_{\mu_1} & \cdots & \delta^{\nu_n}{}_{\mu_n} \end{vmatrix} \\
\frac{1}{(n-p)!}\,\epsilon^{\nu_1\dots\nu_p\nu_{p+1}\dots\nu_n}\,\epsilon_{\mu_1\dots\mu_p\nu_{p+1}\dots\nu_n} \;&=\; (-1)^{n_-}\delta^{\nu_1\dots\nu_p}_{\mu_1\dots\mu_p} \qquad \text{(unrestricted sums)}
\end{aligned} \tag{C.2}$$

In a Euclidean 3-dim space with an orthonormal metric, $n_- = 0$, 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.

D Three-dim Inhomogenous Maxwell Equations in the p-form Formalism

Going from the simple 4-dim formalism to three dimensions is more complicated than for the homogeneous equations, because the Hodge dual in the 4-divergence 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 $F^{\mu\nu} = g^{\mu\alpha}g^{\nu\beta}F_{\alpha\beta}$:

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

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. (1.52) yields the two 3-dim Maxwell field equations:

$$\vec{\mathbf{d}}{}^*\boldsymbol{\mathcal{E}} = 4\pi\boldsymbol{\rho}, \qquad \vec{\mathbf{d}}{}^*\boldsymbol{\mathcal{B}} = \mathbf{j} + \partial_t{}^*\boldsymbol{\mathcal{E}} \tag{D.1}$$

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 CHAPTER II — A BRIEF INTRODUCTION TO GROUP THEORY

One of the most beautiful and useful concepts in physics and, indeed, mathematics, **symmetry** identifies patterns connected with a characteristic behaviour of objects, usually an invariance, under transformations. A problem where a symmetry exists 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 is understood. Equally important, a continuous symmetry 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 gives rise to energy conservation; and isotropy of space (invariance under rotations) to angular-momentum conservation. Conservation of electric charge is embodied in the **local gauge invariance** of Maxwell's equations.

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.

2.1 Introducing the Notion of Group (BF 10.1)

2.1.1 Some basic definitions

Definition 2.1. Let G be a set of *distinct* objects endowed with an *associative*, but not necesssarily commutative, binary composition law, or operation, denoted by $\star$ (or $\circ$). We say that G is a **group** if:

- $\forall\,(a,\,b)\,\in\,G,\;\; a\star b\,\in\,G$ (this is called **closure**);
- 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 $a^{-1}\star a = a\circ a^{-1} = e$.

The composition law is often called **group multiplication**, a term we shall try to avoid because it almost irresistibly evokes the much narrower ordinary multiplication. There immediately follows a constraint on any composition law. Let $G = \{a_i\}$, with i a positive integer or continuous index. Then, for a fixed element a_i, the set $\{a_i\star a_j\}$, with j running over all the elements of G, must itself be G, ie., it must contain all elements of the group once, and only once. Indeed, suppose that $a_i\star a_j = a_i\star a_k$ for some $j,\;k$. Since a_i must have a unique inverse, this forces $a_j = a_k$. A similar argument can be made for fixed a_j in $\{a_i\star a_j\}$.

Definition 2.2. When $\star$ is commutative, ie. $a\star b = b\star a,\;\forall\,(a,\,b)\,\in\,G$, 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. It is **discrete** if it is **countable**, ie., if each element can be associated with a unique positive integer. All finite groups are discrete, but infinite discrete groups exist. Non-discrete infinite groups are called **continuous**.

Example 2.1. Consider the set $\{e,\,a,\,a^2,\,\ldots,\,a^{n-1}\}$, where $a^p \coloneqq a\star a\star\ldots$, and n is the smallest integer such that $a^n = e$. The set is closed under $\star$, and $(a^p)^{-1} = a^{n-p}\;\forall\,p$. All a^p are distinct, for supposing $a^p = a^q$, we would have $a^{p-q} = e$, with $p-q<n$, and n would not be the smallest integer such that $a^n = e$. We conclude that the set is a group called Z_n (sometimes C_n), the **cyclic group** of order n. When n is even, only $a^{n/2}$ is its own inverse; when n is odd, each element other than e is paired with a distinct inverse. Thus, Z_n can have at most one self-inverse element.

Let g belong to a group G of order n. There must be an integer m such that $g^m = e$. Then we say that g itself is **of order** m. If $m<n$ the group is not cyclic, but $\{e,\,g,\,\ldots,\,g^{m-1}\}$ is a group Z_m. g and its inverse have the same order, $\forall\,g\,\in\,G$. If all elements of a group are their own inverse (order 2), the group is Abelian (EXERCISE).

Some other groups: $\mathbb{C}$ under addition ($e=0$, $a^{-1}=-a$); $\mathbb{C}-\{0\}$ under multiplication ($e=1$, $z^{-1}=1/z$); the set, $GL(n,\,\mathbb{C})$, of all complex $n\times n$ matrices with non-zero determinant under *matrix* multiplication; the n complex roots of 1 under multiplication. Exercise: spot any discrete and cyclic groups in these examples.

It is important to keep in mind that a given set may be a group under one operation, but not under another. Thus, $\mathbb{Z}$ is a group under addition with $e=0$ and $a^{-1}=-a$, but it is not a group under multiplication.

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 $(ij)^{\text{th}}$ 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.

To satisfy the above constraint on the group composition law, any column or row of a Cayley table must contain all elements of the group once, and only once. The ordering of the rows and columns is arbirary.

Constructing Cayley tables for finite groups is easy, if tedious for large groups. Let us do it for $n = 2,\, 3,\, 4$:

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

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

The tables for $n = 2$ and 3 are the only ones possible. Thus, finite groups of order 2 and 3 are cyclic. The case $n = 4$, however, opens up more possibilities: Choosing $a \star a = a^2 = b$, as above, the table is that of the cyclic group Z_4. But we could take $a^2 = c$; or $a^2 = e$, in which case we can further choose $b^2 = a$ or $b^2 = e$, 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 re-labelling $b \longleftrightarrow c$ 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 they are really those of Z_4.

The last table is genuinely different. It belongs to a group $\{e,\, a,\, b,\, a\star b\}$ called the **4-group**—aka Felix Klein's Vierergruppe V—in which every element is its own inverse (so of order 2), with the fourth element constructed out of the other two non-identity elements (otherwise V would be cyclic!). An example is D_2, the symmetry group of a 2-d rectangle centered on the origin:, with the identity, one rotation by π, and two reflections about the axes as elements.

The foregoing illustrates very nicely two important features of groups:

- **Generators of a group**

  Definition 2.4. A set of **generators** of a group G is any subset of G from which all other elements of G can be obtained by repeated compositions of the generators among themselves. G must contain *all* the distinct compositions of its generators, including with themselves.

  For instance, we can say that if a generates Z_n, a^p also generates Z_n provided p and n have no common divisor (EXERCISE). Then any such a^p can be taken on its own as the generator of Z_n. The 4-group is obtained from two generators. EXERCISE: construct a Cayley table for the group: $\{e,\, a,\, b,\, b^2,\, a\star b,\, b\star a\}$.

  Another example is a rotation by $\pi/6$ as the generator of the finite group of rotations by $k\pi/6$ ($0 \le k \le 11$) about the same axis.

  EXERCISE: Is it possible to construct a group of order 6 with all its elements of order 2?

- **Isomorphisms**

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

  Definition 2.5. If there exists a one-to-one mapping between all the elements of one finite group $\{G, \circ\}$ and all the elements of another finite group $\{H, \star\}$ such that under this mapping these groups have identical Cayley tables, then the mapping is an **isomorphism**, and G and H are **isomorphic**: $G \cong H$.

Another definition is more apt for continuous groups, which do not have a Cayley table as such:

Definition 2.6. 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:

- the group of permutations of two objects (S_2), the group of rotations by π around the z axis, and the group $\{1,\ -1\}$ (under multiplication);
- the group of complex numbers and the group of vectors in a plane, both under addition;
- the 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 2.7. 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 group. We look at three of them.

2.2.1 Special Ternary Compositions: Conjugacy Classes

Definition 2.8. 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, 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, since $x \circ y \in G$, b is conjugate to c :

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

Definition 2.9. The subset of elements of a group which are conjugate to one another form a **conjugacy**, or **equivalence**[†], **class**, often abbreviated to just **class**. The systematic way of constructing the class for any element a_i of a group is to form the set:

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

Then e is always in a class by itself, and each element of an Abelian group is the sole element in its class. eg., Z_n and the four-group.

Classes are disjoint: they have no common element (EXERCISE: show this). Thus, they **partition** the group.

Elements in the same class share some properties. In particular, they must all be of the same order (EXERCISE). 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: $\{e,\ a,\ b,\ b^2,\ a \star b,\ b \star a\}$.

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

2.2.2 Subgroups

Definition 2.10. 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: $H \subseteq G$. H is **proper** if it is **non-trivial** (ie. not e) and if $H \subset G$ (ie. $H \neq G$). The subgroups of a group may have more elements than e in common.

We have already seen that *any element g of order $m < n$ of G generates a cyclic subgroup $Z_m \subset G$.*

Example 2.2. The four-group V has the proper Z_2 subgroups: $\{e,\, a\}$, $\{e,\, b\}$, and $\{e,\, c = a \star b\}$, which are isomorphic. By inspection, the group of order 6 $\{e,\, a,\, b,\, b^2,\, a\star b,\, b\star a\}$ contains the proper subgroup $Z_3 = \{e,\, b,\, b^2\}$.

Notation alert: Henceforth, we drop the cumbersome star (circle) whenever there is no risk of confusion with usual multiplication. Also, if H and H' are two subsets of $\{G, \star\}$, we can write $H\,H'$ for $\{h\,h'\}$ $h \in H,\ h' \in H'$.

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

Definition 2.11. A subgroup $N \subseteq G$ is **invariant** (or **normal**) if $N = G\,N\,G^{-1}$ or, more precisely, if $g\,h\,g^{-1} \in N\ \forall\, h \in N$ and $\forall\, g \in G$. Alternate notation: $N \lhd G$, $G \rhd N$.

EXERCISE: Show that $\forall\, g_i \in G$ the set with distinct elements $g_i^{-1} g_j^{-1} g_i g_j$ forms an invariant subgroup of G.

Definition 2.11 is sometimes written $G\,N = N\,G$, but it does not mean that an invariant subgroup must be Abelian (though it *can* be). It means that if $h_i \in N$ and $g \in G$, there is *some* element $h_j \in N$ such that $g\,h_i = h_j\,g$.

Example 2.3. Because the four-group V is Abelian, its non-trivial subgroups, $\{e,\, a\}$, $\{e,\, b\}$, $\{e,\, a\,b\}$, are all invariant. Subgroups of any Abelian group are invariant.

Since classes and normal groups are both defined by conjugation, it is hardly surprising that they are related. Indeed, 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; this holds for all classes, which are disjoint; so only complete classes can be in H. Conversely, let a subgroup $H \subset G$ be a union of complete classes; therefore, $g\,h\,g^{-1} \in H\ \forall\, g \in G$, which is precisely the definition of a normal subgroup.

Definition 2.12. A **simple group** has no invariant subgroup other than itself and the identity.

2.2.3 Cosets (BF 10.3)

Definition 2.13. 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 $g\,H$ 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. 2.11. In particular, the right and left cosets of any Abelian subgroup are identical.

Example 2.4. 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 parallel to H, and we would say in this language that it is a left (or right) coset of H 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. Thus, we can say that the coset space of a subgroup $H \subset G$ provides a **partition** of G.

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.

It follows (why?) that $e\,H = H$ is the only coset of a subgroup H that is a group.

2.2.4 Lagrange's Theorem and quotient groups

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, 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, eg., Z_5, Z_7, and that such a group has no non-trivial subgroup.

Also, if its order n is odd, a group (not only the cyclic ones!) cannot contain any self-inverse element, for such an element must generate a Z_2 subgroup whose order, 2, is forbidden by Lagrange's Theorem.

The converse of Lagrange's Theorem does not hold in general: eg., the group of even permutations of four objects, A_4, with 12 elements, has no subgroup of order 6. The theorem only gives the *possible* orders of subgroups. There are stronger conditions on the order and number of the subgroups of G stipulated by the **Sylow Theorems**, which lack of time prevents us from exploring.

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

> **Definition 2.14.** The set of all left cosets of $H \subset G$, each considered as a whole, is called a **factor space** for H. *Note that the elements of this space are the cosets themselves, each considered as a whole*, not any individual element within a coset.

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

> **Definition 2.15.** To an invariant subgroup N of G is associated a **factor group** of G, G/N. Its elements are N and all its cosets as sets (not the elements of N or of the cosets!). Its order is the order of G divided by the order of N, hence the name **quotient group** often used for G/N.

To show that the factor space of an invariant subgroup is a group, we note that for any coset $g\,N$, $(g\,N)\,N = g\,N\,N = g\,N$, and $N\,(g\,N) = g\,N\,N = g\,N$, where we have used the associativity of the group product and the invariant nature of N. This establishes N as 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 we do not need to distinguish between the elements of subgroups of a group. Much of the usefulness of normal subgroups comes from providing a quick way to find factor groups.

2.2.5 Direct Products

> **Definition 2.16.** Let H_1 and H_2 be subgroups of G with $H_1 \cap H_2 = e$, 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.
>
> **Example 2.5.** The four-group introduced in section 2.1.2 can be seen as $Z_2 \times Z_2$, or $\{e,\, a\} \times \{e,\, b\} = \{e,\, a,\, b,\, a\,b\}$. But $Z_4 \neq Z_2 \times Z_2$, even though Z_2 is a normal subgroup of Z_4, with $Z_2 = Z_4/Z_2$!

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:

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

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 (with normal multiplication as group product).

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 2.17.

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

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. 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 i after l steps. When this occurs, we speak of a l**-cycle**. More precisely:

Definition 2.18. 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. Warning: do not confuse the symbols in a l-cycle with the *outcome* of a permutation in S_n!

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. An element of S_n and its inverse have the same cycle structure.

Definition 2.19. 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.

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

Definition 2.20. A **cyclic permutation of length** l has a single cycle of length $l > 1$.

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), where we start the cycle does not matter! It can be shown that S_3 and Z_6 are the only groups of order 6, up to isomorphisms.

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

2.3.2 Some 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 do not form a group, because their product is an even permutation. Other important subgroups of S_n are the cyclic groups of order n, generated by the permutations of all objects which return 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.

Transpositions being self-inverse, the other (isomorphic!) subgroups of S_3 are $\{e, \pi_2\}$, $\{e, \pi_3\}$, and $\{e, \pi_4\}$.

2.3.3 Cayley table of S_3 as an example

The group-product table of S_3 contains 36 entries, 25 of which are non-trivial. But I claim that no more than one product needs to be worked out "from scratch", with explicit permutations.

Indeed, the entries of the 2×2 sub-table for the π_5 and π_6 rows and columns must be even permutations (they are the group product of even permutations). The diagonals cannot be e, so as to avoid repetition. Next, the non-diagonal elements of rows and columns corresponding to π_2, π_3 and π_4 must be π_5 or π_6, the only even permutations other than e. To fill in this sector only requires calculating one group product, say, $\pi_2\,\pi_3$:

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

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 table-building rules. Similarly, $\pi_2\,\pi_3\,\pi_3 = \pi_2 = \pi_5\,\pi_3$, and the rest of the π_5 and π_6 rows is determined. The comes (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 Cayley 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:

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

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

What we have shown is that $\pi_{a_i}\,\pi_{a_j} \;=\; \pi_{a_i a_j}$; in other words, by definition 2.6, 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.*

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

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:

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

How did we obtain $\left(\begin{smallmatrix} a_1 & a_2 & \dots & a_n \\ b_{a_1} & b_{a_2} & \dots & b_{a_n} \end{smallmatrix}\right)$ in the second line from $\left(\begin{smallmatrix} 1 & 2 & \dots & n \\ b_1 & b_2 & \dots & b_n \end{smallmatrix}\right)$ in the last member of the first line? Well, a_1 must occur in some slot on the top line of the latter; since the order of the slots in the permutation is arbitrary, we move that slot to first position and rename 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}.

This leads to the important 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.* Classes being disjoint, each class of S_n is associated with a unique cycle structure of its elements. But 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. $A_4 \subset S_4$, $Z_4 \subset S_4$). Also, all elements of a class of S_n must have their inverse in the same class; can you see why?

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 $\mathcal{C}_3 = \{(1\,3\,2),\,(1\,2\,3)\}$. Since $A_3 = \{e,\,(1\,2\,3),\,(1\,3\,2)\}$ is the only non-trivial subgroup of S_3 that is the sum of *complete* classes, $\mathcal{C}_1 + \mathcal{C}_3$, A_3 is the only normal[‡] 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.

It is an instructive EXERCISE to show that A_4 has no subgroup of order 6 despite this being allowed by Lagrange's Theorem. There are several ways to proceed. Here is one that requires no result we have not seen:
Hint: after convincing yourself that such a group would take the form $\{e,\,a,\,b_1,\,b_1^{-1},\,b_2,\,b_2^{-1}\}$, with a a double transposition and $(b_1,\,b_2)$ 3-cycles, show that because the latter are of order 3, $b_1\,b_2$ is neither $b_1^{-1} = b_1^2$ nor b_2^{-1}, which leaves only one possibility since b_1 and b_2 cannot be each other's inverse. Is that possibility allowed?

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 is the number of i-cycles in the class.

Start with e, whose cycle structure can be written as a product of n 1-cycles: $e = (1)\,(2)\,\cdots\,(n)$. So its class, which always exists, would be denoted by 1^n. A transposition has 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 $\sum_i^n i\,\alpha_i = n$. In that sense the cycle structure of a class corresponds to a partition of n.

Once we have noticed this correspondence, it becomes rather easy to find the number and cycle ctructure of S_n classes. 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 sum up to n. Thus, the only possible partitions of S_3 lead 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 possible partitions of 4 give rise to the five classes (1^4), $(2\,1^2)$, (2^2), $(3\,1)$, and (4).

It is important not to confuse the cycle notation we first introduced with this standard notation which lists all the cycles in a class as a whole, including 1-cycles when they occur (whereas the other one ignores them).

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

To find the number of elements in a class of S_n, count the *distinct* ways of partitioning n numbers into its cycle structure:

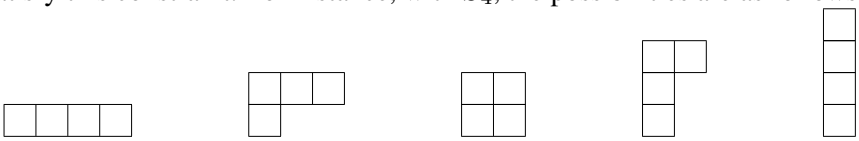
$$\frac{n!}{\alpha_1! \ldots \alpha_n!\, 1^{\alpha_1} \ldots n^{\alpha_n}} \tag{2.1}$$

where $\alpha_i!$ is the number of non-distinct ways of ordering α_i commuting cycles of a given length, and i^{α_i} is the number of equivalent orderings of the symbols inside each i-cycle occurring α_i times. 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 writing down its $24{\times}24$ Cayley table!

2.3.6 Graphical representation of classes: Young frames

A useful and visual 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 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. 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

Finding the left cosets of the subgroups of S_3 is as easy as reading rows in its Cayley table. Take the subgroup $H = \{e,\, \pi_2\}$; its left cosets by π_k are $\pi_k \{e,\, \pi_2\} = \{\pi_k,\, \pi_k\, \pi_2\}$ $(1 \leq k \leq 6)$. Only three are *distinct*: $\{e,\, \pi_2\}$, $\{\pi_3,\, \pi_6\}$, $\{\pi_4,\, \pi_5\}$. Following Definition 2.14, this set of cosets is the factor space for H. The same arguments apply to the subgroups $\{e,\, \pi_3\}$ and $\{e,\, \pi_4\}$.

Turn now to the remaining non-trivial proper subgroup, $A_3 = \{e,\, \pi_5,\, \pi_6\}$, of all even permutations in S_3. Its left cosets are $\{\pi_k, \pi_k\, \pi_5, \pi_k\, \pi_6\}$. For instance, $\pi_2 \{e,\, \pi_5,\, \pi_6\} = \{\pi_2,\, \pi_3,\, \pi_4\}$, 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 these left and right cosets are identical, another way of saying that $\{e,\, \pi_5,\, \pi_6\}$ is invariant, as we had found by 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 Cayley 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 isomorphic to Z_2. 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. We wish to flesh out the rather abstract ideas and tools we have introduced. We shall find that linear operators on vector spaces (most often, on a Hilbert space) provide us with this connection.

2.4.1 Action of a group from the left and from the right

Let G be a group of linear transformations $\mathcal{T}_{\mathrm{g}}$ on square-integrable functions that live in a space called the **carrier** space. Introduce a set of operators, $\{T_{\mathrm{g}}\}$, with each T_{g} acting on a set of parameters that characterise $g \in G$.

> **Definition 2.21.** We distinguish between an **action from the left**, $[\mathcal{T}_{\mathrm{g}} f](\mathbf{x}) \coloneqq f(\mathrm{g}^{-1}\mathbf{x})$, and an **action from the right**, $[\mathcal{T}_{\mathrm{g}} f](\mathbf{x}) \coloneqq f(\mathbf{x}\,\mathrm{g})$, $\forall\, f$. Note that, here, the operators $\mathcal{T}_{\mathrm{g}}$ act on the *functions* (not on $\mathbf{x}$!). In what follows we always use the left action.

Why did we define the left action of $g \in G$ as $\mathrm{g}^{-1}\,\mathbf{x}$, and not $\mathrm{g}\,\mathbf{x}$? Denote by $\mathcal{T}_{\mathrm{g}_i\mathrm{g}_j}$ the transformation associated with $\mathrm{g}_i\,\mathrm{g}_j \in G$. Then, with $\mathrm{g}_i = i$ and $\mathrm{g}_j = j$ in subscripts so as to declutter the notation:

$$\big[\mathcal{T}_{ij} f\big](\mathbf{x}) \;=\; f\big((\mathrm{g}_i\,\mathrm{g}_j)^{-1}\mathbf{x}\big) \;=\; f(\mathrm{g}_j^{-1}\,\mathrm{g}_i^{-1}\,\mathbf{x}) \;=\; \big[\mathcal{T}_j\, f\big](\mathrm{g}_i^{-1}\,\mathbf{x}) \;=\; \big[\mathcal{T}_i\mathcal{T}_j\, f\big](\mathbf{x})$$

which means that the $\mathcal{T}$ operators do form a group; but what if instead:

$$\big[\mathcal{T}_{ij} f\big](\mathbf{x}) \;=\; f(\mathrm{g}_i\,\mathrm{g}_j\,\mathbf{x}) \;=\; \big[\mathcal{T}_i\, f\big](\mathrm{g}_j\,\mathbf{x}) \;=\; \big[\mathcal{T}_j\mathcal{T}_i\, f\big](\mathbf{x})$$

Something awkward has happened: if we *write* the left action as $\mathrm{g}\,\mathbf{x}$, the associated transformations do *not* form a group! So, as a matter of *notational* consistency, we should always write $\mathrm{g}^{-1}\,\mathbf{x}$ for the left action, which is indeed what BF do, without much explanation.

2.4.2 Matrix representations of a group (BF10.4)

> **Definition 2.22.** A **representation** D of a group G is a homomorphic mapping of the group elements 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 rank of its matrices or, equivalently, the dimension of the carrier space on which it acts.

Even addition can be represented by matrix multiplication: $\mathsf{D}_\alpha\,\mathsf{D}_\beta = \mathsf{D}_{\alpha+\beta}$, with α and β two values of a group parameter, eg. the matrix $\mathsf{D}_v = \left(\begin{smallmatrix} 1 & 0 \\ v & 1 \end{smallmatrix}\right)$. Do you recognise the transformation that applies it to the vector $\binom{t}{x}$?

The matrices $GL(n,\mathbb{C})$ of rank n can be thought of as the set of all invertible linear transformations on a vector space of complex-valued functions $\mathcal{V} = \{f(\mathbf{x})\}$. We have: $\mathbf{x} = x^i\,\mathbf{e}_{\mathrm{i}}$, with $\{\mathbf{e}_{\mathrm{i}}\}$ a basis for $\mathcal{V}$ and x^i the components of $\mathbf{x}$ in the basis; the subscript on basis vectors labels a whole vector, not a component of the vector.

Let us focus on the transformations $T_{\mathrm{g}}(\mathbf{x})$. Then the (left) action of $\mathrm{g} \in G$ is expressed as:

$$T_{\mathrm{g}}(\mathbf{x}) \;=\; \mathrm{g}^{-1}\,\mathbf{x} \;=\; x^i\,\mathrm{g}^{-1}\,\mathbf{e}_{\mathrm{i}} \;=\; x^i\left[\mathbf{e}_{\mathrm{j}}\,(D^{\mathrm{L}}_{\mathrm{g}^{-1}})^j{}_i\right] \;=\; \mathbf{e}_{\mathrm{j}}\left[(D^{\mathrm{L}}_{\mathrm{g}^{-1}})^j{}_i\,x^i\right] \tag{2.2}$$

2.4.3 Non-unicity of group representations

One might hope to define an algorithm that would churn out *the* representation of each element 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, of dimension 1, by mapping these matrices to the number 1. 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 one-dim representation. Of course, no one will claim 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 $\mathsf{D'}$ obey the same product rules as the D matrices.

Definition 2.23. Representations connected by a similarity transformation are said to be **equivalent** if the transformation matrix is *the same for all group elements*. They differ only by a choice of basis.

Example 2.6. Consider the continuous group, called $SO(2)$, of rotations in a plane. We parametrise a rotation by $g = R_\alpha$ such that $R_\alpha\,\phi = \phi - \alpha$. This corresponds to a counterclockwise rotation of the standard basis in $\mathbb{R}^2$ by α (passive transformation), so that a vector initially at angle ϕ is at angle $\phi - \alpha$ in the rotated basis; equivalently, rotate the vector by $-\alpha$ in the *initia*l basis.

We find representations for the left action. One method uses Def. 2.21 (with $\mathcal{T}_{\rm g} = \mathcal{R}_\alpha$) and eq. (2.2):

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

We look for a set of functions of ϕ which, under $\mathcal{R}_\alpha$, transform into linear combinations of themselves. 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 left $\mathbf{D}(\alpha)$ matrix:

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

Well, that's the 2-dim left **defining (fundamental)** representation for $SO(2)$, probably the most often used. But it is not the only one! If instead $f_1 = \mathrm{e}^{\mathrm{i}\phi}$, $f_2 = \mathrm{e}^{-\mathrm{i}\phi}$, the same procedure would yield:

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

so here is another two-dim representation. But it is equivalent because the transformation $\mathbf{S}^{-1}\,\mathbf{D}^{(1)}\,\mathbf{S}$, with the matrix $\mathbf{S} = \begin{pmatrix} 1 & \mathrm{i} \\ \mathrm{i} & 1 \end{pmatrix}/\sqrt{2}$, diagonalises $\mathbf{D}^{(1)}$ into $\mathbf{D}^{(2)}$, $\forall\,\alpha$, ie. for all elements of $SO(2)$.

And there are more: *each* linearly independent function $\mathrm{e}^{\mathrm{i}\alpha}$ 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 $\mathbf{D}^{(4)}$. Obviously, there is something special about $\mathrm{e}^{\pm\mathrm{i}\alpha}$. Before we discover what it is, let us look at another instructive example.

Example 2.7. 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, ie., those whose transpose is their inverse. In fact, $S_3 \subset O(3)$! Then, from eq. (2.2), $\pi_k^{-1}\mathbf{x} = x^i\mathbf{e}_j D^j{}_i(\pi_k^{-1}) = \big(x^i D_i{}^j(\pi_k)\big)\mathbf{e}_j$ so as to view the permutations as a shuffling of the *components* of $\mathbf{x}$ (written as one-row matrices), and we have:

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

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

Such a faithful, n-dim left defining (fundamental) representation can be constructed for any S_n.

Now, I claim that there exists another (two-dimensional!) representation of S_3, which is not faithful:

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

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

Indeed, the products of these matrices are consistent with the group product of S_3 in its Cayley table.

Even less faithful, but no less acceptable, is the one-dim representation of S_n obtained by mapping its permutations to their parity values. For S_3:

$$\mathbf{D}^{(3)}(\pi_1) \;=\; \mathbf{D}^{(3)}(\pi_5) \;=\; \mathbf{D}^{(3)}(\pi_6) \;=\; 1$$
$$\mathbf{D}^{(3)}(\pi_2) \;=\; \mathbf{D}^{(3)}(\pi_3) \;=\; \mathbf{D}^{(3)}(\pi_4) \;=\; -1$$

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.

2.4.4 The regular representation of finite groups

Definition 2.24. The **regular** representation of the left action of a finite group G is the set of matrices $\mathbf{D}^{\mathrm{L}}_{\mathrm{g}}$, with $\mathrm{g} \in G$, derived from a group product, such that:

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

The regular representation is seen to be closely related to the Cayley table of the group. Its dimension is equal to N_{G}, the order of the group, and it is faithful. We can also see that $D^j{}_i(e) = \delta^j{}_i$, ie. $\mathbf{D}^{\mathrm{L}}(e) = \mathbf{I}$. 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.

A word of caution: do not confuse the dimension of a representation, ie. of its 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 Unitary representations (BF10.6)

A representation $\mathbf{D}_{\mathrm{g}}$ is unitary if $\mathbf{D}^{\dagger}_{\mathrm{g}} = \mathbf{D}^{-1}_{\mathrm{g}}$), $\forall\, \mathrm{g} \in G$. In terms of matrix elements, $D_{ij}(\mathrm{g}^{-1}) = D^*_{ji}(\mathrm{g})$. For example, $\mathbf{D}^{(3)}$ and $\mathbf{D}^{(4)}$ for $SO(2)$ are unitary. Regular representations are also unitary.

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 the positive semi-definite (ie., with eigenvalues $\lambda_n > 0$) matrix: $\mathbf{S}^2 = \sum_{\mathrm{g}} \mathbf{D}^{\dagger}_{\mathrm{g}} \mathbf{D}_{\mathrm{g}}$, such that $\mathbf{D}'_{\mathrm{g}} = \mathbf{S}\,\mathbf{D}_{\mathrm{g}}\,\mathbf{S}^{-1}$ is unitary $\big($EXERCISE—first, show that $\mathbf{D}^{\dagger}_{\mathrm{g}'} \mathbf{S}^2 \mathbf{D}_{\mathrm{g}'} = \mathbf{S}^2$, then apply $\mathbf{S}^{-1}$ on the left and on the right$\big)$. *Any representation of a finite group is equivalent to a unitary representation*. This is also true for certain infinite (continuous) groups, such as compact Lie groups.

2.4.6 Invariant Spaces and Kronecker sum

To understand what relationship may exist between representations, it is time to bring in a very useful concept:

Definition 2.25. Let $\{f^{(i)}\}$ be a subspace $\mathcal{H}^{(i)}$ of the carrier space $\mathcal{H}$ of functions on which the linear transformations $\mathcal{T}_{\mathrm{g}}$ associated with a group G act. If, $\forall\, f^{(i)} \in \mathcal{H}^{(i)}$ and $\forall\, \mathrm{g} \in G$, $\mathcal{T}_{\mathrm{g}}\, f^{(i)} \in \mathcal{H}^{(i)}$, the subspace is **invariant under** G. Also, it can be shown that if a subspace of the carrier space is invariant under a unitary representation, its complement must also be invariant.

Definition 2.26. 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 zero intersection. 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 $\mathcal{H}^{(2)}$.

2.4.7 Reducible and irreducible representations (BF10.5)

Definition 2.27. If some function space $\mathcal{H}$ has a *proper* subspace invariant under a representation $\mathbf{D}$ of G, in the sense of Def. 2.25, then the representation is said to be **reducible**.

When the n-dimensional function space $\mathcal{H}$ has proper invariant subspaces, it means that there are at least two subspaces in $\mathcal{H}$, each of which has its own set of linearly independent functions that 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 $\mathcal{H}^B$ is invariant, as it always is in cases of interest to physics (see section 2.4.5 just below), then a block-diagonal representation matrix $\mathbf{D}_\mathrm{g}$, with block submatrices $\mathbf{D}^A_\mathrm{g}$ and $\mathbf{D}^B_\mathrm{g}$, maps vectors $\begin{pmatrix} A \\ B \end{pmatrix}$ to $\begin{pmatrix} A' \\ B' \end{pmatrix}$ where $A' \in \mathcal{H}^A$, $B' \in \mathcal{H}^B$. Also, since:

$$\begin{pmatrix} \mathbf{D}^A_\mathrm{g} & 0 \\ 0 & \mathbf{D}^B_\mathrm{g} \end{pmatrix} \begin{pmatrix} \mathbf{D}^A_{\mathrm{g}'} & 0 \\ 0 & \mathbf{D}^B_{\mathrm{g}'} \end{pmatrix} = \begin{pmatrix} \mathbf{D}^A_\mathrm{g}\,\mathbf{D}^A_{\mathrm{g}'} & 0 \\ 0 & \mathbf{D}^B_\mathrm{g}\,\mathbf{D}^B_{\mathrm{g}'} \end{pmatrix} = \begin{pmatrix} \mathbf{D}^A_{\mathrm{g}\,\mathrm{g}'} & 0 \\ 0 & \mathbf{D}^B_{\mathrm{g}\,\mathrm{g}'} \end{pmatrix}$$

$\mathbf{D}^A_\mathrm{g}$ and $\mathbf{D}^B_\mathrm{g}$ do preserve the group product, as they should. $\mathbf{D}^A_\mathrm{g}$ has dimension d, and $\mathbf{D}^B_\mathrm{g}$ dimension n - d.

Then, if all the matrices $\mathbf{D}_\mathrm{g}$ in a representation can be brought into diagonal-block form by the *same* similarity transformation, the representation is reducible to lower-dimensional representations composed of the block matrices.

If no such proper subspace exists. the representation will be called **irreducible**.

Definition 2.28. 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 $\mathbf{D}_\mathrm{g}$ is **fully reducible**, in the sense that:

$$\mathbf{D}_\mathrm{g} \;=\; \bigoplus_{i=1}^{N} a_i\,\mathbf{D}^{(i)} \;= a_1\,\mathbf{D}^{(1)}_\mathrm{g} \,\oplus\, a_2\,\mathbf{D}^{(2)}_\mathrm{g} \,\oplus\, \cdots \oplus\, a_N\,\mathbf{D}^{(N)}_\mathrm{g} \tag{2.3}$$

where a_i is the number of times (its **multiplicity**) the irreducible representation $\mathbf{D}^{(i)}_\mathrm{g}$ occurs in the direct sum, and N is the number of different irreducibl;e representations in the direct sum.

It can be shown that every representation of a *finite* group is either irreducible or fully reducible.

Going back to $SO(2)$, the $\mathbf{D}^{(2)}$ representation we have obtained is clearly fully reducible to the irreducible representations $\mathbf{D}^{(3)} = \mathrm{e}^{\mathrm{i}\alpha}$ and $\mathbf{D}^{(4)} = \mathrm{e}^{-\mathrm{i}\alpha}$, so we can write it as $\mathbf{D}^{(2)} = \mathbf{D}^{(3)} \oplus \mathbf{D}^{(4)}$.

Example 2.8. The 5-dimensional representation, $\mathbf{D}^{(4)}$, we have constructed for S_3 in Example 2.7 is (by construction) reducible since it is in block-diagonal form, so $\mathbf{D}^{(4)} = \mathbf{D}^{(1)} \oplus \mathbf{D}^{(2)}$. What about that last two-dimensional representation $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$? The first is already in block-diagonal form and the second can be diagonalised to $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$, Therefore, we obtain two 1-dim irreducible representations, one identical to the identity representation $\mathbf{D}^{(5)} = 1$, and the other the "parity" representation $\mathbf{D}^{(3)}$. Then we can write: $\mathbf{D}^{(4)} = \mathbf{D}^{(1)} \oplus \mathbf{D}^{(3)} \oplus \mathbf{D}^{(5)}$. What about the (left) defining representation of S_3, $\mathbf{D}^{(1)}$: is it reducible?

The defining representation of S_N has dimension N. This always reducible representation reduces to a 1-d representation and a (N - 1)-dimensional irreducible representation. To see how this comes about, let $(x_1, \ldots, x_N)$ be a set of coordinates in the carrier space of a 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; the subspace thus qualifies as the carrier space of the 1-dim irreducible representation of S_N that in section 2.4.8 will be 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:

$$Y_{j-1} = \frac{x_1 + \ldots + x_{j-1} - (j-1)\,x_j}{\sqrt{j(j-1)}} = \frac{(x_1 - x_j) + \ldots + (x_{j-1} - x_j)}{\sqrt{j(j-1)}} \qquad 2 \le j \le N$$

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 j^{th} one. This allows us to identify the representation with another irreducible representation of S_N that we will label (N-1 1), and the defining representation can be written as $(N) \oplus (N\text{-}1\ 1)$.

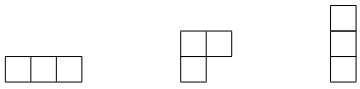
The defining representation, $\mathbf{D}^{(1)}$, of S_3 is reducible to two irreducible representations, $\mathbf{D}^{(5)} = 1$ and a set of six 2-dim orthogonal matrices, three with determinant $+1$ (rotations in a plane by angles 0, $\pm\pi/3$) and three with determinant -1, thus showing that $S_3 \subset O(2)$!. As expected, $\mathbf{D}^{(4)}$ is fully reducible. Can you see why these irreducible representations could not all be one-dimensional?

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

2.4.8 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, as will be shown later, 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 Young diagrams for irreducible respresentations of S_N look exactly like those for 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 take them as their number in rows from top to bottom. For S_3, this gives:



(3) (2 1) 1^3

The sequence of representation labels is 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 of S_N. 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 $\psi_{(1\ldots N)} \coloneqq \psi_1(x_1)\cdots\psi_1(x_N)$ spans a one-dimensional subspace which contains functions which are obviously completely symmetric and invariant under *any* of the N! possible permutations of the variables. Thus, our subspace qualifies as an invariant subspace for the regular representation, and it makes sense to associate it with the 1-dim irreducible identity representation which has the same matrix, 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-dim 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. 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 2.9. 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 any column and to the right along any row. The number of ways this can be done is the dimension of the representation. For instance, the $(2\,1)$ Young diagram of S_3 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, however, a much more 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 2.29. 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 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 an irreducible representation is the order of S_N, $N!$, divided by the product of the N hook lengths for the associated diagram.

Definition 2.30. 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.

The Young diagram of a **self-conjugate** irreducible representation is identical to its transpose.

2.5 Schur's Lemmas and Symmetry in the Language of Group Theory (BF10.6)

We now present two fundamental results of group theory which provide useful criteria for the irreducibility of representations as well as insight into symmetries, and which lead to relations that help to classify representations.

2.5.1 What is a symmetry in the language of group theory?

Consider a linear operator L such that, $\forall\, f \in \mathcal{H}$, $[L_{\mathbf{x}}\, f](\mathbf{x}) \;=\; h(\mathbf{x}) \in \mathcal{H}$. Under a group G, $[\mathcal{T}_{\mathrm{g}} L_{\mathbf{x}} \mathcal{T}_{\mathrm{g}}^{-1}][\mathcal{T}_{\mathrm{g}}\, f](\mathbf{x}) \;=\; [\mathcal{T}_{\mathrm{g}} h](\mathbf{x})$, and $[L_{\mathbf{x}'} f](\mathbf{x}') = h(\mathbf{x}')$, so that L transforms under G as: $L_{\mathbf{x}'} = \mathcal{T}_{\mathrm{g}}\, L_{\mathbf{x}}\, \mathcal{T}_{\mathrm{g}}^{-1}$.

Definition 2.31. When $\mathcal{T}_{\mathrm{g}}\, L_{\mathbf{x}}\, \mathcal{T}_{\mathrm{g}}^{-1} \;=\; L_{\mathbf{x}}$, $\forall\, \mathrm{g} \in G$, L is said to be **invariant under the action of group** G. Since this condition can also be written as $\mathcal{T}_{\mathrm{g}}\, L \;=\; L\, \mathcal{T}_{\mathrm{g}}$, $\forall\, \mathrm{g} \in G$, then an operator that is invariant under a group of transformations *must commute with all those transformations*. If also $[\mathcal{T}_{\mathrm{g}} f](\mathbf{x}) = f(\mathbf{x})$, f is invariant under G itself as well (eg., $f(r)$ in polar coordinates under rotations).

If L has eigenvalues and eigenfunctions and is invariant under G, then there should exist a set $\{f^i\}$ such that:

$$L\,(\mathcal{T}_{\mathrm{g}}\, f^i) \;=\; \mathcal{T}_{\mathrm{g}}\, L\, f^i \;=\; \lambda\,(\mathcal{T}_{\mathrm{g}}\, f^i) \tag{2.4}$$

Thus, if f^i is an eigenfunction of L, so is $\mathcal{T}_{\mathrm{g}}\, f^i$, with *the same eigenvalue*. Therefore, the distinct $\mathcal{T}_{\mathrm{g}}\, f^i$ *are all degenerate* with respect to λ. If λ is degenerate, there are N degenerate f^i for that λ, which form a basis for a N-dim subspace of functions, characterised by λ. The $\mathcal{T}_{\mathrm{g}}\, f^i$ will then be some linear combination of $\{f^j\}$. so that the transformed eigenfunctions $\mathcal{T}_{\mathrm{g}} f^i$ also form a basis for the same subspace of functions as that spanned by the eigenfunctions of L: the subspace is invariant under the action of the group, in the sense of Def. 2.25! With summation up to N over repeated indices implied:

$$\mathcal{T}_{\mathrm{g}}\, f^i \;=\; f^j\, (D_{\mathrm{g}})_j{}^i \tag{2.5}$$

Whenever we find (or observe) a set of degenerate eigenfunctions for some operator, *the operator is invariant under the action of a group, and these functions live in an variant subspace connected with an irreducible representation of the group*.

2.5.2 Schur's Lemmas

Consider a matrix $\mathbf{M} \in GL(n, \mathbb{C})$ with eigenvectors $\mathbf{A}$ belonging to eigenvalue λ. Let $\mathbf{M}$ and $\mathbf{D}_{\mathrm{g}}$ commute $\forall\, \mathrm{g} \in G$. The same argument as in the previous section shows that $\mathbf{D}_{\mathrm{g}}\mathbf{A}$ are also eigenvectors of $\mathbf{M}$ spanning the same subspace H as the eigenvectors $\mathbf{A}$.

If $\mathbf{D}_{\mathrm{g}}$ is irreducible, H has no proper subspace invariant under G and the $\mathbf{D}_{\mathrm{g}}\mathbf{A}$ form a basis of H. In other words, $\forall\, \psi \in H$:

$$\mathbf{M}\,\psi = \sum_{\mathrm{g}} a_{\mathrm{g}}\, \mathbf{M}\, \mathbf{D}_{\mathrm{g}}\, \mathbf{A} \;=\; \lambda \sum_{\mathrm{g}} a_{\mathrm{g}}\, \mathbf{D}_{\mathrm{g}}\, \mathbf{A} \;=\; \lambda\, \psi$$

so that *all* transformed vectors in H are eigenvectors of $\mathbf{M}$, with the *same* eigenvalue λ. This can happen only if $\mathbf{M} = \lambda\,\mathbf{I}$, and there comes **Schur's First Lemma**:

The only complex matrix $\mathbf{M}$ that commutes with all the matrices of a given *irreducible* representation $\mathbf{D}_{\mathrm{g}}$ is a multiple of the identity matrix.

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. This provides one handy test for reducibility.

From this Lemma follows an immediate consequence for Abelian groups, where any matrix $\mathbf{D}_{\mathrm{g}}$ in a given representation commutes with the matrices for *all* other group elements in this representation. Assuming a $(n > 1)$-dim irreducible representation, the Lemma requires that $\mathbf{D}_{\mathrm{g}} = \lambda\,\mathbf{I}$, $\forall\,\mathrm{g} \in G$. But the $n \times n$ identity matrix, which is diagonal, cannot be irreducible if it represents *all* group elements, contradicting our assumption. We conclude that *all irreducible representations of an Abelian group are one-dimensional*.

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 can be proved (pp. BF615–617) by assuming unitary representations. This makes for no loss of generality for finite or compact Lie groups, since these $\big($eg. $O(n)\big)$ have finite-dimensional representations.

2.5.3 An orthogonality relation for the matrix elements of irreducible representations (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 $\mathbf{D}^{(\alpha)}$ and $\mathbf{D}^{(\beta)}$ are *inequivalent* irreducible representations:

$$\sum_{\mathrm{g}}^{N_{\mathrm{G}}} \big(D^{(\alpha)}_{\mathrm{g}}\big)^{i}{}_{k}\,\big(D^{(\beta)}_{\mathrm{g}^{-1}}\big)^{l}{}_{j} \;=\; \frac{N_{\mathrm{G}}}{n_\alpha}\,\delta^{i}{}_{j}\,\delta_{k}{}^{l}\,\delta_{\alpha\beta} \tag{2.6}$$

where N_{G} is the order of the group and n_α is the dimension of $\mathbf{D}^{(\alpha)}$. The sum is not matrix multiplication! Each term is the product of some ik entry of $D^{(\alpha)}_{\mathrm{g}}$ and lj entry of $D^{(\beta)}_{\mathrm{g}^{-1}}$, with ik and lj the same for each term.

In the usual case of unitary representations, this relation simplifies to:

$$\sum_{\mathrm{g}}^{N_{\mathrm{G}}} \big(D^{(\alpha)}_{\mathrm{g}}\big)^{i}{}_{k}\,\big(D^{(\beta)\,*}_{\mathrm{g}}\big)_{j}{}^{l} \;=\; \frac{N_{\mathrm{G}}}{n_\alpha}\,\delta^{i}{}_{j}\,\delta^{l}{}_{k}\,\delta_{\alpha\beta} \tag{2.7}$$

These relations set powerful constraints on the matrix elements of representations.

Eq. (2.6) is so important that it deserves a proof. Fortunately, this proof is not too hard. Construct a matrix:

$$\mathbf{M} \;=\; \sum_{\mathrm{g}}^{N_{\mathrm{G}}} \mathbf{D}^{(\alpha)}_{\mathrm{g}}\,\mathbf{X}\,\big[\mathbf{D}^{(\beta)}_{\mathrm{g}}\big]^{-1} \tag{2.8}$$

where $\mathbf{D}^{(\alpha)}$ and $\mathbf{D}^{(\beta)}$ are m-dim and n-dim inequivalent irreducible matrix representations of G, and X is any arbitrary operator represented by a $m \times n$ matrix $\mathbf{X}$. Then, for some $\mathrm{g}' \in G$,

$$\mathbf{D}^{(\alpha)}_{\mathrm{g}'}\,\mathbf{M}\,\big[\mathbf{D}^{(\beta)}_{\mathrm{g}'}\big]^{-1} \;=\; \sum_{\mathrm{g}}^{N_{\mathrm{G}}} \mathbf{D}^{(\alpha)}_{\mathrm{g}'\mathrm{g}}\,\mathbf{X}\,\big[\mathbf{D}^{(\beta)}_{\mathrm{g}'\mathrm{g}}\big]^{-1}$$

The sum on the right-hand side is just a different rearrangement of the sum that defines $\mathbf{M}$, so that:

$$\mathbf{M} \;=\; \mathbf{D}^{(\alpha)}_{\mathrm{g}'}\,\mathbf{M}\,\big[\mathbf{D}^{(\beta)}_{\mathrm{g}'}\big]^{-1}$$

Thus, $\mathbf{D}_{\mathrm{g}}^{(\alpha)}\,\mathbf{M} = \mathbf{M}\,\mathbf{D}_{\mathrm{g}}^{(\beta)}\;\forall\,\mathrm{g}\in G$, and $\mathbf{M}$ meets the condition for Schur's Second Lemma. Now let us choose $\mathbf{X}$ to be a matrix whose only non-zero element, 1, is its $(kl)^{\text{th}}$ entry. We can write this formally as: $(X_k^l)^m{}_n = \delta^m{}_k\,\delta_n{}^l$. Inserting gives:

$$(M_k^l)^i{}_j \;=\; \sum_{\mathrm{g}}^{N_{\mathrm{G}}} \left(\mathbf{D}_{\mathrm{g}}^{(\alpha)}\right)^i{}_m\,(X_k^l)^m{}_n\,\left(\mathbf{D}_{\mathrm{g}^{-1}}^{(\beta)}\right)^n{}_j \;=\; \sum_{\mathrm{g}}^{N_{\mathrm{G}}} \left(\mathbf{D}_{\mathrm{g}}^{(\alpha)}\right)^i{}_k\,\left(\mathbf{D}_{\mathrm{g}^{-1}}^{(\beta)}\right)^l{}_j$$

When $\alpha\neq\beta$ schur's Second Lemma requires that $\mathbf{M}_k^l = 0$. When $\alpha=\beta$, Schur's First Lemma requires that $\mathbf{M}_k^l = \lambda_k{}^l\,\mathbf{I}$, leading to:

$$(M_k^l)^i{}_j \;=\; \sum_{\mathrm{g}}^{N_{\mathrm{G}}} \left(\mathbf{D}_{\mathrm{g}}^{(\alpha)}\right)^i{}_k\,\left(\mathbf{D}_{\mathrm{g}^{-1}}^{(\alpha)}\right)^l{}_j \;=\; \lambda^l{}_k\,\delta^i{}_j$$

Setting $i=j$ and interchanging the $\mathbf{D}$ factors to get a matrix product, there comes:

$$\sum_{\mathrm{g}}^{N_{\mathrm{G}}} \left(\mathbf{D}_{\mathrm{g}^{-1}}^{(\alpha)}\right)^l{}_j\,\left(\mathbf{D}_{\mathrm{g}}^{(\alpha)}\right)^j{}_k \;=\; \sum_{\mathrm{g}}^{N_{\mathrm{G}}} \left(\mathbf{D}_{\mathrm{g}^{-1}\mathrm{g}}^{(\alpha)}\right)^l{}_k \;=\; N_{\mathrm{G}}\,\delta^l{}_k \;=\; \lambda^l{}_k\,n_\alpha$$

Thus, we find: $\lambda^l{}_k = N_{\mathrm{G}}\delta^l{}_k/n_\alpha$. Combining the results for $\alpha=\beta$ and $\alpha\neq\beta$ gives eq, (2.6) or (2.7).

This means that the matrix elements $\sqrt{\frac{n_\alpha}{N_{\mathrm{G}}}}\,(D_{\mathrm{g}}^{(\alpha)})^i{}_j$ of a unitary irreducible representation must be orthonormal functions of the group elements g; they are the components of a N_{G}-dim vector orthogonal to the similarly built vectors for any other irreducible representation. Therefore, these vectors are linearly independent. Also, they form a complete set.

EXERCISE: Show that for a finite group the sum over all elements g of the matrix elements $(D_{\mathrm{g}})^i{}_j$ (i and j fixed) of any *irreducible* representation other than the identity 1-dim representation (eg., (N) for S_N) is zero. This property can provide a useful check.

2.5.4 Characters of a representation (BF10.7); first orthogonality relation for characters

It turns out that a surprising large amount of information about representation matrices is encoded in their trace. Very often this trace can be found without knowing the full matrix.

Definition 2.32. The **character** of a representation $\mathbf{D}_{\mathrm{g}}$ of a group G is defined as a map from G to $\mathbb{C}$:

$$\chi_{\mathrm{g}} \;=\; \mathbf{Tr}\,\mathbf{D}_{\mathrm{g}}$$

Characters of reducible representations are **compound**; those of irreducible representations are called **simple**. Language alert: Mathematicians speak of the "character" of a representation as the *set* of traces of the matrices in the representation.

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 $\mathbf{D}_{\mathrm{g}}$, or χ. Since matrices for equivalent representations have the same character, any statement about characters is basis-independent!

Now set $k=i$ and $l=j$ in eq. (2.7):

$$\sum_{\mathrm{g}}^{N_{\mathrm{G}}} \left(D_{\mathrm{g}}^{(\alpha)}\right)^i{}_i\,\left(D_{\mathrm{g}}^{(\beta)\,*}\right)^j{}_j \;=\; \frac{N_{\mathrm{G}}}{n_\alpha}\,\delta^i{}_j\,\delta_i{}^j\,\delta_{\alpha\beta} \;=\; \frac{N_{\mathrm{G}}}{n_\alpha}\,\delta^i{}_i\,\delta_{\alpha\beta}$$

where repeated indices are summed over. Since $\delta^i{}_i = n_\alpha$, this can be rewritten as:

$$\sum_{\mathrm{g}}^{N_{\mathrm{G}}} \chi_{\mathrm{g}}^{(\alpha)}\,\chi_{\mathrm{g}}^{*(\beta)} \;=\; N_{\mathrm{G}}\,\delta_{\alpha\beta} \tag{2.9}$$

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

This provides our first orthogonality relation between the characters of irreducible representations. It can be viewed as an inner product on the space of functions of the $N_{\rm G}$-dim " character vectors".

Some of the terms in this sum will be identical since they correspond to group elements in the same class. So we can collect all terms belonging to the same class, which we label with k, and instead sum over the classes:

$$\sum_{k=1}^{N_{\rm c}} n_k \, \chi_k^{(\alpha)} \, \chi_k^{*(\beta)} \;=\; N_{\rm G} \, \delta_{\alpha\beta} \tag{2.10}$$

with n_k the number of elements in class k and $N_{\rm c}$ the number of classes in the group. This looks for all the world like an orthogonality relation between two vectors, $\sqrt{n_k/N_{\rm G}}\,\chi^{(\alpha)}$ and $\sqrt{n_k/N_{\rm G}}\,\chi^{(\beta)}$, each of dimension $N_{\rm c}$.

For a given irreducible representation, eq. (2.10) becomes:

$$\sum_{k=1}^{N_{\rm c}} n_k \left|\chi_k^{(\alpha)}\right|^2 \;=\; N_{\rm G} \tag{2.11}$$

This is a necessary and sufficient condition for the representation to be irreducible!

> **Example 2.10.** Take for instance the 3×3 representation of S_3 found in section 2.4.3. The identity, with trace 3, is in its own class, the three transpositions are in another class with trace 1, and the two cyclic permutations) have trace 0. Eq. (2.11) gives: $n_1\chi_1^2+n_2\chi_2^2+n_3\chi_3^2 = 1\,(3)^2 + 3\,(1)^2 + 2\,(0)^2 = 12$. Since this is not equal to 6, the number of elements in S_3, the representation must be reducible.

According to eq. (2.10), the " character vectors" of the $N_{\rm r}$ different irreducible representations are orthogonal. There are $N_{\rm r}$ such orthogonal vectors, and their number may not exceed the dimensionality of the space, $N_{\rm c}$, so that $N_{\rm r} \leq N_{\rm c}$. We will need this result a little later.

2.5.5 Multiplicity of irreducible representations and a sum rule for their dimension

Now consider the decomposition of a fully reducible representation into a direct sum of irreducible ones, given in eq. (2.3). Taking its trace yields an equation for the compound character $\chi_{\rm g}$: $\chi_{\rm g} = a_\alpha \, \chi_{\rm g}^{(\alpha)}$, where the sum runs over the $N_{\rm r}$ irreducible representations of the group. The compound character is seen to be a linear combination of simple characters with positive coefficients equal to the multiplicity of each irreducible representation.

Multiplying this relation by $\chi^{\star\,(\beta)}({\rm g})$ and summing over group elements, we find from eq. (2.9):

$$\sum_{\rm g}^{N_{\rm G}} \chi_{\rm g} \, \chi_{\rm g}^{\star\,(\beta)} \;=\; a_\alpha \sum_{\rm g}^{N_{\rm G}} \chi_{\rm g}^{(\alpha)} \, \chi_{\rm g}^{*\,(\beta)} \;=\; a_\alpha \, N_{\rm G} \, \delta_{\alpha\beta} \;=\; a_\beta \, N_{\rm G}$$

Thus, the multiplicity of each irreducible representation in the decomposition of a reducible representation is:

$$a_\alpha \;=\; \frac{1}{N_{\rm G}} \sum_{\rm g}^{N_{\rm G}} \chi_{\rm g} \, \chi_{\rm g}^{*\,(\alpha)} \;=\; \frac{1}{N_{\rm G}} \sum_{k}^{N_{\rm c}} n_k \, \chi_k \, \chi_k^{*\,(\alpha)} \tag{2.12}$$

Also, we can exploit the regular representation to obtain other *general* results for irreducible representations.

As we have seen in section 2.4.4, the matrix elements 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, the characters all vanish except for $\chi(e) = N_{\rm G}$.

Now, with n_α the dimension of the $\alpha^{\rm th}$ irreducible representation and ${\rm g} = e$, eq. (2.12) gives:

$$a_\alpha \;=\; \frac{1}{N_{\rm G}} \, \chi_e \, \chi_e^{*\,(\alpha)} \;=\; \chi_e^{*\,(\alpha)} \;=\; n_\alpha$$

Only $\chi(e)$ can contribute to the sum since $\chi_{\rm g} = 0$ in the regular representation when ${\rm g} \neq e$.

Therefore, *the multiplicity of an irreducible representation in the decomposition of the regular representation is its dimension, and it is never zero*. All the irreducible representations of a group must appear in the decomposition of its regular representation.

Next, taking the trace of the Kronecker decomposition (2.3) for the identity element in the regular representation yields: $N_{\rm G} = \sum_\alpha a_\alpha\, n_\alpha$. Combining those results, there comes an important **sum rule**:

$$N_{\rm G} = \sum_\alpha n_\alpha^2 \tag{2.13}$$

This powerful constraint tells us that $n_\alpha \le \sqrt{N_{\rm G}}$ so that any representation of dimension larger than $\sqrt{N_{\rm G}}$ must be reducible. When $N_{\rm G} = 2$ or 3, all irreducible representations are one-dimensional. When $N_{\rm G} = 4$, we can have only four inequivalent 1-d irreducible representations; $n_\alpha = 2$ is ruled out because there would be no identity 1-d representation. When $N_{\rm G} = 5$, eq. (2.13) does allow the identity representation plus one 2-d irreducible representation; but we know that this group, Z_5, is Abelian, and so admits only five inequivalent 1-d irreducible representations. For $N_{\rm G} = 6$, six 1-d, or two 1-d plus one 2-d irreducible representations, are allowed.

2.5.6 Another orthogonality relation

Here is another orthogonality relation, whose more complicated proof has been consigned to Appendix E at the end of this chapter for those who may be interested:

$$\sum_{\alpha=1}^{N_{\rm r}} \frac{n_k}{N_{\rm G}}\, \chi_k^{(\alpha)}\, \chi_{k'}^{*\,(\alpha)} = \delta_{k'k} \tag{2.14}$$

Thus, the characters in a given class k can be considered as components of $N_{\rm c}$ vectors forming a basis of a space whose dimension is $N_{\rm r}$, the number of irreducible representations, and which, according to eq. (2.14), are orthogonal. But in a $N_{\rm r}$-dimensional space there cannot be more than $N_{\rm r}$ orthogonal vectors, so $N_{\rm c} \le N_{\rm r}$.

In section 2.5.4, however, we had argued that $N_{\rm r} \le N_{\rm c}$. These results together lead to the important statement:

The number of inequivalent irreducible representations of a group is equal to the number of classes: $N_{\rm r} = N_{\rm c}$.

Now it can be shown (see Appendix F) that the direct 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 identity). 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 an 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 a given class in a pair of 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 the characters of the classes in a group's irreducible representations. Each row contains the characters of all classes in a representation, and each column the characters of a class in all representations.

The first row corresponds to the identity 1-dim irreducible representation, (N); all its entries must be 1. 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 in each representation) for the row.

If we are dealing with S_N, there is another 1-dim irreducible **antisymmetric** representation (called the **sign representation** by mathematicians), (1^N), conjugate to (N), whose 1×1 matrices, and therefore characters, are the parities ± 1 of its classes. We choose to place this representation at the bottom of the table.

What about the other entries? 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 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 representations, any character sitting in the same column as a -1 in the last row must be zero.

We have shown in Example 2.8 that the defining representation of S_N reduces to (N) and $(N\text{-}1\ 1)$. The characters of this $(N\text{-}1\ 1)$ 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. Since these compound characters are the sum of the characters of $(N\text{-}1\ 1)$ and (N), we find that the characters of each class labelled by α in the $(N\text{-}1\ 1)$ irreducible representation are just $\alpha - 1$.

When S_N has an N-dim irreducible representation, the permutations in the (N) class shuffle *all* N objects,. The N-dimensional matrices representing them must have diagonal entries 0, resulting in a character that is 0.

Next, we let eq. (2.10) and (2.14) provide constraints on the remaining unknowns:

- The first says that *complete* rows in the table (each for a different representations) are orthogonal, 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 orthogonal.

Now, if β refers to the identity representation, then, for any irrep α *other* than the identity, eq. (2.10) becomes:

$$\sum_{k=1}^{N_c} n_k \, \chi_k^{(\alpha)} \; = \; 0 \tag{2.15}$$

When invoking the orthogonality constraints to find characters, it is best to apply the linear ones first. 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. (2.15) and (2.14), and there still remain unknowns, one can try to spot reasonably simple quadratic relations from eq. (2.10) as well as using the normalisation of rows and columns.

Two last but important remarks: the characters of any 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 (2.15), as well as the other orthogonality rules, are automatically satisfied. Collecting yields:

	(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 2.11. 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 $-l \leq m \leq l$, 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, as well as of the Hamiltonian, transform under a rotation by some angle α. Using eq. (2.5), 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. $SO(3)$ will be discussed in chapter 3.

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 the z-axis. It will be sufficient to calculate the trace of the matrix representing rotations around that axis. To find this matrix, notice that $[\mathcal{R}_\alpha Y_{lm}](\theta,\phi) = \mathrm{e}^{\mathrm{i}m\alpha}Y_{lm}(\theta,\phi) = Y_{lm}(\theta,\phi+\alpha)$ because the dependence of the spherical harmonics on ϕ is $\mathrm{e}^{\mathrm{i}m\phi}$. Therefore, $\mathbf{D}^{(l)}(\alpha) = \mathrm{diag}(\mathrm{e}^{-\mathrm{i}l\alpha}, \mathrm{e}^{-\mathrm{i}(l+1)\alpha}, \ldots, \mathrm{e}^{\mathrm{i}l\alpha})$, and its character is not hard to compute:

$$\chi^{(l)}(\alpha\neq 0) \;=\; \sum_{m=-l}^{l}(\mathrm{e}^{\mathrm{i}\alpha})^m \;=\; \mathrm{e}^{-\mathrm{i}l\alpha}\sum_{n=0}^{2l}(\mathrm{e}^{\mathrm{i}\alpha})^n \;=\; \mathrm{e}^{-\mathrm{i}l\alpha}\left(\frac{1-\mathrm{e}^{\mathrm{i}(2l+1)\alpha}}{1-\mathrm{e}^{\mathrm{i}\alpha}}\right) \;=\; \frac{\sin\big[(l+1/2)\alpha\big]}{\sin(\alpha/2)} \tag{2.16}$$

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 fully rotationally-invariant, but still retains invariance under rotations by a *restricted*, finite set of angles, which we collectively denote by β. This would happen, for instance, if we embed our spherically-symmetric atom in a crystal lattice. Suppose this restricted set of rotations 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 (2.13) that applies to the finite subgroup.

The set of Y_{lm} transform as: $\mathcal{R}_\beta Y_{lm} = Y_{lm'}\big(D^{(l)}\big)^{m'}{}_{m}(\beta)$, with summation over repeated indices implied. If the induced representation $\mathbf{D}$ of the restricted-symmetry subgroup is reducible, there exists a matrix $\mathbf{S}$ independent of β which transforms all its matrices into block-diagonal matrices $\mathbf{D'} = \mathbf{S}\,\mathbf{D}\,\mathbf{S}^{-1}$, something which was impossible when there was no restriction on the angles.

But we do not have to know $\mathbf{S}$ to extract useful information. Indeed, because $\mathbf{D}$ and $\mathbf{D'}$ have the same trace, 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 restricted-symmetry 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 dimension of any irreducible representation of the subgroup. Then all the corresponding $\mathbf{D}^{(l)}(\beta)$ will be reducible to a direct sum of the irreducible representations of the subgroup, given by eq. (2.3), with the multiplicity of each irrep calculable from eq. (2.12).

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 (vertices).
- 6 rotations by π around the six axes intersecting the centre of two diagonally opposed edges;

[1] See, eg: `http://demonstrations.wolfram.com/RotatingCubesAboutAxesOfSymmetry3DRotationIsNonAbelian/`.

With the identity rotation, these add up to 24 elements forming a subgroup of $SO(3)$ isomorphic to S_4. The correspondence between rotations and permutations is obtained by considering each rotation as a shuffling of the four pairs of diagonally opposed vertices (or the four principal diagonals through the centre), each pair labelled 1 to 4. The five classes of S_4 are (1^4) (e), (4) (6 rotations by $\pm\pi/2$), (2^2) (3 rotations by π) , $(3\,1)$ (8 rotations by $\pm 2\pi/3$), and $(2\,1^2)$ (6 rotations by π). The character table S_4 is:

	(1^4)	$(2\,1^2)$	(2^2)	$(3\,1)$	(4)
n_k	1	6	3	8	6
(4)	1	1	1	1	1
(1^4)	1	−1	1	1	−1
(2^2)	2	0	2	−1	0
$(3\,1)$	3	1	−1	0	−1
$(2\,1^2)$	3	−1	−1	0	1

Here, the irreps of S_4 (or of the group of rotational symmetries of the cube) are ordered by increasing dimension instead of their mixed-symmetry structure. With eq. (2.16), 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:

	(1^4)	$(2\,1^2)$	(2^2)	$(3\,1)$	(4)
$\mathbf{D}^{(l=1)}$	3	1	−1	0	−1

	(1^4)	$(2\,1^2)$	(2^2)	$(3\,1)$	(4)
$\mathbf{D}^{(l=2)}$	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 representation $(3\,1)$ 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. (2.12), we calculate the following multiplicity for each irrep of S_4 that can appear in the decomposition of $\mathbf{D}^{(l=2)}(\beta)$: $a_{(4)} = a_{(1^4)} = a_{(2\,1^2)} = 0$, and $a_{(2^2)} = a_{(3\,1)} = 1$. Then we have the S_4 decomposition:

$$\mathbf{D}^{(l=2)}(\beta) \;=\; \mathbf{D}_{(2^2)}(\beta) \oplus \mathbf{D}_{(3\,1)}(\beta)$$

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.

Another example illustrating how symmetry-breaking can remove degeneracy, at least in part, can be found in Appendix G.

Appendices

E Proof of the Second orthogonality Relation for Characters

Our first orthogonality relation for characters, eq. (2.7). says that the set $\{\sqrt{n_\alpha/n_{\rm G}}\big(D^{(\alpha)}_{\rm g}\big)^i{}_j\}$, with i and$j$ fixed, of an irreducible representation α, can be viewed as the $N_{\rm G}$ components of a vector orthogonal to any other such vector corresponding to other matrix elements, whether or not they belong to the same representation as that of the first vector. There are $N_{\rm G}$ such vectors and they form a complete set with completeness relation expressed as:

$$\sum_{\alpha}^{N_{\rm r}}\sum_{i,\,j}^{n_\alpha}\frac{n_\alpha}{N_{\rm G}}\big(D^{(\alpha)}_{\rm g}\big)^i{}_j\,\big(D^{(\alpha)\,*}_{\rm g'}\big)^i{}_j \;=\; \delta_{\rm g'g} \tag{E.1}$$

where N_r is the number of irreducible representations. Again, the left-hand side is not matrix multiplication.

Take the equation for each element g of some class k, and sum over all elements of the class; we can also do this with g$'$ 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.

$\sum_{\rm g}^{n_k}\big(D^{(\alpha)}_{\rm g}\big)^i{}_j$ in the now quadruple sum on the left-hand side is an element of the matrix $\mathbf{M}$ constructed by summing all the matrices $\mathbf{D}_{\rm g}$ in the representation that correspond to elements g of class k: $\mathbf{M}\;=\;\sum_{\rm g}^{n_k}\mathbf{D}_{\rm g}$.

If g$'$ is some arbitrary element of G, we have:

$$\mathbf{D}_{\rm g'}\,\mathbf{M}\,\mathbf{D}_{\rm g'^{-1}}\;=\;\sum_{\rm g}\mathbf{D}_{\rm g'}\,\mathbf{D}_{\rm g}\,\mathbf{D}_{\rm g'^{-1}}\;=\;\sum_{\rm g}\mathbf{D}_{\rm g'\,g\,g'^{-1}}\;=\;\mathbf{M}$$

where the last equality results from the fact that, since $\rm g'\,g\,g'^{-1}$ is in class k, the left-hand side of the last equality is just a rearrangement of the sum defining $\mathbf{M}$. Thus, $\mathbf{D}_{\rm g}\,\mathbf{M}\;=\;\mathbf{M}\,\mathbf{D}_{\rm g}\;\forall\,{\rm g}\in G$, and, from Schur's First Lemma, $\mathbf{M}=\lambda\,\mathbf{I}$, with λ a constant that depends on the class and on the n-dim representation. Then $\mathbf{Tr}\;\mathbf{M}=n\lambda$.

Because all matrices in a class for a given representation must have the same trace, we have: $\mathbf{Tr}\;\mathbf{M}=n_k\chi_k$. Since that trace is also $n\lambda$, we find:

$$\mathbf{M}\;=\;\frac{n_k}{n}\,\chi\,\mathbf{I} \tag{E.2}$$

With two of its four sums replaced by matrix elements of $\mathbf{M}$, the completeness relation (E.1) now reads:

$$\sum_{\alpha}^{N_{\rm r}}\sum_{i,\,j}^{n_\alpha}\frac{n_\alpha}{N_{\rm G}}\big(M^{(\alpha)}_{k}\big)^i{}_j\,\big(M^{(\alpha)\,*}_{k'}\big)^i{}_j \;=\; n_k\delta_{k'k}$$

Inserting $\mathbf{M}^{(\alpha)}=(n_k/n_\alpha)\chi^{(\alpha)}\,\mathbf{I}$ and carrying out the sums over i and j gives another orthogonality relation:

$$\sum_{\alpha=1}^{N_{\rm r}}\frac{n_k}{N_{\rm G}}\,\chi^{(\alpha)}_{k}\,\chi^{*\,(\alpha)}_{k'} \;=\; \delta_{k'k} \tag{E.3}$$

F Direct Product of Representations

Let $D^i{}_j$ and $D'^\alpha{}_\beta$ be the entries of respresentation matrices $\mathbf{D}$, of rank m, and $\mathbf{D}'$, of rank n, for some goup element g. We define the **direct-product** representation constructed out of $\mathbf{D}$ and $\mathbf{D}'$ as the $(mn)\times(mn)$ matrix $\mathbf{M} = \mathbf{D}\otimes\mathbf{D}'$ with entries $M^{i\alpha}{}_{j\beta} \coloneqq D^i{}_j D'^\alpha{}_\beta$, where the superscript $i\alpha$ and subscript $j\beta$, each running from 1 to (mn) just indicate which matrix elements to multiply.

$\mathbf{M}$ can be seen to be a representation because the ordinary matrix product $\mathbf{M}_{\mathrm{g}_1}\mathbf{M}_{\mathrm{g}_2}$ has entries:

$$\begin{aligned}(M_{\mathrm{g}_1})^{i\alpha}{}_{j\beta}(M_{\mathrm{g}_2})^{j\beta}{}_{k\gamma} &= (D_{\mathrm{g}_1})^i{}_j(D'_{\mathrm{g}_1})^\alpha{}_\beta(D_{\mathrm{g}_2})^j{}_k(D'_{\mathrm{g}_2})^\beta{}_\gamma = (D_{\mathrm{g}_1})^i{}_j(D_{\mathrm{g}_2})^j{}_k(D'_{\mathrm{g}_1})^\alpha{}_\beta(D_{\mathrm{g}_2})'^\beta{}_\gamma \\ &= (D_{\mathrm{g}_1\mathrm{g}_2})^i{}_k(D'_{\mathrm{g}_1\mathrm{g}_2})^\alpha{}_\gamma = (M_{\mathrm{g}_1\mathrm{g}_2})^{i\alpha}{}_{k\gamma}\end{aligned}$$

Therefore, the matrix product of direct-products for two group elements is the direct product of the matrix product for the representations of the sane two elements. Also, the character of the direct-product for a group element, $M^{i\alpha}{}_{i\alpha} = D^i{}_i D'^\alpha{}_\alpha$, is obviously the product of the characters of the representations in the product.

When $\mathbf{D}$ is irreducible, its direct product with a 1-dim (and thus irreducible) representation, results in an irreducible representation of the same dimension as $\mathbf{D}$, which we call its conjugate representation.

G A Second Example of Symmetry-Breaking Lifting a Degeneracy

Take six identical bodies arranged on a circle 60° apart, each subject to an identical external linear restoring force giving rise to *small* oscillations about their equilibrium position and tangent to the circle. Let $\mathbf{X}$ be their displacement vector, with components: $x_1, \ldots, x_6$ their displacements away from their respective equilibrium position. This vector is a solution of Newton's 2^{nd} Law for the system, $\ddot{\mathbf{X}} = -\mathbf{M}^{-1}\mathbf{K}\mathbf{X}$, where: $\mathbf{M} = \mathrm{diag}(m, m, \ldots, m)$, and $\mathbf{K} = \mathrm{diag}(k, \ldots, k)$, 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, all bodies oscillate at the same frequency $\omega_0 = \sqrt{k/m}$. The space of solutions is spanned by the six eigenvectors belonging to the *same* eigenvalue ω_0. This is a six-fold degeneracy.

Now we couple the bodies by an identical weak interaction to their nearest neighbours $\pm 60^\circ$ away; similarly, couple them 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.

The eigenvectors of S now satisfy $\mathsf{S}\mathbf{A} = \lambda\mathbf{A}$. But since $\mathsf{S}^6 = \mathsf{I}$, 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 $A^j_{(m)} = \lambda^{j-1}_{(m)} = \mathrm{e}^{\mathrm{i}m(j-1)\pi/3}$.

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.

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 CHAPTER III — LIE GROUPS

3.1 Definitions

In this chapter 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 now we put in an extra requirement: that each group element g_p, or $\mathrm{g}(P)$, be in correspondence with a point P in some manifold, with the index "p" taken as a set of continuous, real variables. We 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 invertible mappings $P = \phi_i(P_1, P_2)$. Let $\mathrm{g}(P_1) \star \mathrm{g}(P_2) = \mathrm{g}(P)$ be the group product of an infinite group G. If the maps ϕ_i and their 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 said to be **path-connected** if any pair of points on its manifold is connected by a continuous path.

A Lie group is **compact** when the volume of its manifold is finite.

The subset of all elements in a Lie group whose corresponding points in M^n are connected by a continuous path to the identity is a subgroup. Thus, a Lie group that is not path-connected must contain a path-connected subgroup.

Example 3.1. An infinite line with coordinate $-\infty < x < \infty$ ($x \in \mathbb{R}$) 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: $z = \phi(x, y) = x + y$.

Example 3.2. Restrict $\theta = x \in \mathbb{R}$ with $0 \leq \theta < 2\pi$, and define group elements $\mathrm{g}(\theta) = \mathrm{e}^{\mathrm{i}\theta}$ with 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, whose points are each parametrised by *real* angle θ, and $\phi(\theta_1, \theta_2) = \theta_1 + \theta_2$. Its elements are complex, but the group is real! It is Abelian, and connected.

Example 3.3. Real invertible 2×2 matrices form a group whose elements can be written as $\mathbf{g}(x) = \begin{pmatrix} x_1 & x_2 \\ x_3 & x_4 \end{pmatrix}$. Constraining the matrices to be **unimodular** (to have determinant 1) lowers the number of parameters by 1. The group product is:

$$\begin{pmatrix} x_1 & x_2 \\ x_3 & \frac{1+x_2\,x_3}{x_1} \end{pmatrix} \begin{pmatrix} y_1 & y_2 \\ y_3 & \frac{1+y_2\,y_3}{y_1} \end{pmatrix} = \begin{pmatrix} z_1 & z_2 \\ z_3 & \frac{1+z_2\,z_3}{z_1} \end{pmatrix}$$

Compute the set of three functions $z_i = \phi_i(x_1, x_2, x_3, y_1, y_2, y_3)$ consistent with this group product:

$$z_1 \;=\; x_1\,y_1 \;+\; x_2\,y_3 \qquad z_2 \;=\; x_1\,y_2 \;+\; x_2\,\frac{1 + y_2\,y_3}{y_1} \qquad z_3 \;=\; x_3\,y_1 \;+\; y_3\,\frac{1 + x_2\,x_3}{x_1}$$

In this parametrisation, the mappings ϕ_i are all differentiable only off the $x_1 = 0$ and $y_1 = 0$ planes. Whatever the associate manifold is—see later—it cannot be covered with just this coordinate patch.

The inverse mapping corresponding to $\mathrm{g}^{-1}(x)$ can be read off the inverse matrix $\mathbf{g}^{-1}$.

Example 3.4. If we demand instead that invertible complex 2×2 matrices be not only unimodular, but also *unitary*, we obtain the group $SU(2)$, and the treatment is simpler. Introduce the parametrisation:

$$\begin{pmatrix} z & w \\ -w^* & z^* \end{pmatrix} = \begin{pmatrix} a^0 + \mathrm{i}\, a^3 & a^2 + \mathrm{i}\, a^1 \\ -(a^2 - \mathrm{i}\, a^1) & a^0 - \mathrm{i}\, a^3 \end{pmatrix}$$

The condition $|z|^2+|w|^2 = a_0^2+a_1^2+a_2^2+a_3^2 = 1$ emsures 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 real coordinates $(a_0,\, a_1,\, a_2,\, a_3)$ the components of a 4-dim unit vector; this is a *real* 3-dim Lie group. This time, there is a smooth *invertible* map (**diffeomorphism**) between $SU(2)$ and S^3.

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.

Following standard usage, we introduce the diagonal matrix I_p^q with p entries $+1$, and q entries -1, where $p+q=n$. In this notation I_n is the n-dim identity matrix, and also the Cartesian metric in Euclidean n-dim space.

One useful 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 identify interesting subgroups by constraining its elements. Here, we focus on two types of constraints: bilinear and unimodular.

3.2.1 Bilinear or quadratic constraints: the metric (or distance)-preserving groups

Definition 3.3. Unitary transformations T of a complex matrix $\mathsf{M}\in GL(n,\mathbb{C})$ are defined by:

$$\mathsf{M} \;\mapsto\; \mathsf{T}\,\mathsf{M}\,\mathsf{T}^\dagger$$

where the subgroup of matrices T leaves the Cartesian n-dim metric $\mathsf{M}=\mathsf{I}_n$ invariant: $\mathsf{T}\mathsf{I}_n\mathsf{T}^\dagger = \mathsf{T}\mathsf{T}^\dagger = \mathsf{I}_n$. Thus, as expected, $\mathsf{T}^{-1}=\mathsf{T}^\dagger$, and we call that subgroup $U(n)\subset GL(n,\mathbb{C})$: We say that $U(n)$ is **unitary**. Example 3.2 referred to $U(1)$.

Definition 3.4. Orthogonal transformations T of a real matrix $\mathsf{M}\in GL(n,\mathbb{R})$ are defined by:

$$\mathsf{M} \;\mapsto\; \mathsf{T}\,\mathsf{M}\,\mathsf{T}^{\mathrm{T}}$$

(T^{T} is the transpose of T), such that T leaves I_n invariant: $\mathsf{T}\,\mathsf{I}_n\mathsf{T}^{\mathrm{T}} = \mathsf{T}\mathsf{T}^{\mathrm{T}} = \mathsf{I}_n$, that is, $\mathsf{T}^{-1}=\mathsf{T}^{\mathrm{T}}$. It is not hard to show (EXERCISE) that such **orthogonal** matrices form a group, called $O(n)$.

Be aware that n in $O(n)$ or $U(n)$ refers to the dimension of the *matrices*, not that the group which is the number of coordinates on its manifold! $O(n)$ 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 path-connected; neither is $U(n)$.

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 I_p^q, defined by the transformations:

$$\mathsf{T}\,\mathsf{I}_p^q\,\mathsf{T}^{\mathrm{T}} = \mathsf{I}_p^q \qquad O(p,q) \tag{3.1}$$

$$\mathsf{T}\,\mathsf{I}_p^q\,\mathsf{T}^{\dagger} = \mathsf{I}_p^q \qquad U(p,q) \tag{3.2}$$

A famous example is $O(3,1)$, aka the **full Lorentz group**, that leaves the mostly positive Minkowski 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: $\mathsf{T}\,\mathsf{I}_3^1\,\mathsf{T}^{\mathrm{T}} = \mathsf{I}_3^1$, 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.

3.2.2 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 all unimodular matrices.

Example 3.3 actually referred to $SL(2,\mathbb{R})$ with the constraint $x_1x_4 - x_2x_3 = 1$, 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—and metrics!

The intersection of special linear and metric-preserving groups can form important subgroups: eg., $SO(n) = O(n) \cap SL(n,\mathbb{R})$ and $SU(n) = U(n) \cap SL(n,\mathbb{C})$. These groups are compact. Example 3.4 was about $SU(2)$.

We stated earlier that $O(n)$ and $U(n)$ are not path-connected, but we know that they must have path-connected subgroups, ie. groups with elements connected to the identity by a continuous path. These are $SO(n)$ and $SU(n)$.

3.2.3 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 matrix Lie groups.

1. **Translations** Let f be an analytic function acting on $\mathbb{R}^n$. The left action on f of the operator $\mathcal{T}_\mathbf{a}$ associated with $T_\mathbf{a}\mathbf{x} = \mathbf{x} + \mathbf{a}$ is:

$$[\mathcal{T}_\mathbf{a}\, f](\mathbf{x}) \;=\; f(T_\mathbf{a}^{-1}\mathbf{x}) \;=\; f(\mathbf{x} - \mathbf{a}) \qquad\qquad \mathbf{a} \in \mathbb{R}^n$$

   Except for the identity ($\mathbf{a} = 0$), such transformations leave no $\mathbf{x}$ invariant and are called inhomogeneous.

2. **Rotations**

   Parametrise 3-dim rotations in the $z = 0$ plane of a vector $\mathbf{x} \in \mathbb{R}^3$ by R_α, with $R_\alpha\phi = \phi + \alpha$, with $[\mathcal{R}_\alpha\, f](\phi) = f(\phi - \alpha)$, and $-\pi < \phi \leq \pi$. In terms of the left action on the components of $\mathbf{x}$: $\mathbf{x}' = R_\alpha\,\mathbf{x}$ (ie. $\mathbf{x}'$ obtained by rotating $\mathbf{x}$ by $+\alpha$ in the $z = 0$ plane), the matrix associated with R_α is:

$$\begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad \text{Then}: [\mathcal{R}_\alpha\, f](\mathbf{x}) = f(R_\alpha^{-1}\mathbf{x}) \;=\; f(x\cos\alpha + y\sin\alpha,\; -x\sin\alpha + y\cos\alpha,\; z)\,.$$

   In terms of example 2.6, this corresponds to rotating the basis by $-\alpha$.

   Arbitrary rotations in 3-dim space: can be written as $[\mathcal{R}\, f](\mathbf{x}) = f(R^{-1}\mathbf{x})$, where $\mathcal{R}$ can be factored 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 will not write down a matrix representation. It leaves lengths invariant, is unimodular, and thus is an element of $SO(3)$. Rotations in a plane are isomorphic to the one-parameter group $SO(2)$ whose group manifold is the circle, S^1 (see example 3.2). Is the group manifold of $SO(3)$ the sphere S^2? How many parameters does S^2 have?

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 (think of zooming in or out). 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 $\mathbb{R}^4$ coordinates ct and $\mathbf{x}$:

$$ct' \;=\; \gamma(ct - \boldsymbol{\beta}\cdot\mathbf{x}) \qquad\qquad \mathbf{x}' \;=\; \mathbf{x} + \frac{\gamma - 1}{\beta^2}\,(\boldsymbol{\beta}\cdot\mathbf{x})\,\boldsymbol{\beta} \;-\; \gamma\boldsymbol{\beta}(ct)$$

   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. 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 $O(3,1)$ (or $O(1,3)$ in a mostly negative metric), the full Lorentz group. The transformation is a homogeneous one, which in the 4-vector formalism is written: $\mathbf{a}' = \Lambda\mathbf{a}$, where $\mathbf{a}$ is a *any* 4-vector (not necessarily position).

We can extend the full Lorentz transformations to include space-time translations $\mathbf{t}$:

$$\mathbf{x}' \;=\; \Lambda\mathbf{x} \,+\, \mathbf{t}$$

Whereas the homogeneous transformations left the norm of a 4-vector invariant, these inhomogeneous transformations leave invariant only the norm of the *difference* between two 4-vectors.

If we call $\mathbf{\Lambda}$ the full Lorentz transformation matrix, we can construct the matrix for these transformations by adding to $\mathbf{\Lambda}$ a fifth row and column whose last element is a 1 that does not do anything, that is:

$$\begin{pmatrix}\mathbf{x}'\\ 1\end{pmatrix} \;=\; \begin{pmatrix}\mathbf{\Lambda} & \mathbf{t}\\ 0 & 1\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ 1\end{pmatrix}$$

These matrices form the 10-parameter inhomogeneous Lorentz group, or Poincaré group, $ISO(3,1)$. Incidentally, setting $\mathbf{\Lambda} = \mathbf{I}$ gives a matrix realisation of the 4-dim translation group.

These examples illustrate the isomorphism between physical transformations and matrix Lie groups. We can then identify, say, a rotation with a $SO(3)$ matrix, and even call $SO(3)$ the rotation group.

3.2.4 Differential-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 few transformations as *differential operators*, a far from gratuitous exercise.

1. **Translations**

   We can first look just at smooth functions $f(x)$, $x \in \mathbb{R}$. Then the result of a translation $T_a x = x + a$, $a \in \mathbb{R}$, on f, with $a \ll x$, can be Taylor-expanded about x:

   $$[\mathcal{T}_a f](x) \;=\; f(T_a^{-1}x) \;=\; f(x-a) \;=\; \big[(1 \,-\, a\,\mathrm{d}_x \,+\, \ldots)\, f\big](x) \;=\; \big[\exp(-a\,\mathrm{d}_x)\, f\big](x)$$

   In $\mathbb{R}^3$ this generalises to:

   $$[\mathcal{T}_{\mathbf{a}} f](\mathbf{x}) \;=\; f(T_{\mathbf{a}}^{-1}\mathbf{x}) \;=\; f(\mathbf{x}-\mathbf{a}) \;=\; \left[\sum_{n=0}^{\infty}\frac{1}{n!}\,(-a^i\,\partial_i)^n\, f\right](\mathbf{x}) \;=\; \big[\exp(-a^i\,\partial_i)\, f\big](\mathbf{x}) \tag{3.3}$$

   The operators $-\partial_i$ are called the **infinitesimal generators** of translations. Quantum mechanics uses instead the Hermitian momentum operator $\mathbf{p} = -\mathrm{i}\hbar\boldsymbol{\partial}$ and writes the translation operator as: $\mathcal{T}_{\mathbf{a}} = \mathrm{e}^{-\mathrm{i}\mathbf{a}\cdot\mathbf{p}/\hbar}$.

   We note that the *Cartesian* infinitesimal generators $-\partial_i$ (or p_i) commute amongst themselves.

2. **Rotations**

   For rotations $R_\alpha\phi = \phi + \alpha$ in the $(z=0)$ plane by a small angle α:

   $$[\mathcal{R}_\alpha f](\phi) \;=\; f(R_\alpha^{-1}\phi) \;=\; f(\phi-\alpha) \;=\; \big[(1 \,-\, \alpha\,\mathrm{d}_\phi \,+\, \ldots)\, f\big](\phi) \;=\; \big[\exp(-\alpha\,\mathrm{d}_\phi)\, f\big](\phi)$$

   As we have seen in the last section, in $\mathbb{R}^3$ with Cartesian coordinates, this gives for the left action of a rotation $R_\alpha\mathbf{x} = (x\cos\alpha - y\sin\alpha,\; x\sin\alpha + y\cos\alpha,\; z)$: $f(R_\alpha^{-1}\mathbf{x}) \;=\; f(x\cos\alpha + y\sin\alpha, -x\sin\alpha + y\cos\alpha, z)$. If we Taylor-expand the right-hand side we obtain:

   $$[\mathcal{R}_\alpha\, f](\mathbf{x}) \;=\; \big[\big(1 \,+\, \alpha\,(y\,\partial_x \,-\, x\,\partial_y) \,+\, \ldots\big)\, f\big](\mathbf{x}) \;=\; \big[\exp(\alpha\, M_z)\, f\big](\mathbf{x}) \tag{3.4}$$

   where $M_z = y\,\partial_x - x\,\partial_y$. Similarly for rotations about the x and y axes, the general rotation operator is: $\mathcal{R}_{\alpha,\beta,\gamma} = \exp(\alpha M_x)\exp(\beta M_y)\exp(\gamma M_z)$, where:

   $$M_x = z\,\partial_y \,-\, y\,\partial_z,\;\; M_y = x\,\partial_z \,-\, z\,\partial_x,\;\; M_z = y\,\partial_x \,-\, x\,\partial_y \tag{3.5}$$

or: $M_i = -\epsilon_{ijk} x^j \partial^k = \frac{1}{2}\epsilon_{ijk} J^{jk}$, where $J_{jk} \coloneqq x_{[k}\partial_{j]}$, with x_j and ∂_k defining a 2-dim plane of rotation. The pseudovector operator $\mathbf{M}$ is the Hodge dual of the more natural simple 2-form operator $\mathbf{J}$. In quantum mechanics, it is redefined as $\mathbf{L} = \mathrm{i}\hbar\mathbf{M}$ and interpreted as the (Hermitian) angular-momentum operator.

These infinitesimal generators do not commute. Indeed: $[M_i,\, M_j] = \epsilon_{ij}{}^k\, M_k\,, \quad$ or $\quad [L_i,\, L_j] = \mathrm{i}\hbar\, \epsilon_{ij}{}^k\, L_k$.

(Note: we could have written — some do! — defined $\mathbf{M}$ as the negative of the above. The cost, however, would be an extra minus sign in the commutation relations.)

3. **Dilations or scale transformations**

In a n-dim space with Cartesian coordinates x^μ, a **scale transformation** is: $T_\kappa x^\mu = (1+\kappa)x^\mu$. In the limit of small κ, $T_\kappa^{-1} x^\mu \approx (1-\kappa)x^\mu$. Again we Taylor-expand a function $f(T_\kappa^{-1}x^\mu)$ in the small parameter κ:

$$f(T_\kappa^{-1}x^\mu) \;=\; \big[(1 \,-\, \kappa\, x^\mu\partial_\mu \,+\, \cdots)\, f\big](x^\mu) \;=\; \big[\exp(-\kappa\, x^\mu\partial_\mu)\, f\big](x^\mu) \tag{3.6}$$

We identify $D = -x^\mu\partial_\mu = -\mathbf{x}\cdot\boldsymbol{\partial}$ (compare with $\mathbf{M} = -\mathbf{x}\times\boldsymbol{\partial}$) as the infinitesimal generator of dilations.

We can now find the infinitesimal generators of an arbitrary group of transformations with m parameters a^i near 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 one nearby that can be described by the same coordinate chart.

Let the transformations act (left action!) on a space (aka carrier space) of differentiable functions f on M^n:

$$[\mathcal{T}_{\mathbf{a}}\, f](\mathbf{x}) \;=\; f(T_{\mathbf{a}}^{-1}\mathbf{x})$$

Focus on $\mathcal{T}_{\mathbf{a}}\, f$, and take f as a function of the *parameters* a^i. As before, Taylor-expand the right-hand side to first order around the identity parametrised by $\mathbf{a} = 0$:

$$[\mathcal{T}_{\mathbf{a}}\, f](\mathbf{x}) \;=\; \Big[\big(1 \,+\, a^i\, \partial_{a^i}(T_{\mathbf{a}}^{-1}x)^j\Big|_{\mathbf{a}=0}\partial_j \,+\, \ldots\big)\, f\Big]\,(\mathbf{x})$$

where i runs over the number of parameters, ie. the dimension of the group, and j from 1 to the dimension of the space on which the functions f act.

Definition 3.5. The operators:

$$X_i \;=\; \partial_{a^i}(T_{\mathbf{a}}^{-1}x)^j\Big|_{\mathbf{a}=0}\partial_j \tag{3.7}$$

are called **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) to ensure hermiticity.

For example, rotations in the $z = 0$ plane in Cartesian $\mathbb{R}^3$ involve one parameter (angle) $a^1 = \alpha$, and only x and y derivatives can occur since z does not depend on α. Then the second term in the square bracket of eq. (3.4) is recovered.

3.2.5 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:

$$\mathbf{M}_{\mathbf{a}} \;=\; \mathbf{I}_n \,+\, a^i\,\mathbf{X}_i \qquad\qquad \mathbf{X}_i \;=\; \partial_{a^i}\mathbf{M}_{\mathbf{a}}\Big|_{\mathbf{a}=0} \tag{3.8}$$

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. Let $\mathsf{M}_\theta \in SO(2)$: $\left(\begin{smallmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{smallmatrix}\right)$, for $0 \le \theta < 2\pi$, that effects rotations in a plane.

Taylor-expand to first order: $\mathsf{M}_\theta \approx \begin{pmatrix}1 & -\theta\\ \theta & 1\end{pmatrix} = \mathsf{I}_2 + \theta\,\mathsf{X}$

Then the infinitesimal generator of $SO(2)$ is:

$$\mathsf{X} = \partial_\theta \mathsf{M}_\theta\big|_{\theta=0} = \begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}$$

a matrix fully consistent with the constraints on $SO(n)$ generators as we shall discover in section 3.3.4. 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. Similarly, using the parametrisation of example 3.4, we see that an element of $SU(2)$ may be written as $a^0\mathsf{I}_2 + a^i\mathsf{X}_i$, where $\mathsf{X}_i = \mathrm{i}\,\boldsymbol{\sigma}_i$ are the generators of $SU(2)$, with $\boldsymbol{\sigma}_i$ the Pauli matrices.

When the group matrices are not known we must resort to other methods to be discussed a little later.

An infinitesimal generator is an operator that effects an infinitesimal transformation away from the identity. We would like to reconstruct a *finite* transformation out of a succession of infinitesimal transformations that use only the generators, ie., the *first-order* contribution in the expansion of a transformation written as an exponential:

$$\mathrm{e}^A = \sum_{n=0}^{\infty} \frac{A^n}{n!} \tag{3.9}$$

which holds for any (well-behaved) linear operator A. Here, $A = a^i\mathsf{X}_i$. Higher powers of A do contribute, but they end up being either multiples of the identity matrix or of X. The series breaks into series of powers of the parameters, which can usually be identfied as functions. This is what happens for the one generator of $SO(2)$: we have $\mathsf{X}^{2n} = \mathsf{I}$, and so $\mathsf{X}^{2n+1} = \mathsf{X}$. The exponential series breaks into a cosine series in θ multiplying the identity, and a sine series multiplying X.

This **exponential map**, then, is the tool that reconstructs finite transformations from infinitesimal ones. But it must be handled with some care as we shall discover.

Note that the inverse of a group element e^{A} is $\mathrm{e}^{-\mathsf{A}}$, and that a generator matrix $\mathsf{A} = a^i\mathsf{X}_i$ need not be invertible.

3.3 Lie Algebras

3.3.1 Linearisation of a Lie group product

To understand the importance of infinitesimal generators, notice that linear combinations of group elements may not be in the group: for instance, linear combinations of $SO(2)$ matrices are not elements of $SO(2)$. 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. Then an arbitrary element in the space can always be written as b^iX_i, with b_i the group parameters.

That vector space arises from linearising the product of a Lie group G around the identity. The result can considerably simplify the study of the group. First, write $(\mathrm{g},\ \mathrm{g}') \in G$ in the neighbourhood of the identity as $\mathrm{g} \approx e + a\epsilon X$ and $\mathrm{g}' \approx e + b\epsilon Y$, where ϵ is an arbitrarily small positive real number and a and b are real, but arbitrary as well. X and Y are infinitesimal generators of g and g′, respectively. Expand $\mathrm{g}\,\mathrm{g}' \in G$ to first order in ϵ:

$$\mathrm{g}\,\mathrm{g}' \approx e + \epsilon\,(aX + bY) + \ldots$$

Manifestly, $aX + bY$ is a generator for the product $\mathrm{g}\,\mathrm{g}'$, and the generators indeed form a linear vector space.

Now expand the product $h = \mathrm{g}\,\mathrm{g}'\,(\mathrm{g}'\,\mathrm{g})^{-1} \in G$ to first non-vanishing order, this time writing $\mathrm{g} \approx e + \epsilon_1 X + \epsilon_1^2 X^2/2$, and $\mathrm{g}' \approx e + \epsilon_2 Y + \epsilon_2^2 Y^2/2$, with $(\epsilon_1,\, \epsilon_2)$ arbitrarily small:

$$\begin{aligned}\mathrm{g}\,\mathrm{g}'\,(\mathrm{g}'\,\mathrm{g})^{-1} \;&\approx\; \big(e + \epsilon_1\, X + \frac{1}{2}\epsilon_1^2\, X^2\big)\big(e + \epsilon_2\, Y + \frac{1}{2}\epsilon_2^2\, Y^2\big)\big(e \,-\, \epsilon_1\, X \,+\, \frac{1}{2}\epsilon_1^2\, X^2\big)\big(e \,-\, \epsilon_2\, Y \,+\, \frac{1}{2}\epsilon_2^2\, Y^2\big) \,+\, \ldots \\ &\approx\; e \,+\, \epsilon_1\, \epsilon_2\, (XY \,-\, YX) \,+\, \ldots\end{aligned}$$

All other contributions of order ϵ_i^2 and $\epsilon_1\epsilon_2$ cancel out. We define $[X,\, Y] \coloneqq XY - YX$, the commutator of the generators X and Y. As the generator for $\mathrm{g}\,\mathrm{g}'\,(\mathrm{g}'\,\mathrm{g})^{-1}$, $[X,\, Y]$ must be an element of the same vector space as X and Y. When $h = e$, $\mathrm{g}\,\mathrm{g}' = \mathrm{g}'\,\mathrm{g}$, and the commutator of the generators vanishes. Thus, mathematicians often refer to $\mathrm{g}\,\mathrm{g}'\,(\mathrm{g}'\,\mathrm{g})^{-1}$ as the "commutator" for the *group* product, but we shall reserve the term for $[X,\, Y]$.

It is straightforward to show that the **Jacobi identity** holds, just by expanding it:

$$\big[X,\, [Y,\, Z]\big] \,+\, \big[Y,\, [Z,\, X]\big] \,+\, \big[Z,\, [X,\, Y]\big] \;=\; 0 \tag{3.10}$$

3.3.2 Definition of a Lie algebra

Now we are ready for an important definition that collects and generalises our findings:

> **Definition 3.6.** An **algebra** $\mathfrak{g}$ is a vector space over a field $(\mathbb{R},\, \mathbb{C})$, equipped with, on top of the generic addition operation, a *bilinear product* $\mathfrak{g} \times \mathfrak{g} \longmapsto \mathfrak{g}$. Algebras may be associative or commutative.
>
> When the product is the **Lie bracket** $[\cdot,\, \cdot]$, which:
>
> - is **antisymmetric**: $[X,\, Y] \;=\; -\,[Y,\, X]$;
> - satisfies the Jacobi identity: $\big[X,\, [Y,\, Z]\big] \,+\, \big[Y,\, [Z,\, X]\big] \,+\, \big[Z,\, [X,\, Y]\big] \;=\; 0$.
>
> we say that $\mathfrak{g}$ is a **Lie algebra**. In physics, the Lie bracket is the commutator $XY - YX$. Many, because they always deal with the algebra, not the group, use G to denote $\mathfrak{g}$, which can be confusing. Because $\big[X,\, [Y,\, Z]\big] - \big[[X,\, Y],\, Z\big] \neq 0$, Lie algebras are associative only if $\big[X,\, [Y,\, Z]\big] = 0 \;\; \forall X, Y, Z \in \mathfrak{g}$. Mathematicians like to think of $\mathfrak{g}$ as the tangent space of a group at its identity.

It is crucial to keep in mind that the action of a Lie-algebra element X on another one, Y, is *not* XY, but their *commutator*! The closure property of a Lie group in effect translates into the existence of its algebra.

The algebra $\pm\mathrm{i}\,\mathfrak{g}$ is said to be **essentially real**. Example: the linear and orbital angular-momentum operators of quantum mechanics related to real infinitesimal generators.

Sometimes, however, it proves very convenient to construct a **complex extension** of a real or essentially real algebra, by allowing basis redefinitions that involve complex coefficients. For instance, we might wish to construct $J_\pm = J_x \pm \mathrm{i} J_y$. This provides more flexibility in constructing useful bases.

The dimension n of a Lie algebra is the number of parameters of its associated group.

> **Definition 3.7.** A **subalgebra** of an algebra $\mathfrak{g}$ is just a subspace that closes under commutation. A subalgebra $\mathfrak{g}_{\mathrm{sub}}$ is **invariant** if $[\mathfrak{g}_{\mathrm{sub}},\, \mathfrak{g}] \subseteq \mathfrak{g}_{\mathrm{sub}}$, ie. if, $\forall\, X \in \mathfrak{g}_{\mathrm{sub}}$ 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 that commutes with *all* elements of the algebra. The centre of a commutative (Abelian) algebra is itself. $\mathfrak{z}$ is always an Abelian invariant subalgebra.

3.3.3 Structure constants of a Lie algebra

The commutators of its n infinitesimal generators X_i 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:

$$[X_i,\, X_j] \;=\; C_{ij}{}^k\, X_k \tag{3.11}$$

The coefficients ${C_{ij}}^k$ are called the **structure constants** of the Lie algebra, whose **structure** they are said to specify. In fact, with some rarely relevant caveats, 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, up to a redefinition (eg. rescaling) of their generators.

The structure constants inherit the antisymmetry of the commutators: ${C_{ji}}^k = -{C_{ij}}^k$. When the structure constants all vanish, ie., when $[X,\, Y] = 0 \;\forall\, (X,\, Y) \in \mathfrak{g}$, we say that the algebra is **Abelian**.

The Jacobi identity on elements of an algebra induces (EXERCISE) a relation between the structure constants:

$$ {C_{ij}}^l\, {C_{kl}}^m \,+\, {C_{jk}}^l\, {C_{il}}^m \,+\, {C_{ki}}^l\, {C_{jl}}^m \;=\; 0 \qquad \Longleftrightarrow \qquad {C_{[ij}}^l\, {C_{k]l}}^m \;=\; 0 \tag{3.12} $$

Defining a matrix $(\mathbf{D}_i)_j{}^k = -{C_{ij}}^k$, we find (EXERCISE) that $\mathbf{D}$ satisfies the commutation relation (3.11). If we can take the group's representations to be unitary, as for compact groups such as $SU(n)$ and $SO(n)$, the corresponding representations of the algebra are anti-Hermitian and we immediately find (EXERCISE), since they must satisfy the commutation relations, that the structure constants are real.

The structure constants for the essentially real algebra $\pm\mathrm{i}\,\mathfrak{g}$ are just (exercise) $\pm\mathrm{i}\,{C_{ij}}^k$. Quite 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 parametric form for the matrix realisation of a Lie group. All we know are the constraints on the group elements. This is sufficient to find the Lie algebra and then reconstruct the group matrix.

First, linearise the constraints. At the beginning of section 3.2 we found that for Cartesian 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), $\mathbf{M}\,\mathbf{I}^q_p\,\mathbf{M}^\dagger \;=\; \mathbf{I}^q_p$, with $p+q=n$.

Linearising for the compact groups, we get: $(\mathbf{I}_n + \epsilon\,\mathbf{A})(\mathbf{I}_n + \epsilon\mathbf{A})^\dagger \approx \mathbf{I}_n + \epsilon\,(\mathbf{A}^\dagger + \mathbf{A}) = \mathbf{I}_n$ 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, with $n(n-1)/2$ independent parameters. Thus, the $\mathfrak{o}(n)$ algebra 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 3.2.4), then $\mathbf{M} = I_n + \mathrm{i}\epsilon\mathbf{A}$, and the $\mathbf{A}$ matrices are Hermitian: $\mathbf{A}^\dagger = \mathbf{A}$.

If the group is an indefinite orthogonal group, which 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: $\mathbf{S}^\dagger \;=-\mathbf{S}$, $\mathbf{V}^\dagger = -\mathbf{V}$, $\mathbf{T}^\dagger = \mathbf{U}$. 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. The number of parameters of the indefinite orthogonal group $O(p,q)$ is then $p(p-1)/2 + q(q-1)/2 + pq = n(n-1)/2$, the same as for the compact orthogonal group $O(n)$.

There only remains to notice that the non-zero elements of the infinitesimal generator matrices can only be ±1 (over $\mathbb{R}$) and also $\pm\,\mathrm{i}$ (over $\mathbb{C}$) because of the linearisation.

Another important constraint can be imposed on a group matrix $\mathbf{M}$: $\det\mathbf{M} = 1$, which defines $SL(n, \mathbb{R}\,\mathrm{or}\,\mathbf{C})$. Since the determinant of a product of matrices is equal to the product of the determinants of the matrices, and because—when a matrix $\mathbf{A}$ is diagonalisable— there exists a similarity transformation $\mathbf{S}\mathbf{A}\mathbf{S}^{-1}$ which takes $\mathbf{A}$ to $\mathbf{A}' = \mathrm{diag}(\lambda_1,\,\ldots,\lambda_i,\,\ldots)$, we conclude that $\det\mathbf{A}$ is equal to the product of the eigenvalues of $\mathbf{A}$.

Also, if $\mathbf{M} = \mathrm{e}^{\mathbf{A}}$, it transforms as:

$$ \mathbf{S}\,\mathrm{e}^{\mathbf{A}}\,\mathbf{S}^{-1} \;=\; \mathbf{S}\,\mathbf{I}\mathbf{S}^{-1} + \mathbf{S}\,\mathbf{A}\,\mathbf{S}^{-1} + \frac{1}{2!}\,\mathbf{S}\,\mathbf{A}\,\mathbf{S}^{-1}\,\mathbf{S}\,\mathbf{A}\,\mathbf{S}^{-1} + \ldots \;=\; \mathbf{I} + \mathbf{A}' + \frac{1}{2!}\,(\mathbf{A}')^2 + \ldots \;=\; \mathrm{e}^{\mathbf{A}'} $$

where $\mathrm{e}^{\mathbf{A}'}$ is a diagonal matrix with e^{λ_i} as entries. In other words, the eigenvalues of $\mathrm{e}^{\mathbf{A}}$ are just e^{λ_i}. Then:

$$\det \mathrm{e}^{\mathbf{A}} \;=\; \prod_i \mathrm{e}^{\lambda_i} \;=\; \exp \sum_i \lambda_i \;=\; \mathrm{e}^{\mathrm{Tr}\,\mathbf{A}'}$$

But $\mathrm{Tr}\,\mathbf{A}' = \mathrm{Tr}(\mathbf{S}\mathbf{A}\mathbf{S}^{-1}) = \mathrm{Tr}\,\mathbf{A}$. We obtain via this elegant (but limited to diagonalisable matrices!) derivation an important *basis-independent* relation, valid for *any* square matrix:

$$\det \mathbf{M} \;=\; \det\left(\mathrm{e}^{\mathbf{A}}\right) \;=\; \mathrm{e}^{\mathrm{Tr}\,\mathbf{A}} \tag{3.13}$$

This extends to $\det\left(\mathrm{e}^{\mathbf{A}}\mathrm{e}^{\mathbf{B}}\cdots\right) = \mathrm{e}^{\mathrm{Tr}\,(\mathbf{A}+\mathbf{B}+\ldots)}$, and since all $SL(n,\,\mathbb{R})$ matrices can be written as a prodiuct $\mathrm{e}^{\mathbf{A}}\mathrm{e}^{\mathbf{B}}$ (to be shown later), we immediately deduce that all matrices in the algebra $\mathfrak{sl}(n,\,\mathbb{R})$ must have vanishing trace, including those in $\mathfrak{su}(n)$ and $\mathfrak{so}(n)$. Thus, it can be said that $\mathfrak{sl}(n,\,\mathbb{R})$ is the set of all traceless matrices of rank n.

Since antisymmetric real matrices are 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 path-connected. Spatial inversions cannot be linearised; one cannot invert axes by a "small" amount! So the infinitesimal generators of $O(3)$ are those of its path-connected $SO(3)$ subgroup of rotations.

We quote 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:

$$\mathrm{e}^{\mathbf{A}}\mathrm{e}^{\mathbf{B}} = \mathrm{e}^{\mathbf{C}} \qquad \mathbf{C} \;=\; \mathbf{A}+\mathbf{B}+\frac{1}{2}[\mathbf{A},\mathbf{B}]+\frac{1}{12}\Big(\big[\mathbf{A},[\mathbf{A},\mathbf{B}]\big]+\big[[\mathbf{A},\mathbf{B}],\,\mathbf{B}\big]\Big)+\ldots \tag{3.14}$$

Example 3.6. To find the matrix realisation of the generators of $SO(3)$, which live in a three-parameter algebra, 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 $\mathbf{x}$, with only vectors along the axis unchanged. 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 an element of the $\mathfrak{so}(3)$ algebra. How does this compare to the operator algebra as laid out in eq. (3.5)? By identifying $\alpha = \theta_z$, 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 indeed the $\mathfrak{so}(3)$-algebra matrix. A rotation by a *finite* angle θ of a vector around axis $\hat{\mathbf{n}}$ can be written as: $R(\boldsymbol{\theta}) = \mathrm{e}^{\theta^k\mathbf{M}_k}$, with generators:

$$\mathbf{M}_x \;=\; \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad \mathbf{M}_y \;=\; \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \qquad \mathbf{M}_z \;=\; \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

The operator and matrix algebras have the same commutator structure, $[M_i,\,M_j] = \epsilon_{ij}{}^k M_k$, establishing their isomorphism. When rotating a function, the M generators of the operator realisation would be used—see eq. (3.5).

Often, $SO(3)$ generators are written as $J_{ij} = \epsilon_{ijk}M^k$, which is arguably more natural. Since $(\mathbf{M}_i)_{jk} = -\epsilon_{ijk}$, the matrix elements are: $(J_{ij})^{lm} = -\epsilon_{ijk}\epsilon^{klm} = -(\delta_i{}^l\delta_j{}^m - \delta_i{}^m\delta_j{}^l)$. The labels

(ij), $i<j$ for J refer to the *plane* of rotation. To obtain their commutators, compute (EXERCISE): $\big[J_{mn},\,J_{pq}\big]^{i}{}_{j} = (J_{mn})^{i}{}_{k}(J_{pq})^{k}{}_{j} - (J_{pq})^{i}{}_{l}(J_{mn})^{l}{}_{j}$ and rearrange the eight resulting terms, yielding[†]:

$$[J_{mn},J^{pq}] \;=\; \delta_{m}{}^{p}\,J_{n}{}^{q} \;-\; \delta_{m}{}^{q}\,J_{n}{}^{p} \;-\; \delta_{n}{}^{p}\,J_{m}{}^{q} \;+\; \delta_{n}{}^{q}\,J_{m}{}^{p} \qquad 1\le m<n\le N,\; 1\le p<q\le N \tag{3.15}$$

Only one term on the right can contribute (EXERCISE), and the commutator vanishes unless one (and only one) number in the pair (mn) is equal to one (and only one) number in the pair (pq).

This result is important because it applies to rotations in dimensions $N>3$, for which a plane of rotation does not uniquely define an axis, as it does for $N=3$. But a rotation in a 2-dim plane in N-dim space is always about a well-defined point where all axes perpendicular to the plane meet.

Two other important and often useful results: scalar operators, ie., those that are invariant under 3-dim rotations, must commute with the $SO(3)$ generators (eg., the Hamiltonian for a spherically-symmetric potential). As for a vector operator $\mathbf{V}$, ie., one that transforms as a vector under rotations, it is shown in Appendix H that it satisfies $[M_i,V_j]=\epsilon_{ijk}V^k$, or $[L_i,V_j]=\mathrm{i}\,\epsilon_{ijk}V^k$.

Example 3.7. The 6-dimensional $\mathfrak{so}(4)$ Lie algebra is the set of all antisymmetric 4×4 real matrices, which can be parametrised in the following way:

$$\mathfrak{so}(4) \;=\; a^{i}M_i \,+\, b^{i}N_i \;=\; \begin{pmatrix} 0 & -a_3 & a_2 & -b_1 \\ a_3 & 0 & -a_1 & -b_2 \\ -a_2 & a_1 & 0 & -b_3 \\ b_1 & b_2 & b_3 & 0 \end{pmatrix}$$

It is now appropriate to use the $4(4-1)/2=6$ J_{ij} generators, introduced in example 3.6, that generate rotations in the (ij)-plane. With eq. (3.15) it is easy to compute the nine non-trivial $\mathfrak{so}(4)$ commutators, by taking $J_{i4}=N_i$ and $J_{ij}=\epsilon_{ijk}M^k$ ($1\le i,\,j<k\le3$). Alternatively, we could use the isomorphism with differential operators. With $\mathbb{R}^4$ coordinates $x,\;y,\;z,\;u$, there are six of these:

$$M_1 \;=\; z\,\partial_{y} \,-\, y\,\partial_{z}, \qquad M_2 \;=\; x\,\partial_{z} \,-\, z\,\partial_{x}, \qquad M_3 \;=\; y\,\partial_{x} \,-\, x\,\partial_{y}$$

$$N_1 \;=\; x\,\partial_{u} \,-\, u\,\partial_{x}, \qquad N_2 \;=\; y\,\partial_{u} \,-\, u\,\partial_{y}, \qquad N_3 \;=\; z\,\partial_{u} \,-\, u\,\partial_{z}$$

Whether with eq. (3.15) or the operator realisation, we obtain:

$$[M_i,\,M_j] \;=\; \epsilon_{ij}{}^{k}\,M_k, \qquad [M_i,\,N_j] \;=\; \epsilon_{ij}{}^{k}\,N_k, \qquad [N_i,\,N_j] \;=\; \epsilon_{ij}{}^{k}\,M_k \tag{3.16}$$

The generators can be decoupled by transforming to the basis: $Y_i=\dfrac{1}{2}\,(M_i+N_i)$, $Z_i=\dfrac{1}{2}\,(M_i-N_i)$, from which we immediately obtain the *decoupled* relations:

$$[Y_i,\,Y_j] \;=\; \epsilon_{ij}{}^{k}\,Y_k, \qquad [Y_i,\,Z_j] \;=\; 0, \qquad [Z_i,\,Z_j] \;=\; \epsilon_{ij}{}^{k}\,Z_k \tag{3.17}$$

By inspection, the Y_i and Z_i are generators of two separate $\mathfrak{su}(2)$ (or $\mathfrak{so}(3)$) algebras, and $\mathfrak{so}(4)=\mathfrak{su}(2)\oplus\mathfrak{su}(2)$. In terms of dimensions, $\mathbf{6}=\mathbf{3}\oplus\mathbf{3}$. At group level, we say that $SO(4)$ is *locally* isomorphic to the direct product $SU(2)\times SU(2)$; globally, however, $SU(2)\times SU(2)$ double-covers $SO(4)$ since pairs of elements of $SU(2)$ and pairs of their negatives map to the same $SO(4)$ rotation.

This makes eminent sense. Indeed, consider a transformation $SU(2)\times SU(2)$ acting on a matrix $C\in SU(2)$: $C'=A^{\dagger}\,C\,B\in SU(2)$, with $(A,\,B)\in SU(2)$. From example 3.4, both C and C' are associated with unit vectors in Euclidean $\mathbb{R}^4$. Thus, the $SU(2)\times SU(2)$ action is a rotation in $SO(4)$.

Then $\mathfrak{so}(4) \;=\; a^{i}\,Y_i \,+\, b^{i}\,Z_i$, and since $[Y_i,Z_j]=0$, an element of $SO(4)$ takes the form: $\mathrm{e}^{a^{i}Y_i}\,\mathrm{e}^{b^{i}Z_i}$.

[†]This relation has a short form: $[J_{mn},J^{pq}]=\delta_{[m}{}^{[p}\,J_{n]}{}^{q]}$, that is helpful as a mnemonic device. Just start from: $\delta_{m}{}^{p}\,J_{n}{}^{q}$, and generate the other three terms by atisymmetrising with respect to p and q, then m and n, and finally both pairs together.

Example 3.8. The $\mathfrak{so}(3,\,1)$ algebra of the group $SO(3,1)$ derived from the metric-preserving constraint is:

$$\mathfrak{so}(3,1) \;=\; \begin{pmatrix} 0 & \zeta_x & \zeta_y & \zeta_z \\ \zeta_x & 0 & -\theta_z & \theta_y \\ \zeta_y & \theta_z & 0 & -\theta_x \\ \zeta_z & -\theta_y & \theta_x & 0 \end{pmatrix} \;=\; \theta^\mu\,\mathbf{M}_\mu \,+\, \zeta^\nu\,\mathbf{K}_\nu \tag{3.18}$$

where the infinitesimal generators can be read off:

$$\begin{aligned} \mathbf{M}_x &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \qquad \mathbf{M}_y = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix} \qquad \mathbf{M}_z = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \\ \mathbf{K}_x &= \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \qquad \mathbf{K}_y = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \qquad \mathbf{K}_z = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \end{aligned} \tag{3.19}$$

One shows (EXERCISE) that the commutators of the infinitesimal generators are:

$$[M_i,\,M_j] \;=\; \epsilon_{ij}{}^k\,M_k \qquad [M_i,\,K_j] \;=\; \epsilon_{ij}{}^k\,K_k \qquad [K_i,\,K_j] \;=\; -\,\epsilon_{ij}{}^k\,M_k \tag{3.20}$$

Notice that although the three $\mathbf{M}$ rotation generators form a subalgebra, the $\mathbf{K}$ generators do not, because they do not close under commutation.

Although the number of generators is identical to $\mathfrak{so}(4)$, there is an important difference between these relations and the ones derived in example 3.7: the minus sign in the relation for the K, which can also be obtained by letting $N \to \mathrm{i}\,K$. Then 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 tells us that complexified $\mathfrak{so}(3,1)$ is isomorphic to $\mathfrak{su}(2) \oplus \mathrm{i}\,\mathfrak{su}(2)$.

As in example 3.6, by defining $J_{ij} = \epsilon_{ij}{}^k M_k$ and $J_{0i} = K_i$, $1 \le i \le 3$, one rewrites the commutator relations (3.20) as a relation valid for *any* $\mathfrak{so}(p,q)$ algebra in $N = p + q$ dimensions.

$$\big[J_{\mu\nu},\,J_{\alpha\beta}\big] \;=\; \eta_{\mu\alpha}J_{\nu\beta} \,+\, \eta_{\nu\beta}J_{\mu\alpha} \,-\, \eta_{\mu\beta}J_{\nu\alpha} \,-\, \eta_{\nu\alpha}J_{\mu\beta} \qquad 0 \le (\mu,\,\nu) \le N-1 \tag{3.21}$$

where $J_{\nu\mu} = -J_{\mu\nu}$ (subscripts label generators J, not their components!), and $\eta_{\mu\nu}$ is the Cartesian Minkowski metric: diag $(\mp 1, \pm 1, \ldots, \pm 1)$, depending on the metric sign convention.

One very important realisation of this algebra interprets θ_i as the three angles rotating around Cartesian axis $1 \le i \le 3$, and $\zeta_i = \hat{\beta}_i\,\tanh^{-1}\beta$ the rapidity parameters for pure Lorentz boosts along the x, y and z axes, written in terms of the relative velocity $\boldsymbol{\beta}$ betweeen two inertial frames. Then $\mathfrak{so}(3,\,1)$ is called the **Lorentz algebra** for Minkowski spacetime. The relation (3.20) can also be derived (EXERCISE) in the differential-operator realisation: $J_{\mu\nu} = x_\nu\partial_\mu - x_\mu\partial_\nu$.

Note that finite Lorentz boosts, $\mathrm{e}^{-\zeta^i K_i}$, do not form a group. These matrices are symmetric but a product of symmetric matrices is symmetric only if they commute, which is not the case for boosts.

Symmetries under rotations and Lorentz transformations, as well as under translations, are prime example of **global** symmettries, in the sense that the transformations have the same *form* at all points. **Local** symmetries involve transformations that can vary arbitrarily from one point to another, so are said to be point-dependent.

3.3.5 Hard-nosed questions about the exponential map — the fine print

Three theorems by Lie, which we have implicitly used, show that for any Lie group an algebra can be found, characterised by the structure constants. At best only the *path-connected* part of a group can be recovered from its algebra. We have relied on the exponential map to do this, but it is not always possible, at least with just one map.

Here is a counter-example (provided by Cartan). Take: $\mathsf{Z} = \left(\begin{smallmatrix} x_1 & x_2-x_3 \\ x_2+x_3 & -x_1 \end{smallmatrix}\right) \in \mathfrak{sl}(2,\mathbb{R})$, whose trace vanishes. Exponentiating gives (EXERCISE):

$$\mathrm{e}^{\mathsf{Z}} = \sum_n \frac{1}{n!}\mathsf{Z}^n = \begin{cases} \mathsf{I}_2 \cosh r + \mathsf{Z}\dfrac{\sinh r}{r} & r^2 > 0 \\ \mathsf{I}_2 + \mathsf{Z} & r^2 = 0 \\ \mathsf{I}_2 \cos r + \mathsf{Z}\dfrac{\sin r}{r} & r^2 < 0 \end{cases}$$

where $r^2 = x_1^2 + x_2^2 - x_3^2 = -\det \mathsf{Z}$, which makes the results basis-independent. The structure is reminiscent of the light-cone structure obtained by endowing the parameter space $\mathbb{R}^3$ with an indefinite metric invariant under $\mathrm{SO}(2,1)$. 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, r can grow without restriction and maps to non-compact elements of $\mathrm{SL}(2,\mathbb{R})$.

So far, so good. But a glance at the above expressions shows that $\mathrm{Tr}\,\mathrm{e}^{\mathsf{Z}} \geq -2$ always. Yet $\mathrm{SL}(2,\mathbb{R})$ has a large subset of elements with trace smaller than -2: matrices of the type $\left(\begin{smallmatrix} -\lambda & 0 \\ 0 & -1/\lambda \end{smallmatrix}\right)$ ($\lambda > 1$), for instance. These cannot be reached with the above exponential map.

Cartan argued that all the group elements could nevertheless be reached by writing:

$$\mathsf{Z} = \mathsf{Z}_a + \mathsf{Z}_b = \begin{pmatrix} x_1 & x_2 \\ x_2 & -x_1 \end{pmatrix} + \begin{pmatrix} 0 & -x_3 \\ x_3 & 0 \end{pmatrix}$$

and taking the product of the exponentials of Z_a and Z_b, which is not e^{Z} since $[\mathsf{Z}_a, \mathsf{Z}_b] \neq 0$. Then (EXERCISE):

$$\mathrm{e}^{\mathsf{Z}_a}\,\mathrm{e}^{\mathsf{Z}_b} = \begin{pmatrix} z+y & x \\ x & z-y \end{pmatrix}\begin{pmatrix} \cos x_3 & -\sin x_3 \\ \sin x_3 & \cos x_3 \end{pmatrix}$$

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.

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 Z_b subalgebra to $\mathrm{SO}(2) \subset \mathrm{SL}(2,\mathbb{R})$, whose manifold is S^1. We conclude that $\mathrm{SL}(2,\mathbb{R})$ is non-compact, and that its manifold is $\mathbb{R}^2 \times S^1$. Every point is path-connected to the origin ($x_1 = x_2 = 0$) of $\mathbb{R}^2$ and $x_3 = 0$ on S_1, so $\mathrm{SL}(2,\mathbb{R})$ is path-connected.

3.4 Representations of Lie Groups and Algebras

3.4.1 Representations of Lie Groups

Definition 3.8. As with 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 matrices $GL(\mathcal{V})$ acting on a space $\mathcal{V}$, 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, it may well happen that $\mathcal{H}$ is infinite-dimensional.

Of special interest are irreducible representations. They satisfy Schur's lemma: A unitary representation T_{g} is irreducible if, and only if, the only operator A on $\mathcal{H}$ such that: $A\,\mathsf{T}_{\mathrm{g}} = \mathsf{T}_{\mathrm{g}}\,A\ \forall \mathrm{g} \in G$ 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 is equivalent to an unitary representation. All unitary representations of a compact Lie group are finite-dimensional. Thus, so are all irreducible representations..
- 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.

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 a Hilbert space $\mathcal{H}$. We 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, $\forall\,(X,\ Y)\in\mathfrak{g}$: $\mathsf{T}_{[X,Y]}=\left[\mathsf{T}_X,\ \mathsf{T}_Y\right]$.

3.4.3 The regular (adjoint) representation and the classification of Lie algebras

We have already noted how eq. (3.12) for the structure constants could be written as the commutator of matrices which 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 R_Z (or ad_Z) such that $R_Z(X_i)=[Z,\ X_i]=X_j\big(R_Z\big)^j{}_i$, where the X_i are the *basis* generators of the algebra. $\big($Some authors use the definition $[Z,\ X_i]=\big(R_Z\big)_i{}^jX_j.\big)$

Clearly, the regular representation of a basis generator is just the structure constants: $[X_i,\ X_j]=\big(R_{X_i}\big)^k{}_j\,X_k=C_{ij}{}^k\,X_k$. Its dimension is that of the algebra, the number of generators (or parameters).

We confirm that R is a representation (EXERCISE, with the Jacobi identity): $[\mathsf{R}_{X_i},\ \mathsf{R}_{X_j}]X_k=\mathsf{R}_{[X_i,\,X_j]}X_k$.

Example 3.9. Take the defining, two-dimensional representation of the essentially real version of the $\mathfrak{su}(2)$ algebra with basis elements $S_i=\boldsymbol{\sigma}_i/2$, where $\boldsymbol{\sigma}_i$ are the three *Hermitian* Pauli matrices, and whose commutators are: $[S_i,\ S_j]=\mathrm{i}\,\epsilon_{ij}{}^kS_k$. Then $(\mathrm{ad}_{S_i})^k{}_j=\mathrm{i}\,\epsilon_{ij}{}^k$, and we have[†];

$$\mathrm{ad}_{S_1}=\begin{pmatrix}0&0&0\\0&0&-\mathrm{i}\\0&\mathrm{i}&0\end{pmatrix}\qquad \mathrm{ad}_{S_2}=\begin{pmatrix}0&0&\mathrm{i}\\0&0&0\\-\mathrm{i}&0&0\end{pmatrix}\qquad \mathrm{ad}_{S_3}=\begin{pmatrix}0&-\mathrm{i}&0\\\mathrm{i}&0&0\\0&0&0\end{pmatrix}$$

Then a generic element $Z=a^iS_i$ of $\mathfrak{su}(2)$ has the Hermitian regular representation:

$$\mathsf{R}_Z=\begin{pmatrix}0&-\mathrm{i}\,a_3&\mathrm{i}\,a_2\\\mathrm{i}\,a_3&0&-\mathrm{i}\,a_1\\-\mathrm{i}\,a_2&\mathrm{i}\,a_1&0\end{pmatrix}$$

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 3.11. 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.

[†]The cummutation relations for the adjoint representation are: $\left[\mathrm{ad}_{S_i},\,\mathrm{ad}_{S_j}\right]=\mathrm{i}\,\epsilon_{ij}{}^k\mathrm{ad}_{S_k}$. With our convention, $\big(\mathrm{ad}_{X_i}\big)^k{}_j=C_{ij}{}^k$, for the adjoint representation, the structure constants for adjoint and defining represeentations are always identical. With the other convention, $\big(\mathrm{ad}_{X_i}\big)_j{}^k=C_{ij}{}^k$, they would differ by a minus sign.

Definition 3.12. If an algebra has no non-trivial invariant subalgebra, its *regular* representation is irreducible (it leaves no *proper* subspace of its carrier space invariant), and the algebra is called **simple**.

Definition 3.13. A Lie algebra that contains no *Abelian, invariant* subalgebra is said to be **semisimple**, ie. it has zero centre (no non-zero element commutes with all other elements). A semisimple algebra is either simple or the sum of simple algebras (that may occur more than once in the sum). A semisimple algebra always has at least two complementary invariant subalgebras, and there is a basis in which all the generators of one commute with all the generators of the other(s), but not amongst themselves.

From these two definitions it follows that all simple algebras are semisimple since they are already in (single) block form. *Non*-simple semisimple algebras 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 example 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 3.10. 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 6-dim 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 semisimple, but not simple.

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 3.14. The **Cartan-Killing form** (CK-form) is a symmetric, bilinear form whose components are the inner product of all pairs of elements of a Lie algebra in their adjoint representation:

$$(Y,\, Z) \;\coloneqq\; \mathrm{Tr}\left(\mathbf{R}_Y\, \mathbf{R}_Z\right) \;=\; (\mathbf{R}_Y)^k{}_l\, (\mathbf{R}_Z)^l{}_k \tag{3.22}$$

The CK-form for basis generators X_i is easily calculated: $(X_i,\, X_j) = C_{il}{}^k\, C_{jk}{}^l$. If the algebra has n parameters, the CK-form has $n(n+1)/2$ components.

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 after writing: $\mathrm{g} = \mathrm{e}^{\epsilon Z}$, we obtain (EXERCISE):

$$\left([Z,\, X],\, Y\right) \;+\; \left(X,\, [Z,\, Y]\right) \;=\; 0 \tag{3.23}$$

This, of course, is consistent with the cyclicity of the trace.

[†]For a given i, j Y_i and Z_j in the decoupled basis of eq. (3.17) form an Abelian subalgebra, but it is not invariant.

Definition 3.15. A CK-form is **degenerate** (or **singular**) if there exists at least one element Z in the algebra for which $(Z,\, Y) = 0\ \forall\, Y \in \mathfrak{g}$; equivalently—and more usefully—if the matrix $(X_i,\, X_j)$ for the basis generators has a row and column entirely populated with zeros, which forces its determinant to vanish. Otherwise, the CK-form is **non-degenerate**.

Equivalently, a CK-form is non-degenerate if there exists a basis in which it is diagonal with all entries non-zero. 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 that basis. 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. 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 raise and lower indices. In particular, introduce $f_{\mu\nu\lambda} \coloneqq C_{\mu\nu}{}^{\rho} g_{\rho\lambda}$. Inserting $g_{\rho\lambda} = (X_\rho,\, X_\lambda)$, one can show (EXERCISE) with eq. (3.23) or, equivalently, the cyclicity of the trace, that $f_{\mu\nu\lambda}$ is antisymmetric.

Now, if an algebra has a non-zero centre $\mathfrak{z}$ (ie. an Abelian invariant subalgebra that commutes with all the elements of the algebra), its CK-form is degenerate because the adjoint representation of any element of $\mathfrak{z}$ vanishes trivially. The converse is also true, which leads to a more useful result, **Cartan's criterion** for semismplicity:

Definition 3.16. A Lie algebra is **semisimple** if, and only if, its CK-form is non-degenerate,

Example 3.11. $x^i\partial_j$ 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. Therefore, it is not semisimple.

Example 3.12. In example 3.9, we have already obtained the adjoint representation for the generators of $\mathfrak{su}(2)$ — and the one for $\mathfrak{so}(3)$ because the structure constants for the two algebras are now identical. With $\mathbf{S}$ in the adjoint representation, eq. (3.22) then gives:

$$(S_i,\, S_j) \;=\; \mathrm{Tr}\,(\mathbf{S}_i\mathbf{S}_j) \;=\; 2\,\delta_{ij}$$

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 $f_{ijk} = \mathrm{i}\epsilon_{ijk}$.

When the number of basis generators is large, calculating all the independent components of a Cartan metric can be tedious. But some useful information about the metric can still be extracted from $(\mathbf{R}_Z,\, \mathbf{R}_Z)$, the adjoint representation of an element $Z = a^\mu X_\mu$, with a^μ the parameters:

$$(\mathbf{R}_Z,\, \mathbf{R}_Z) \;=\; a^\mu\, a^\nu\, \mathrm{Tr}\,(\mathbf{X}_\mu\, \mathbf{X}_\nu) \;=\; a^\mu\, a^\nu\, g_{\mu\nu} \;=\; a^\mu\, a_\mu \tag{3.24}$$

$(\mathbf{R}_Z,\, \mathbf{R}_Z)$, and therefore $a^\mu a_\nu$, is not so hard to find, as illustrated in the following example.

Example 3.13. Go back to the defining representation used 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 structure constants are: $C_{12}{}^3 = 2$, $C_{31}{}^2 = -2$, and $C_{23}{}^1 = -2$. From these we build the adjoint-representation matrix:

$$\mathbf{R}_Z \;=\; \begin{pmatrix} 0 & 2x_3 & -2x_2 \\ -2x_3 & 0 & 2x_1 \\ -2x_2 & 2x_1 & 0 \end{pmatrix}$$

Now, we only need to calculate the diagonal elements of $\mathbf{R}_Z^2$ and sum them to get: $(\mathbf{R}_Z, \mathbf{R}_Z) = 8(x_1^2+x_2^2-x_3^2)$. We deduce that the algebra is non-compact. That X_1 and X_2 are non-compact, while X_3 is compact, was determined earlier in section 3.3.5.

Interestingly enough, using the defining representation directly, we would find (EXERCISE) $2(x_1^2 + x_2^2 - x_3^2)$. This is because for semisimple algebras the defining and regular representations are both faithful, and thus contain the same information, opening up the possibility of calculating $a^\mu\, a_\mu$ in eq. (3.24) directly from the defining representation instead of the more unwieldy regular representation.

3.4.5 Cartan subalgebra of a semismple algebra

Now we would very much like to find whether some elements H_i of a semisimple algebra have a *diagonalisable* adjoint-representation matrix, and thus satisfy the eigenvalue equation:

$$R_{H_i}(Y) \;=\; [H_i,\, Y] \;=\; \lambda_Y\, Y \tag{3.25}$$

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 elements of an algebra that commute between themselves, thus forming an *Abelian* (non-invariant!) subalgebra.

Definition 3.17. 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 unique up to isomorphism. The elements of a Cartan subalgebra are called its **Cartan generators**. Being Abelian, its irreducible representations are one-dimensional, and there exists a basis in which all Cartan generators are diagonal.

Example 3.14. An ordered basis of the complex extension of $\mathfrak{su}(2)$ (Example 3.9) *in its defining representation* is $\{S_-,\, S_0,\, S_+\}$, where $S_\pm = \frac{1}{\sqrt{2}}(S_1 \pm \mathrm{i}S_2)$ and $S_0 = S_3$, with $[S_i,\, S_j] = \mathrm{i}\epsilon_{ij}{}^k S_k$, or: $[S_0,\, S_\pm] = \pm S_\pm$, and $[S_+,\, S_-] = S_0$. Then the adjoint representation for S_0 and $S_\pm$ is:

$$\mathrm{ad}_{S_0} \;=\; \begin{pmatrix} -1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad \mathrm{ad}_{S_+} \;=\; \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix} \qquad \mathrm{ad}_{S_-} \;=\; \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix}$$

Because ad_{S_0} is diagonal, S_0 is a Cartan generator; comparing with eq. (3.25), ad_{S_0} has a *complete* set $\{S_-,\, S_0,\, S_+\}$ of eigengenerators for the corresponding eigenvalues $\{-1,\, 0,\, 1\}$, which form a basis of the algebra. But neither S_+ nor S_- is diagonalisable and they are not Cartan generators, Thus, the algebra contains only one Cartan generator and is of rank 1.

Another important thing we learn from this is that the structure constants in the complex extension of an algebra can be quite different from those of the algebra itself, even in its essentially real version. Indeed, the adjoint representation of S_3 found in example 3.9 is not diagonal, and has only zeros on its diagonal, in contrast with with ad_{S_0}, although ad_{S_3} does diagonalise to ad_{S_0}. Of course, this does not affect the CK-form which, being a trace, is basis-independent.

It can be shown that the rank of a $\mathfrak{su}(n)$ algebra is $n-1$; also, $\mathfrak{so}(2n)$ and $\mathfrak{so}(2n+1)$ have rank n.

3.5 Weights and Roots of a Representation of a Compact Semisimple Algebra

Definition 3.18. 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 generators. Then $H_i|\mu\rangle_{\mathbf{D}} = \mu_i|\mu\rangle_{\mathbf{D}}$. The set $\{\mu_i\}$ corresponding to each eigenvector can themselves be viewed as the components of a r-dimensional vector called a **weight** $\boldsymbol{\mu}$ of the representation. The number of these weights is the dimension of $\mathbf{D}$.

To find the n weights (often called a **multiplet**) of a representation $\mathbf{D}$ with matrices of rank n, simply identify a set of r Cartan generators H_i in $\mathbf{D}$, and diagonalise them if they are not in diagonal form. The i^{th} $(1 \le i \le r)$ component of the j^{th} weight $(1 \le j \le n)$ is the $(jj)^{\text{th}}$ entry of the $n \times n$ matrix representing H_i. These weights correspond to a point on a r-dimensional **weight diagram**, or **lattice**.

Definition 3.19. In a semisimple algebbra there exists a basis in which the non-Cartan generators (aka Weyl generators) $E_{\boldsymbol{\alpha}}$ of a semisimple algebra satisfy: $[H_i,\, E_{\boldsymbol{\alpha}}] = \alpha_i E_{\boldsymbol{\alpha}}$, $1 \leq i \leq r$ The $E_{\boldsymbol{\alpha}} \in \mathfrak{g}$ are then *eigengenerators* (often confusingly called **root vectors** by mathematicians) of the element H_i of the Cartan subalgebra. Then the set $\{\alpha_i\}$ of eigenvalues can be viewed as the components of a r-dimensional vector called the **root** $\boldsymbol{\alpha}$. We can also write $[\mathbf{H},\, E_{\boldsymbol{\alpha}}] = \boldsymbol{\alpha} E_{\boldsymbol{\alpha}}$. In any representation (defining, adjoint), this basis is called the **Cartan-Weyl basis**.

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, whose eigenvalues are the components of the weights. Also, *the roots do not depend on the representation* D, whereas the weights do. Indeed, one often speaks of the weights of D as being the representation itself.

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 define the **root system** of the algebra in a r-dim space.

As we are soon to discover, all the information about a semisimple algebra is encoded in its root system. A *Euclidean* metric is induced on their space by the 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. Being of smaller dimension, this root space is almost always much easier to work with than the algebra itself.

3.5.1 Properties of eigengenerators in the Cartan-Weyl basis

Those eigengenerators of H_i, $E_{\boldsymbol{\alpha}} \in \mathfrak{g}$ (all of them generators in 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 3.19, there comes:

$$\mathrm{ad}_{H_i}\big([E_{\boldsymbol{\alpha}},\, E_{\boldsymbol{\beta}}]\big) \;=\; \big[H_i,\, [E_{\boldsymbol{\alpha}},\, E_{\boldsymbol{\beta}}]\big] \;=\; \big[[H_i,\, E_{\boldsymbol{\alpha}}],\, E_{\boldsymbol{\beta}}\big] \,+\, \big[E_{\boldsymbol{\alpha}},\, [H_i,\, E_{\boldsymbol{\beta}}]\big] \;=\; (\alpha_i \,+\, \beta_i)\,[E_{\boldsymbol{\alpha}},\, E_{\boldsymbol{\beta}}]$$

If $\boldsymbol{\alpha} + \boldsymbol{\beta}$ is not a root, $[E_{\boldsymbol{\alpha}},\, E_{\boldsymbol{\beta}}] = 0$. Otherwise, $[E_{\boldsymbol{\alpha}},\, E_{\boldsymbol{\beta}}]$ is an eigengenerator with root $\boldsymbol{\alpha} + \boldsymbol{\beta}$, so we can write:

$$[E_{\boldsymbol{\alpha}},\, E_{\boldsymbol{\beta}}] \;=\; C_{\boldsymbol{\alpha}\boldsymbol{\beta}} E_{\boldsymbol{\alpha}+\boldsymbol{\beta}} \qquad\qquad C_{\boldsymbol{\alpha}\boldsymbol{\beta}} \in \mathbb{R} \tag{3.26}$$

We quote without proof two useful expressions for the CK-form of a semisimple algebra:

$$(H_i,\, E_{\boldsymbol{\alpha}}) \;=\; 0, \qquad\qquad (E_{\boldsymbol{\alpha}},\, E_{\boldsymbol{\beta}}) \;= 0 \quad \text{unless } \boldsymbol{\alpha} \,+\, \boldsymbol{\beta} \,=\, 0$$

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 that $[H_i,\, E_{\boldsymbol{\alpha}}^\dagger] = -\alpha_i E_{\boldsymbol{\alpha}}^\dagger$, so that $E_{\boldsymbol{\alpha}}^\dagger = E_{-\boldsymbol{\alpha}}$. Thus, non-Cartan generators and non-zero roots always come in pairs, $E_{\boldsymbol{\alpha}}$ and $E_{-\boldsymbol{\alpha}}$. In fact, $-\boldsymbol{\alpha}$ is the *only* possible multiple of $\boldsymbol{\alpha}$ which is a root; it always exists, otherwise $(E_{\boldsymbol{\alpha}},\, Z) \;= 0\; \forall\, Z \in \mathfrak{g}$, and the CK-form would be degenerate. Now we know how to compute the non-Cartan generators in the Cartan-Weyl basis from the pairs X_k and X_l of non-Cartan generators of the algebra: $E_{\pm\boldsymbol{\alpha}} = A(X_k \pm \mathrm{i} X_l)$, with A a normalisation contant.

When $\boldsymbol{\beta} \;=\; -\boldsymbol{\alpha}$, eq. (3.26) maps $[E_{\boldsymbol{\alpha}},\, E_{\boldsymbol{\beta}}]$ to a generator with zero root, ie. one that lives in the Cartan subalgebra. Therefore, $[E_{\boldsymbol{\alpha}},\, E_{-\boldsymbol{\alpha}}] = \lambda^i H_i$ for $1 \leq i \leq r$. Taking the inner product with H_j, one quickly shows (EXERCISE), using eq. (3.23) or the cyclic property of the trace, that $\lambda^i = \alpha^i\big(E_{\boldsymbol{\alpha}},\, E_{-\boldsymbol{\alpha}}\big)$, so that:

$$[E_{\boldsymbol{\alpha}},\, E_{-\boldsymbol{\alpha}}] \;=\; \big(E_{\boldsymbol{\alpha}},\, E_{-\boldsymbol{\alpha}}\big)\, h^{ij} \alpha_j H_i \;=\; \big(E_{\boldsymbol{\alpha}},\, E_{-\boldsymbol{\alpha}}\big)\, k\, \alpha_i H_i$$

where we have noted that $h^{ij} = k\, \delta^{ij}$ for a semisimple algebra. Now $[H_i,\, E_{\boldsymbol{\alpha}}] = \alpha_i E_{\boldsymbol{\alpha}}$ determines $E_{\boldsymbol{\alpha}}$ only up to a normalisation constant, which can be chosen so as to make $\big(E_{\boldsymbol{\alpha}},\, E_{-\boldsymbol{\alpha}}\big)$ cancel k, leaving the more simple:

$$[E_{\boldsymbol{\alpha}},\, E_{-\boldsymbol{\alpha}}] \;=\; \alpha_i\, H_i \;\coloneqq \boldsymbol{\alpha} \cdot \mathbf{H} \qquad\qquad \text{summation implied} \tag{3.27}$$

Now is a good time to discover what those non-Cartan generators do for a living. We have:

$$H_i\left(E_{\pm\boldsymbol{\alpha}}|\mu\rangle\right) \;=\; [H_i,\, E_{\pm\boldsymbol{\alpha}}]|\mu\rangle \;+\; E_{\pm\boldsymbol{\alpha}}H_i|\mu\rangle \;=\; (\mu_i \pm \alpha_i)\big(E_{\pm\boldsymbol{\alpha}}|\mu\rangle\big) \tag{3.28}$$

We see that $E_{\pm\boldsymbol{\alpha}}|\mu\rangle$ is an eigenvector of H_i with eigenvalue $\mu_i \pm\alpha_i$. The $E_{\pm\boldsymbol{\alpha}}$ act as raising/lowering operators on the carrier space of the Cartan generators, changing weights $\boldsymbol{\mu}$ by $\pm\boldsymbol{\alpha}$. Often, the quickest way to obtain the roots is to work out all the possible differences between *neighbouring* weights of a low-dimension representation.

3.6 Irreducible representations of semisimple algebras

Each irreducible representation of a Lie algebra can be labelled with the eigenvalues of some function of the basis generators of the algebra.

3.6.1 Casimir invariant operators

Definition 3.20. A **Casimir invariant operator** C (H. Casimir's thesis, 1931) 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, work by C. Chevalley and Harish-Chandra (1951) guarantees the existence of a set of Casimir operators as polynomials in the generators, whose eigenvalues may be used to label the irreducible representations of the algebra. More precisely, each invariant subspace of the carrier space has a set of basis vectors, each labelled by an eigenvalue of *each* Casimir operator. The number of algebraically independent Casimir operators is the rank of the algebra (sometimes called Racah's theorem).

In other words, if $f(\mathbf{x})$ is in an invariant subspace of the carrier space of the algebra, then for each Casimir operator C_i, $C_i f(\mathbf{x}) = g(\mathbf{x})$ is also in that same invariant subspace.

Because a metric can always be defined for a semisimple algebra, I claim that $C_2 \coloneqq g^{\mu\nu}X_\mu X_\nu$ is a Casimir operator, called the **quadratic Casimir invariant**, and where the X_μ are basis generators of the algebra. Indeed:

$$\begin{aligned}\left[g^{\mu\nu}\,X_\mu\,X_\nu,\,X_\rho\right] \;&=\; g^{\mu\nu}\left(X_\mu\,[X_\nu,\,X_\rho] \;+\; [X_\mu,\,X_\rho]\,X_\nu\right) \;=\; g^{\mu\nu}\,C_{\mu\rho}{}^{\lambda}\,(X_\nu\,X_\lambda \;+\; X_\lambda\,X_\nu)\\ &=\; g^{\mu\nu}\,g^{\alpha\lambda}\,f_{\mu\rho\alpha}\,(X_\nu\,X_\lambda \;+\; X_\lambda\,X_\nu)\\ &=\; 0\end{aligned}$$

since $g^{\mu\nu}\,g^{\alpha\lambda}f_{\mu\rho\alpha}$ is antisymmetric in, and the term in round brackets is symmetric in, ν and λ.

Example 3.15. From example 3.12, the metric for $\mathfrak{so}(3)$ is, up to a constant, $g_{\mu\nu} = \delta_{\mu\nu}$. Then:

$$C_2 \;=\; X^\mu\,X_\mu \;=\; J_x^2 \;+\; J_y^2 \;+\; J_z^2 \;=\; J^2$$

where **J** is the angular momentum operator of quantum mechanics. Since $\mathfrak{so}(3)$ is of rank 1, this is the only Casimir invariant. Then the eigenvalues of J^2 each label an irreducible representation of $\mathfrak{so}(3)$.

Note that because of its construction, C_2 is not in the algebra. In a Cartan-Weyl basis, it takes the form:

$$C_2 \;=\; g^{ij}\,H_iH_j \;+\; \sum_{+\text{roots}}\left(E_{-\boldsymbol{\alpha}}E_{\boldsymbol{\alpha}} \;+\; E_{\boldsymbol{\alpha}}E_{\boldsymbol{-\alpha}}\right) \;=\; g^{ij}\,H_iH_j \;+\; \sum_{+\text{roots}}\left(2\,E_{\pm\boldsymbol{\alpha}}E_{\mp\boldsymbol{\alpha}} \mp \boldsymbol{\alpha}\cdot\mathbf{H}\right) \tag{3.29}$$

3.6.2 Irreducible representations of $\mathfrak{so}(3)$

We now show how working in the Cartan-Weyl basis yields without much effort the irreducible representations of $\mathfrak{so}(3)$, the 3-parameter algebra of the group of 3-dim rotations. All we will need in this approach are the commutation relations in the standard basis: $[J_i,\, J_j] = \mathrm{i}\,\epsilon_{ij}{}^{k}\,J_k$.

Because of these relations, only one generator can be diagonalised. Then the eigenvalues m of this Cartan generator label the weights in each irreducible representation. The other two generators are non-Cartan. The single, one-component root $\boldsymbol{\alpha}$ can be normalised to 1. In a Cartan-Weyl basis, $\{E_{-1},\, H_1,\, E_1\} = \{J_-, J_0, J_+\}$, where $J_\pm = (J_1 \pm \mathrm{i}\,J_2)/\sqrt{2}$. From definition 3.19, $[J_0\,, J_\pm] = \pm J_\pm$. Eq. (3.27) then leads directly to: $[J_+,\, J_-] = J_0$.

We know enough to find the eigenvalues λ of the Casimir operator J^2 which label *all* irreducible representations. By definition of a Casimir operator J_0 and $J_\pm$ commute with J^2, so that: $J^2(J_\pm|\,\lambda\,m\rangle) = J_\pm(J^2|\,\lambda\,m\rangle) = \lambda\,(J_\pm|\,\lambda\,m\rangle)$. Also, eq. (3.28) becomes for $\mathfrak{so}(3)$: $J_0(J_\pm|\,\lambda\,m\rangle) = (m\,\pm\,1)\,(J_\pm|\,\lambda\,m\rangle)$.

Then J_+ raises, and J_- lowers, the weights m by 1, but they cannot transform $|\,m\rangle$ to an eigenvector of J^2 with a different eigenvalue λ. *All* the weights in a given invariant subspace are eigenvectors of J^2 with the *same* λ.

Next, we write relation (3.29) between the C_2 Casimir operator and the generators for $\mathfrak{so}(3)$:

$$J^2 \;=\; 2J_\pm\,J_\mp \,+\, J_0^2 \,\mp\, J_0 \tag{3.30}$$

Since an irreducible representation must be finite-dimensional, we expect that for a given λ there exists a highest weight, $m_{\max} \equiv j$, and also a lowest weight, $m_{\min} \equiv j'$. Then $J_+|\,j\rangle = 0$ and $J_-|\,j'\rangle = 0$. There comes:

$$\begin{aligned} J^2|\,j\rangle \;&=\; j^2|\,j\rangle \,+\, j\,|\,j\rangle \;=\; j(j+1)\,|\,j\rangle \;=\; \lambda\,|\,j\rangle \\ J^2|\,j'\rangle \;&=\; (j')^2|\,j'\rangle \,-\, j'|\,j'\rangle \;=\; j'(j'-1)\,|\,j'\rangle \;=\; \lambda\,|\,j'\rangle \end{aligned}$$

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, and that 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. (3.30), we can now exhibit the full action of J_- on a weight $|\,jm\rangle$ of J^2 and J_0. Let $J_-|\,jm\rangle = c_-\,|\,j, m-1\rangle$. Then, if $|\,jm\rangle$ is normalised:

$$\langle jm|J_+J_-\,|jm\rangle \;=\; c_-^*c_- \;=\; |c_-|^2$$

But since $2J_\pm\,J_\mp = J^2 - J_0^2 \,\pm\, J_0$, we also have that:

$$\langle jm|J_+J_-\,|jm\rangle \;=\; \frac{1}{2}\,\langle jm|(J^2 \,-\, J_0^2 \,+\, J_0)\,|jm\rangle \;=\; \frac{1}{2}\,\big(j(j+1) \,-\, m^2 \,+\, m\big)$$

Comparing yields c_- up to an unimportant phase factor which we put equal to ± 1. We find the coefficient in $J_+|\,jm\rangle = c_+\,|\,j, m+1\rangle$ in a strictly analogous way. Then, up to an arbitrary sign:

$$J_\pm|\,jm\rangle \;=\; \frac{1}{\sqrt{2}}\,\sqrt{j(j+1) \,-\, m(m\,\pm 1)}\,|\,j, m\pm 1\rangle \tag{3.31}$$

Each value of j labels a $2j{+}1$-dim invariant subspace of the carrier space of $\mathfrak{so}(3)$ of which the $|\,jm\rangle$ form a basis.

The entries of the three representation matrices, $D^j_{J_0} = \langle jm'|J_0|\,jm\rangle$ and $D^j_{J_\pm} = \langle jm'|J_\pm\,|jm\rangle$, of the $\mathfrak{so}(3)$ generators are:

$$(D^j_{J_0})^{m'}{}_{m} \;=\; m\,\delta^{m'}{}_{m} \qquad\qquad (D^j_{J_\pm})^{m'}{}_{m} \;=\; \frac{\delta^{m'}{}_{m\pm 1}}{\sqrt{2}}\,\sqrt{(j\mp m)(j\pm m+1)} \qquad |m| \;\le\; j \tag{3.32}$$

This form for the coefficients is often quoted, but the equivalent form in eq. (3.31) 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 the defining representation of $\mathfrak{so}(3)$, labelled by $j = 1$:

$$D^1_{J_+} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \qquad D^1_{J_0} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad D^1_{J_-} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

Any other irreducible representation for integer values of j can be calculated in the same way with eq. (3.32).

Another approach relies on the actual form of the generators. In the defining, irreducible 3-dim representation of the Cartesian basis, the three generators, which we choose to be Hermitian, are:

$$J_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\mathrm{i} \\ 0 & \mathrm{i} & 0 \end{pmatrix} \qquad J_2 = \begin{pmatrix} 0 & 0 & \mathrm{i} \\ 0 & 0 & 0 \\ -\mathrm{i} & 0 & 0 \end{pmatrix} \qquad J_3 = \begin{pmatrix} 0 & -\mathrm{i} & 0 \\ \mathrm{i} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{3.33}$$

Diagonalise, say, J_3 with the transformation $J_i \longmapsto \mathbf{A}^{-1} J_i \mathbf{A}$, where $\mathbf{A}$ is a unitary matrix so as to preserve Hermiticity, and construct the non-Cartan generators in the Cartan-Weyl basis as before:

$$J_+ = \begin{pmatrix} 0 & -\mathrm{i} & 0 \\ 0 & 0 & \mathrm{i} \\ 0 & 0 & 0 \end{pmatrix} \quad J_0 = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad J_- = \begin{pmatrix} 0 & 0 & 0 \\ \mathrm{i} & 0 & 0 \\ 0 & -\mathrm{i} & 0 \end{pmatrix} \qquad \mathbf{A} = \frac{1}{\sqrt{2}} \begin{pmatrix} -\mathrm{i} & 0 & \mathrm{i} \\ 1 & 0 & 1 \\ 0 & \sqrt{2} & 0 \end{pmatrix}$$

Although these generators are different from the $\mathcal{D}^1$ matrices obtained from eq. (3.32), they are perfectly acceptable as an irreducible representation since they satisfy both $[J_0\,, J_\pm] = \pm J_\pm$ and $[J_+,\, J_-] = J_0$. Indeed, any pair of the form $J_+ = \left(\begin{smallmatrix} 0 & a & 0 \\ 0 & 0 & b \\ 0 & 0 & 0 \end{smallmatrix}\right)$ and $J_- = \left(\begin{smallmatrix} 0 & 0 & 0 \\ a^* & 0 & 0 \\ 0 & b^* & 0 \end{smallmatrix}\right)$, wih $(a,\, b) \in \mathbb{C}$, satisfies these commutation relations! All these irreducible representations in the Cartan-Weyl basis are equivalent.

3.6.3 Cartan-Weyl basis for $\mathfrak{su}(2)$ in the defining representation

Take as the defining representation of the semisimple algebra $\mathfrak{su}(2)$ the Hermitian generators $\mathbf{S} = \boldsymbol{\sigma}/2$:

$$S_1 = \frac{1}{2} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad S_2 = \frac{1}{2} \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix} \qquad S_3 = \frac{1}{2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

One Cartan and one pair of non-Cartan generators fit, thus one *independent* 1-dim non-zero root. The diagonal S_3 is identified with the sole Cartan generator s_0. In this representation the weights of s_0 are $1/2$ and $-1/2$, for the doublet of eigenvectors $\left(\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right)$. The roots are all the possible differences between the weights, ie. ± 1.

Fom the definition of roots, $[s_0,\, E_{\pm 1}] = \pm E_{\pm 1}$. The structure of the algebra then determines the non-Cartan generators in the Cartan-Weyl basis. $[s_0,\, E_{\pm 1}] = \pm E_{\pm 1}$ gives, up to a normalisation constant A:

$$s_+ \coloneqq E_1 = A \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = A(S_1 + \mathrm{i}\, S_2), \qquad s_- \coloneqq E_{-1} = A \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = A(S_1 - \mathrm{i}\, S_2)$$

With $h^{11} = 1/h_{11} = (\operatorname{Tr} s_0^2)^{-1} = 2$ in the argument leading to eq. (3.27), we find: $\big(s_+,\, s_-\big) = \operatorname{Tr}(s_+\, s_-) = A^2$. The choice $A = 1$ recovers the commutator $[s_+,\, s_-] = 2s_0$, as in quantum mechanics, whereas the choice $A = 1/\sqrt{2}$ makes $\operatorname{Tr}(s_+\, s_-)$ cancel h^{11}, yielding $[s_+,\, s_-] = s_0$, with the structure constant just the root component. Then the set $\{s_+,\, s_0,\, s_-\}$ forms a Cartan-Weyl basis for the complex extension of $\mathfrak{su}(2)$.

Now our discussion of the $\mathfrak{so}(3)$ algebra in section 3.6.2 allows us to take j to be a multiple of $1/2$. Inserting $j = 1/2$ into eq. (3.32) then yields 2×2 generators identical to those of the defining representation of $\mathfrak{su}(2)$, whether in the standard or the Cartan-Weyl basis. This confirms the isomorphism of the $\mathfrak{su}(2)$ and $\mathfrak{so}(3)$ algebras.

3.6.4 Irreducible epresentations of SU(2) and SO(3)

Finite $SU(2)$ and $SO(3)$ transformations can be reconstructed with the exponential map, corresponding to a rotation parametrised by $\boldsymbol{\theta} = \theta\hat{\mathbf{n}}$: $R(\boldsymbol{\theta}) = \mathrm{e}^{\mathrm{i}\theta\hat{\mathbf{n}}\cdot\mathbf{J}}$ for $SO(3)$, and $S(\boldsymbol{\theta}) = \mathrm{e}^{\mathrm{i}\theta\hat{\mathbf{n}}\cdot\mathbf{S}} = \mathbf{I}\cos\frac{\theta}{2} - 2\mathrm{i}\,(\hat{\mathbf{n}}\cdot\mathbf{S})\sin\frac{\theta}{2}$ (EXERCISE) for $SU(2)$, 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 $SO(3)$! A $SO(3)$ rotation by 2π is identical to the identity, but a $SU(2)$ rotation by 2π is equivalent to *minus* the identity, because of the factor $1/2$ lurking in the $\mathbf{s}$ matrices. We say that $SU(2)$ and $SO(3)$ are *locally* isomorphic.

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 only $SU(2)$ matrices that correspond to integer j are *stricto sensu* representations of $SO(3)$. 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 basis generators of the defining 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 $(\mathcal{D}^j_\theta)^{m'}{}_m = \mathrm{e}^{\mathrm{i}m\theta}\delta^{m'}{}_m$ ($|m| \le j$). They are tabulated in many places for small j and are easily calculated by computer.

3.6.5 $\mathfrak{su}(2)$ substructure of a semisimple algebra and constraints on its root system

Because they live in a r-dim space, only r of the $n - r$ roots of a semisimple algebra can be linearly independent.

> **Definition 3.21.** A **positive** root is one whose first non-zero component is positive; otherwise, it is **negative**. The r positive roots which cannot be obtained from a linear combination of other positive roots are called **simple**, **fundamental**, or **independent**. The other positive roots can be obtained as linear combinations of the simple roots, with *positive* coefficients.

Each pair $e_{\pm\boldsymbol{\alpha}} \coloneqq \sqrt{2}E_{\pm\boldsymbol{\alpha}}/|\boldsymbol{\alpha}|$ of normalised non-Cartan generators of a semisimple algebra, together with the combination: $h_{\boldsymbol{\alpha}} = 2\boldsymbol{\alpha}\cdot\mathbf{H}/|\boldsymbol{\alpha}|^2$, forms a $\mathfrak{su}(2)$ subalgebra. *There is a* $\mathfrak{su}(2)$ *subalgebra for each pair of non-zero roots* (Chevalley 1955). $\{h_{\boldsymbol{\alpha}},\, e_{\pm\boldsymbol{\alpha}}\}$ is called the Chevalley basis of the $\mathfrak{su}(2)$ subalgebra. Indeed, $[e_{\boldsymbol{\alpha}},\, e_{-\boldsymbol{\alpha}}] = h_{\boldsymbol{\alpha}}$, but also:

$$\big[h_{\boldsymbol{\alpha}},,\, e_{\pm\boldsymbol{\alpha}}\big] \;=\; \frac{2\sqrt{2}\,\boldsymbol{\alpha}}{|\boldsymbol{\alpha}|^3}\cdot[\mathbf{H},\, E_{\pm\boldsymbol{\alpha}}] \;=\; \pm\,\frac{2\sqrt{2}\,|\boldsymbol{\alpha}|^2}{|\boldsymbol{\alpha}|^3}\,E_{\pm\boldsymbol{\alpha}} \;=\; \pm\,2\,e_{\pm\boldsymbol{\alpha}}$$

With $h_{\boldsymbol{\alpha}} = 2s_0$ and $e_{\pm\boldsymbol{\alpha}} = s_\pm$, we recover the $\mathfrak{su}(2)$ structure constants in the Cartan-Weyl basis. Thus, a semisimple algebra of dimension n and rank r contains $(n-r)/2$ generally non-distinct $\mathfrak{su}(2)$ subalgebras, each associated with a different root and having as Cartan generator a different element of the Cartan subalgebra, plus two non-Cartan generators corresponding to that root.

Roots are tightly constrained by the $\mathfrak{su}(2)$ substructure described above. Consider some other root $\boldsymbol{\beta}$. Then:

$$\Big[h_{\boldsymbol{\alpha}},\, e_{\pm\boldsymbol{\beta}}\Big] \;=\; \frac{2\sqrt{2}\,\boldsymbol{\alpha}}{|\boldsymbol{\alpha}|^2\,|\boldsymbol{\beta}|}\cdot\big[\mathbf{H},\, E_{\pm\boldsymbol{\beta}}\big] \;=\; \pm\,\frac{2\sqrt{2}\,\boldsymbol{\alpha}\cdot\,\boldsymbol{\beta}}{|\boldsymbol{\alpha}|^2\,|\boldsymbol{\beta}|}\,E_{\pm\boldsymbol{\beta}} \;=\; \pm 2m\,e_{\pm\boldsymbol{\beta}} \qquad\qquad m \coloneqq \frac{\boldsymbol{\alpha}\cdot\,\boldsymbol{\beta}}{|\boldsymbol{\alpha}|^2}$$

Since $h_{\boldsymbol{\alpha}}/2$ is a Cartan generator of $\mathfrak{su}(2)$, we may have found another $\mathfrak{su}(2)$ subalgebra if we can make sense of m. Now let $\boldsymbol{\beta} + k\boldsymbol{\alpha}$ ($k \in \mathbb{Z}$) be a non-zero root. Then, in the same way as above:

$$\Big[\frac{h_{\boldsymbol{\alpha}}}{2},\, e_{\boldsymbol{\beta}+k\boldsymbol{\alpha}}\Big] \;=\; \frac{\boldsymbol{\alpha}\cdot\,(\boldsymbol{\beta}+k\boldsymbol{\alpha})}{|\boldsymbol{\alpha}|^2}\,e_{\boldsymbol{\beta}+k\boldsymbol{\alpha}} \;=\; (m\;+\;k)\,e_{\boldsymbol{\beta}+k\boldsymbol{\alpha}}$$

Now let p and q be two non-negative integers, with p the largest number for which $\boldsymbol{\beta} + p\boldsymbol{\alpha}$ is still a root, and q be the largest number for which $\boldsymbol{\beta} - q\boldsymbol{\alpha}$ is still a root. So we have a string, or chain $e_{\boldsymbol{\beta}-q\boldsymbol{\alpha}}, \ldots, e_{\boldsymbol{\beta}}, \ldots, e_{\boldsymbol{\beta}+p\boldsymbol{\alpha}}$ of $\mathfrak{su}(2)$ generators that act in the *root space*, raising or lowering in *integer* steps. All elements in the set $\{\boldsymbol{\beta} + k\boldsymbol{\alpha}; k = -q, \ldots\, m, \ldots p\}$ are roots. We can associate each of these generators with one in a $(2j+1)$-dim $\mathfrak{su}(2)$ representation labelled by j. It takes p unit steps to go from m to the highest root, and q steps to go to the lowest

root in the chain, so that $-j+q=m$ and $j-p=m$, leading to $q-p=2m$, and $p+q=2j$. As expected for a $\mathfrak{su}(2)$ algebra, m (and j) is an integer or half-integer. We arrive at the **master formula**[†]:

$$-p \leq 2\frac{\boldsymbol{\alpha}\cdot\boldsymbol{\beta}}{|\boldsymbol{\alpha}|^2} = -(p-q) < q \tag{3.34}$$

If we had started instead with $e_{\boldsymbol{\alpha}}$ and added/subtracted integer multiples of $\boldsymbol{\beta}$ to $\boldsymbol{\alpha}$, we would have found that $2\boldsymbol{\beta}\cdot\boldsymbol{\alpha}/|\boldsymbol{\beta}|^2 = -(p'-q')$. Multiplying the two master formulae yields the important expression:

$$\frac{(\boldsymbol{\alpha}\cdot\boldsymbol{\beta})^2}{|\boldsymbol{\alpha}|^2|\boldsymbol{\beta}|^2} = \cos^2\theta_{\boldsymbol{\alpha}\boldsymbol{\beta}} = \frac{1}{4}(p-q)(p'-q') \leq 1 \tag{3.35}$$

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, $\pm(\boldsymbol{\alpha}-\boldsymbol{\beta})$ cannot be a root; otherwise, one of the two must be positive, and a simple root could be constructed out of two different positive roots: eg., $\boldsymbol{\beta}=(\boldsymbol{\beta}-\boldsymbol{\alpha})+\boldsymbol{\alpha}$. Thus, $\boldsymbol{\beta}-k\boldsymbol{\alpha}$ is not a root for any $k\neq 0$, including $k=q$, and $q=0$ for simple roots. Therefore, from the master formula (3.34), the angle between two *simple* roots satisfies $\cos\theta_{\boldsymbol{\alpha}\boldsymbol{\beta}}\leq 0$, so that $\pi/2\leq\theta_{\boldsymbol{\alpha}\boldsymbol{\beta}}\leq\pi$.

Since $(p-q)(p'-q')$ must be an integer, There are only five possible values allowed for $\cos^2\theta_{\boldsymbol{\alpha}\boldsymbol{\beta}}$ in eq. (3.35), and this, for any two roots of any semisimple algebra: $0 \Rightarrow \theta_{\boldsymbol{\alpha}\boldsymbol{\beta}}=\pm 90°$; $1/4 \Rightarrow \theta_{\boldsymbol{\alpha}\boldsymbol{\beta}}=(60°,120°)$; $1/2 \Rightarrow \theta_{\boldsymbol{\alpha}\boldsymbol{\beta}}=(45°,135°)$; $3/4 \Rightarrow \theta_{\boldsymbol{\alpha}\boldsymbol{\beta}}=(30°,150°)$; and $1 \Rightarrow \theta_{\boldsymbol{\alpha}\boldsymbol{\beta}}=180°$ ($0°$ is forbidden because no two roots can be a positive multiple of each other).

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. The root diagrams exhibit a high degree of symmetry. All positive roots can be generated by linear combinations of the simple roots. So-called **Weyl reflections** about hyperplanes perpendicular to the simple roots through the origin generate the rest.

With the subscript denoting the rank of the algebra, the four families of semisimple groups are:

- A_{n-1} $(n>1)$, corresponding to $SU(n)$, $SL(n,\mathbb{R})$, $SU(p,q)$, with $p+q=n$ (not 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, 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. Instead, one uses Dynkin diagrams, which are planar and represent only the simple roots and the angle between them. They are equivalent to a root diagram.

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 3.6.2) 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 starting from the highest weight of the representation, defined as the weight $\boldsymbol{\mu}$ for which $\boldsymbol{\mu}+\boldsymbol{\alpha}$ is not a weight when $\boldsymbol{\alpha}$ is *any* positive root. and applying the lowering non-Cartan generator of $\mathfrak{su}(2)$ to the weights in each $\mathfrak{su}(2)$ multiplet, 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 up from the lowest to the highest weight), is always a root of the semisimple algebra. Needless to say, as one moves parallel to the H_1 axis, all other components in the weights remain the same. Subtracting a simple root from the highest weight yields the highest weight of a neighbouring $\mathfrak{su}(2)$ multiplet.

[†]A derivation that does not rely on the $\mathfrak{su}(2)$ substructure can be found in Appendix I; but it involves rather heavier calculations.

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-dim representation (decuplet) of $\mathfrak{su}(3)$ of rank 2; thus the weights are 2-component vectors. The weights lie on an inverted-triangle lattice with one horizontal $\mathfrak{su}(2)$ quadruplet, triplet, doublet and singlet, in the direction of decreasing H_2 eigenvalues.

3.7 More on finding irreducible representations

3.7.1 Tensor product representations

Definition 3.22. Let $f_{j_1 m_1}$ and $f_{j_2 m_2}$ be two basis functions in the carrier space of irreducible representations $\mathcal{D}^{j_1}_{\rm g}$ and $\mathcal{D}^{j_2}_{\rm g}$, respectively, of ${\rm g} \in SU(2)$ or $SO(3)$, such that:

$$S_{\rm g}\, f_{j_1 m_1} \;=\; f_{j_1 m_1'}\, (\mathcal{D}^{j_1}_{\rm g})^{m_1'}{}_{m_1}\,, \qquad\qquad S_{\rm g}\, f_{j_2 m_2} \;=\; f_{j_2 m_2'}\, (\mathcal{D}^{j_2}_{\rm g})^{m_2'}{}_{m_2}$$

Then we form the **product representation** $\mathcal{D}^{j_1}_{\rm g} \otimes \mathcal{D}^{j_2}_{\rm g}$:

$$S_{\rm g}\, f_{j_1 m_1}\, f_{j_2 m_2} \;=\; f_{j_1 m_1'}\, f_{j_2 m_2'}\, (\mathcal{D}^{j_1})^{m_1'}{}_{m_1}\, (\mathcal{D}^{j_2})^{m_2'}{}_{m_2} \tag{3.36}$$

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. (3.36) using the generic expansion $\mathcal{D} = \mathbf{I} + a^i \mathbf{D}_{X_i}$, where $\mathbf{D}_X$ stands for a generator of $SU(2)$ or $SO(3)$ in that representation. 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:

$$\mathbf{D}^{1\otimes 2}_X (f_{j_1 m_1}\, f_{j_2 m_2}) \;=\; (\mathbf{D}_{X^1}\, f_{j_1 m_1})\, f_{j_2 m_2} \;+\; f_{j_1 m_1}\, (\mathbf{D}_{X^2}\, f_{j_2 m_2}) \tag{3.37}$$

that is: $\mathbf{D}^{1\otimes 2}_X = \mathbf{D}_{X^1} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{D}_{X^2}$ 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:

$$\phi_{jm} \;=\; \sum_{m_1, m_2} \big(j_1 m_1,\, j_2, m_2\, |jm\big)\, f_{j_1 m_1}\, f_{j_2 m_2} \tag{3.38}$$

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.

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. (3.37), 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.

3.7.2 Irreducible 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 3.23. Let $\{T_{jm}\}$ be a set of operators on the carrier space that transform as:

$$R_{\boldsymbol{\theta}}\, T_{jm}\, R^{-1}_{\boldsymbol{\theta}} = T_{jm'}\,(\mathcal{D}^j_{\boldsymbol{\theta}})^{m'}{}_m \tag{3.39}$$

Then we say that they are the components of a rank-j **irreducible (or spherical) tensor**.

If we linearise this equation, we obtain (EXERCISE) a more useful alternative definition of irreducible tensors in terms of generators $J^{(j)}$ of irreducible representations of the algebra, preferably in the Cartan-Weyl basis:

$$[J^{(j)},\, T_{jm}] = T_{jm'}\,(J^{(j)})^{m'}{}_m \qquad \text{(no summation on } j\text{)} \tag{3.40}$$

where j is the label of the irreducible representation. For SU(2) or SO(3):

$$[J^{(j)}_0,\, T_{jm}] = m\,T_{jm}, \qquad [J^{(j)}_\pm,\, T_{jm}] = \sqrt{(j\mp m)(j\pm m+1)}\,T_{j,m\pm1} \tag{3.41}$$

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 a version of the famous vector addition rules.

3.7.3 The Wigner-Eckart theorem

The Wigner-Eckart theorem asserts 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:

$$\langle j_2 m_2|\, T_{jm}\, |j_1 m_1\rangle = \frac{(j_1 m_1,\, jm|j_2 m_2)}{\sqrt{2j_2+1}}\,\langle j_2||\,\mathbf{T}_j\,||j_1\rangle \tag{3.42}$$

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 $SU(2)$. 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 m are just ratios of Clebsch-Gordan coefficients.

Example 3.16. When $\mathbf{T}$ transforms as a scalar under some Lie group, the relevant representation matrix of the group is the identity matrix. For $SU(2)$, $j = m = 0$, and the vector-addition rules collapse the Wigner-Eckart theorem to: $\langle j_2 m_2|\,\mathbf{T}\,|j_1 m_1\rangle = \delta^{j_1}{}_{j_2}\,\delta^{m_1}{}_{m_2}\langle j_2||\,\mathbf{T}\,||j_1\rangle/\sqrt{2j_2+1}$. Then matrix elements of scalar operators between weights of different irreducible representations vanish.

The importance of the Wigner-Eckart theorem resides in its separating symmetry-related ("geometrical") aspects of matrix elements from other ("dynamical") aspects stored in the possibly unknown reduced matrix element.

3.7.4 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. the wave-functions of two particles), each transforming in some known way under 3-dim rotations. We can form tensor products, $f(\mathbf{x})\otimes g(\mathbf{y})$ 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 (with nine weights, or basis vectors for its carrier space) , with components T^{ij}, which under rotations R would tranform as:

$$T'^{ij} = R^i{}_k\, R^j{}_l\, T^{kl} \tag{3.43}$$

The 6-dim symmetric part of T^{ij} rotates into a symmetric object, and the 3-dim antisymmetric part into an antisymmetric one. Thus, we have easily found invariant subspaces. Moreover, the trace of T^{ij}, $T^i{}_i$, is invariant under rotations, forming a 1-dim invariant subspace that 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 that is also invariant under rotation. Thus, we can write:

$$T^{ij} = \frac{1}{2}\left(T^{ij} + T^{ji}\right) + \frac{1}{2}\,\epsilon^{ijk}\,\epsilon_{klm}\,T^{lm} = \frac{1}{2}\left(T^{ij} + T^{ji} - \frac{2}{3}\,g^{ij}\,T^k{}_k\right) + \frac{1}{2}\left(T^{ij} - T^{ji}\right) + \frac{1}{3}\,g^{ij}\,T^k{}_k \quad (3.44)$$

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. (3.43) as a $3 \otimes 3$ exterior direct product of 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, $T^{ij} \longmapsto (R_2\,R_1)^i{}_m (R_2\,R_1)^j{}_n T^{mn}$, where now the 9×9 matrix is formed from the *matrix* product $R_2 R_1$. Being reducible, the representation 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.

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 3.7.1 to obtain a direct sum of three irreducible representations labelled by $j = 2$, $j = 1$, and $j = 0$.

Appendices

H Commutators of Angular Momentum with Vector Operators

Take a unit vector $\hat{\mathbf{u}}$ and a vector operator $\mathbf{V}$ with components V_u with respect to $\hat{\mathbf{u}}$. In example 3.6 we found that under a rotation $R(\boldsymbol{\theta})$ by a small angle θ about an axis $\hat{\mathbf{n}}$, $\hat{\mathbf{u}}' = \hat{\mathbf{u}} + \theta\,\hat{\mathbf{n}} \times \hat{\mathbf{u}}$. Then:

$$V_u' \;=\; \mathbf{V}\cdot\hat{\mathbf{u}}' \;=\; \mathbf{V}\cdot\hat{\mathbf{u}} + \theta\,\mathbf{V}\cdot\hat{\mathbf{n}}\times\hat{\mathbf{u}} \;=\; V_u \;+\; \theta\,\mathbf{V}\cdot\hat{\mathbf{n}}\times\hat{\mathbf{u}}$$

Also,

$$V_u' \;=\; R(\boldsymbol{\theta})\,V_u\,R^\dagger(\boldsymbol{\theta}) \;=\; \mathrm{e}^{-\mathrm{i}\,\theta\hat{\mathbf{n}}\cdot\mathbf{L}}\,V_u\,\mathrm{e}^{\mathrm{i}\,\theta\hat{\mathbf{n}}\cdot\mathbf{L}} \;\approx\; (1-\mathrm{i}\,\theta\hat{\mathbf{n}}\cdot\mathbf{L})\,V_u\,(1+\mathrm{i}\,\theta\hat{\mathbf{n}}\cdot\mathbf{L}) \;\approx\; V_a \;-\; \mathrm{i}\,\theta\,\big[\hat{\mathbf{n}}\cdot\mathbf{L},\,V_u\big] \quad \text{(H.1)}$$

Consistency then demands that: $\big[\hat{\mathbf{n}}\cdot\mathbf{L},\,V_u\big] = \mathrm{i}\,\mathbf{V}\cdot\hat{\mathbf{n}}\times\hat{\mathbf{u}} = \mathrm{i}\,\epsilon_{ijk}\,V^k\,n^i\,u^j$. With $\hat{\mathbf{n}}$ along the the x-axis and $\hat{\mathbf{u}}$ along the y-axis, there comes:

$$\big[L_i,\,V_j\big] \;=\; \mathrm{i}\,\epsilon_{ijk}\,V^k \quad \text{(H.2)}$$

I Alternative Derivation of the Master Formula

This derivation involves finding the (real!) structure constants $C_{\boldsymbol{\alpha\beta}}$ in eq. (3.26), *without explicit calculation of commutators*. They satisfy symmetry relations, such as, from (3.26) and its adjoint:

$$C_{\boldsymbol{\beta\alpha}} \;=\; -\,C_{\boldsymbol{\alpha\beta}} \qquad\qquad C_{-\boldsymbol{\alpha},-\boldsymbol{\beta}} \;=\; -\,C^*_{\boldsymbol{\alpha\beta}} \;=\; -\,C_{\boldsymbol{\alpha\beta}} \quad \text{(I.1)}$$

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. (3.26) and (3.27), leads (EXERCISE) to:

$$(\boldsymbol{\alpha}\,C_{\boldsymbol{\beta\gamma}} \;+\; \boldsymbol{\beta}\,C_{\boldsymbol{\gamma\alpha}} \;+\; \boldsymbol{\gamma}\,C_{\boldsymbol{\alpha\beta}})\cdot\mathbf{H} \;=\; 0$$

The H_i being linearly independent, this 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:

$$C_{\boldsymbol{\beta},-\boldsymbol{\alpha}-\boldsymbol{\beta}} \;=\; C_{-\boldsymbol{\alpha}-\boldsymbol{\beta},\boldsymbol{\alpha}} \;=\; C_{\boldsymbol{\alpha\beta}} \quad \text{(I.2)}$$

Eq. (3.26) leads to: $[E_{\boldsymbol{\alpha}},\,E_{\boldsymbol{\beta}+\boldsymbol{\alpha}}] = C_{\boldsymbol{\alpha},\boldsymbol{\beta}+\boldsymbol{\alpha}}E_{\boldsymbol{\beta}+2\boldsymbol{\alpha}}$, …, $[E_{\boldsymbol{\alpha}},\,E_{\boldsymbol{\beta}+k\boldsymbol{\alpha}}] = C_{\boldsymbol{\alpha},\boldsymbol{\beta}+k\boldsymbol{\alpha}}E_{\boldsymbol{\beta}+(k+1)\boldsymbol{\alpha}}$. But there must exist a value $k = p \geq 0$ such that $\boldsymbol{\beta}+(p+1)\boldsymbol{\alpha}$ is not a root; then $C_{\boldsymbol{\alpha},\boldsymbol{\beta}+p\boldsymbol{\alpha}} = 0$. Similarly, starting 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. (3.26) and (3.27):

$$\begin{aligned}
&\big[E_{\boldsymbol{\alpha}},\,[E_{\boldsymbol{\beta}+k\boldsymbol{\alpha}},\,E_{-\boldsymbol{\alpha}}]\big] \;+\; \big[E_{\boldsymbol{\beta}+k\boldsymbol{\alpha}},\,[E_{-\boldsymbol{\alpha}},\,E_{\boldsymbol{\alpha}}]\big] \;+\; \big[E_{-\boldsymbol{\alpha}},\,[E_{\boldsymbol{\alpha}},\,E_{\boldsymbol{\beta}+k\boldsymbol{\alpha}}]\big] \;=\; 0\\
\Longrightarrow\quad &[E_{\boldsymbol{\alpha}},\,E_{\boldsymbol{\beta}+(k-1)\boldsymbol{\alpha}}]\,C_{\boldsymbol{\beta}+k\boldsymbol{\alpha},-\boldsymbol{\alpha}} \;-\; [E_{\boldsymbol{\beta}+k\boldsymbol{\alpha}},\,\boldsymbol{\alpha}\cdot\mathbf{H}] \;+\; [E_{-\boldsymbol{\alpha}},\,E_{\boldsymbol{\beta}+(k+1)\boldsymbol{\alpha}}]\,C_{\boldsymbol{\alpha},\boldsymbol{\beta}+k\boldsymbol{\alpha}} \;=\; 0\\
\Longrightarrow\quad &C_{\boldsymbol{\alpha},\boldsymbol{\beta}+(k-1)\boldsymbol{\alpha}}\,C_{\boldsymbol{\beta}+k\boldsymbol{\alpha},-\boldsymbol{\alpha}} \;+\; \boldsymbol{\alpha}\cdot(\boldsymbol{\beta}+k\boldsymbol{\alpha}) \;+\; C_{-\boldsymbol{\alpha},\boldsymbol{\beta}+(k+1)\boldsymbol{\alpha}}\,C_{\boldsymbol{\alpha},\boldsymbol{\beta}+k\boldsymbol{\alpha}} \;=\; 0
\end{aligned}$$

Applying relations (I.1) and then (I.2) 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, $C_{\boldsymbol{\alpha},\boldsymbol{\beta}+p\boldsymbol{\alpha}} = 0$. 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 $C_{-\boldsymbol{\alpha},\boldsymbol{\beta}-q\boldsymbol{\alpha}} = -C_{\boldsymbol{\alpha},-(\boldsymbol{\beta}-q\boldsymbol{\alpha})} = -C_{\boldsymbol{\beta}-(q+1)\boldsymbol{\alpha},\boldsymbol{\alpha}} = 0$:

$$0 \;=\; C^2_{\boldsymbol{\alpha},\boldsymbol{\beta}-(q+1)\boldsymbol{\alpha}} \;=\; (p+q+1)\left[\boldsymbol{\alpha}\cdot\boldsymbol{\beta} \;+\; \frac{p-q}{2}\,|\boldsymbol{\alpha}|^2\right]$$

or:

$$2\,\boldsymbol{\alpha}\cdot\boldsymbol{\beta}/|\boldsymbol{\alpha}|^2 \;=\; -\,(p-q) \quad \text{(I.3)}$$

4 CHAPTER IV — 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:

$$f\left(y, \frac{\partial y}{\partial x^i}, \ldots, \frac{\partial^m y}{\partial x^i \partial x^j \ldots}, x^i\right) = 0 \tag{4.1}$$

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 *completely* 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 linear ODEs of first and second order, returning to PDEs later.

4.1 One-dimensional Linear Differential Operators

A differential operator L of order n is said to be **linear** over an interval $a \leq t \leq b$ if its action $L[f]$ on all the functions $f(t)$ in its domain $\mathcal{D}$ is linear in the functions and all their derivatives present in the operator. Linearity means that if f_1 and f_2 are any two functions $f_1, f_2 \in \mathcal{D}$, then $L[c_1 f_1 + c_2 f_2] = c_1 L[f_1] + c_2 L[f_2]$, where c_1 and c_2 are constants. **Formally** (without mention of $\mathcal{D}$), in one dimension L is written as:

$$L = \sum_{j=0}^{n} p_j(t)\, \mathrm{d}_t^j \tag{4.2}$$

L is not considered to be specified until $\mathcal{D}$ is given. At this stage, this consists of all n-times differentiable functions, but we will want to restrict it further. First of all, $\mathcal{D}$ should be a vector space, $\mathcal{H}$, whose elements are all the square-integrable functions on $[a, b]$, using the inner product:

$$(f, g) := \int_a^b f^*(t)\, g(t)\, \mathrm{d}t \tag{4.3}$$

We might want to know whether L has eigenfunctions, ie., whether there exist some $\phi \in \mathcal{D}$ such that L scales ϕ by a constant factor λ. This, of course, is the eigenvalue problem: $L[\phi] = \lambda\, \phi$. It requires $L : \mathcal{H} \to \mathcal{H}$, which is not the case in general for differential operators. Also, if it can be shown that the set $\{\phi_i\}$ of eigenfunctions is a basis for $\mathcal{H}$, then $L[y] \in \mathcal{H}, \forall\, y \in \mathcal{H}$. Achieving this will restrict the formal L itself, not only its domain.

Can L have 0 as eigenvalue? If so, we say that the associated eigenfunction, f_{h}, is a solution to the homogeneous equation $L[y] = 0$. Or, instead, we can restrict the image of L to be a specific function $F(t)$, the **source** or **driving** term, and look for the c orresponding function(s) in $\mathcal{D}$, if any. This means solving the inhomogeneous equation $L[y] = F$, and the existence and uniqueness of a solution, f_{inh}, implies the existence of an inverse operator L^{-1} such that $f_{\mathrm{inh}} = L^{-1}[F]$. As we are soon to discover, the existence and uniqueness of L^{-1} is connected to the eigenvalue problem with $\lambda = 0$.

Also, solving a n^{th}-order ODE requires specification of boundary conditions (B.C.) on the solution Thus, the domain of L also depends on the BC imposed on the functions on which it is allowed to act. Then invertibility of L is also very much dependent on those BC. We address each each of these questions in turn.

4.1.1 Existence of the inverse of a linear differential operator

In linear algebra it is well known that a matrix operator L has an inverse if, and only if, its determinant is non-zero, in which case the equation $\mathsf{LX} = \mathsf{Y}$ has the unique solution $\mathsf{X} = \mathsf{L}^{-1}\mathsf{Y}$. When L is a linear differential operator, we say that it is invertible if it is one-to-one. And it is one-to-one if, and only if, the only function it sends to zero is the zero function.

To prove this, assume that $L[f] = 0$ has for *unique* solution $f = 0$, and that $L[g] = L[h] = F$. Then $L[g] - L[h] = L[g - h] = 0$ by linearity, and we have $g = h$. Therefore, there is only one function mapped to F (including $F = 0$!) by L. Conversely, assume that L is one-to-one, and that there exists a function $f_{\text{h}} \neq 0$ such that $L[f_{\text{h}}] = 0$. But then Cf_{h}, with C an arbitrary constant, is also a solution of this equation, and L is many-to-one, contradicting our assumption. Thus, L is invertible if, and only if, it has no zero eigenvalue.

The existence of a non-trivial solution to the homogeneous equation depends very much on the boundary conditions imposed on the most general solution, which completely determine a unique solution if it exists.

4.1.2 Boundary Conditions

The most general solution to an inhomogeneous equation takes the form $f = f_{\text{h}} + f_{\text{inh}}$, ie., the sum of a homogeneous and an inhomogeneous solution, and where $L[f] = L[f_{\text{h}}] + L[f_{\text{inh}}] = L[f_{\text{inh}}] = F$. We require that the task of satisfying the B.C. fall entirely to f_{h}, so that f_{inh} satisfies *homogeneous* B.C., ie., it contributes nothing to the B.C. on f. As for the source term F, although it may happen that it vanishes at the boundaries, we do not want to constrain it other than being piecewise continuous.

In the theory of linear operators of order n, B.C. are often expressed in the form of n linear combinations of f and all its derivatives of order n-j $(1 \leq j \leq n)$, evaluated at the two end-points a and b, with $b > a$. These can be written as a matrix equation involving two matrices, A and B, with constant coefficients: $\mathsf{A}\mathbf{f}_a + \mathsf{B}\mathbf{f}_b = \mathbf{C}$, where $\mathbf{f}_a$ and $\mathbf{f}_b$ are vectors with components $\{f, \dot{f}, \ldots, \mathrm{d}_t^{n-1} f\}$ evaluated at a and b, respectively, and $\mathbf{C}$ is a given constant vector. When $\mathbf{C} = 0$, we say that the B.C. are homogeneous. To have f_{h} (and its derivatives) zero everywhere in the interval $[a, b]$, both $\mathbf{f}_a$ and $\mathbf{f}_b$ must vanish, which requires the B.C. to be homogeneous. Then L is indeed invertible. If a non-trivial f_{h} existed that satisfied homogeneous B.C., any constant multiple of f_{h} would also satisfy those B.C., and we would lose the uniqueness necessary for L to be invertible.

But what if the B.C. on f_{h} are inhomogeneous (non-zero)? Then we can consider two linear operators with the *same* form L: $L[\mathcal{D}^{\text{h}}]$, where $\mathcal{D}^{\text{h}}$ is the set of functions f_{h} with homogeneous B.C. that satisfy $L[f_{\text{h}}] = 0$, and $L[\mathcal{D}^{\text{inh}}]$, where $\mathcal{D}^{\text{inh}}$ is the set of functions f_{h} with *inhomogeneous* B.C. that satisfy $L[f_{\text{h}}] = 0$. Here, "inhomogeneous" applies to the B.C., not the inhomogeneous solution which always has homogeneous B.C. Only $L[\mathcal{D}^{\text{h}}]$ can be invertible, when $\mathcal{D}^{\text{h}}$ contains only the zero function.

In practice, B.C. are usually expressed as n arbitrary values assigned to f and/or its derivatives at the boundaries, and they come as two main types:

(1) **One-Point (Initial) conditions, aka Initial-Value Problem (IVP)**: In the formal theory, one matrix in $\mathsf{A}\mathbf{f}_a + \mathsf{B}\mathbf{f}_b = \mathbf{C}$, say B, is set to zero, so that only one point, the initial "time" a, is involved, and A is diagonal. Therefore, f and its n-1 derivatives take known values (or can be set arbitrarily) at $t = a$. Then a theorem shows that the solution to the one-dim IVP exists and is unique.

(2) **Two-point boundary conditions, or Boundary-Value Problem (BVP):** this time the n known or specified values of f and its derivatives can be at both end-points a and b. This is a much more complicated situation, with neither existence of a solution nor its uniqueness guaranteed.
In the most prevalent case, $f(a)$ and $f(b)$ are known (**Dirichlet problem**), or its first derivatives at a and b are known (**Neumann problem**). **Periodic** B.C., where $f(a) = f(b)$ and $\dot{f}\big|_a = \dot{f}\big|_b$, can also occur.

4.1.3 First-order linear ODEs

It is not difficult to obtain an *explicit* general solution to a first-order IVP. First, assume that the ODE has been converted to its **normal form**:

$$L[f] \;=\; \mathrm{d}_t f \,+\, \beta(t)\, f \;=\; F(t)$$

Then one shows (EXERCISE) that if $\beta(t)$ is continuous over $[a,\, b]$, the general solution of this IVP is:

$$f(t) \;=\; \frac{i(a)}{i(t)}\, f(a) \,+\, \int_a^t F(t')\, \frac{i(t')}{i(t)}\, \mathrm{d}t' \qquad\qquad i(t) \;=\; \mathrm{e}^{\int^t \beta(t')\, \mathrm{d}t'} \tag{4.4}$$

Notice that $f_\mathrm{h} = i(a)f(a)/i(t)$ solves the homogeneous equation: $\mathrm{d}_t f + \beta(t) f \;= 0$, and that the inhomogeneous term in eq. (4.4) satisfies homogeneous B.C., as expected. Thus, $f_\mathrm{h}(a) = f(a)$. When $f(a) = 0$, $f_\mathrm{h}(t) = 0$.

With $f(a)$ specified, the solution is unique. Indeed, let $g(t)$, with $g(a) = f(a)$, also satisfy $L[g] = F$. Then $h = g - f$ solves the homogeneous ODE with homogeneous B.C. $h(a) = 0$, which forces $h(t) = 0$ for $t > a$.

4.1.4 Second-order linear ODEs

The most general form for a linear, second-order equation over the interval $[a,\, b]$ is:

$$L[f] \;=\; \alpha(t)\, \mathrm{d}_t^2 f \,+\, \beta(t)\, \mathrm{d}_t f \,+\, \gamma(t)\, f \;\;=\; F(t) \tag{4.5}$$

where $\alpha \neq 0$, β and γ are required to be continuous, while F is piecewise continuous.

Introduce the **Wronskian** of two differentiable functions, $f_1(t)$ and $f_2(t)$, defined as: $W(t) \coloneqq f_1 \dot{f}_2 - f_2 \dot{f}_1$. If there exists no constant C such that $f_2 = C f_1 \;\forall\, t$, f_1 and f_2 are said to be **linearly independent**. The Wronskian provides a handy test for linear independence: two differentiable functions *that do no vanish anywhere* in an interval are linearly dependent over that interval if, and only if, their Wronskian vanishes everywhere (EXERCISE).

The Wronskian of two *homogeneous* solutions of eq. (4.5) obeys a first-order differential equation whose solution is **Abel's formula** (EXERCISE):

$$W(t) \;=\; W(t_0)\, \mathrm{e}^{-\int_{t_0}^t [\beta(t')/\alpha(t')]\mathrm{d}t'} \tag{4.6}$$

where t_0 is any point in the interval $[a, b]$. In the important case that $\beta = \dot{\alpha}$, eq. (4.6) leads to αW being constant. Also, if the Wronskian of two homogeneous solutions vanishes anywhere, it vanishes everywhere, because the exponential cannot vanish in a finite interval.

Given one solution, f_1, of eq. (4.5), an immediate useful application of the Wronskian generates a second linearly independent solution. Noticing that $W(t)/f_1^2 = \mathrm{d}_t\big(f_2/f_1\big)$ and integrating, we find with eq. (4.6) that:

$$f_2(t) \;=\; f_1(t) \int_a^t \frac{W(t_0)}{f_1^2(t')}\, \mathrm{e}^{-\int^{t'} (\beta/\alpha)\mathrm{d}t''}\, \mathrm{d}t' \tag{4.7}$$

Ignoring the unknown cconstant $W(t_0)$ and discarding any term proportional to f_1 leaves a solution that is linearly independent from f_1.

And now comes a surprising fact, courtesy also of the Wronskian: given two independent solutions of the *homogeneous* equation, a solution of the *inhomogeneous* equation: $L[f(t)] \;=\; F(t)$, can be generated which satisfies homogeneous B.C. Appendix J presents a simplified version, leading to eq. (J.1), of this **variation of parameters** method discovered by Euler and Lagrange. Shortly, however, we shall explore another method which yields the same results while providing much deeper insight.

Note also that if $\int \beta(t)/\alpha(t)\mathrm{d}t$ exists within the interval of interest, it is always possible to eliminate the first-order derivative term in any linear second-order ODE, with a redefinition of the form $f(t) = g(t)\mathrm{e}^{\mu(t)}$ (the substitution $f(t) = \mu(t)g(t)$ also works), to arrive (EXERCISE) at the **normal Sturm-Liouville** form:

$$\ddot{g}(t) + \frac{1}{\alpha(t)}\left(\gamma - \mathrm{d}_t\Big(\frac{\beta}{2\alpha}\Big) - \frac{1}{4}\frac{\beta^2}{\alpha}\right) g(t) = \frac{F(t)}{\alpha(t)}\,\mathrm{e}^{-\mu(t)} = \frac{F(t)}{\alpha(t)}\exp\left[\int^t \frac{\beta}{2\alpha}\mathrm{d}t'\right] \tag{4.8}$$

as determined by the requirement that the transformed ODE have no first-order derivative of g. In the frequent case $\alpha = 1$ and β and γ constants, this takes the much simpler form: $\ddot{g}(t) + \big(\gamma - (\beta/2)^2\big)g(t) = F(t)\mathrm{e}^{\beta t/2}$.

Let $f_1(t)$ and $f_2(t)$ be two independent solutions of $L[f] = 0$. Then $f_\mathrm{h} = c_1 f_1 + c_2 f_2$, with c_1 and c_2 determined from the B.C. on f_h, is the general solution of the homogeneous equation (principle of linear superposition).

4.1.5 Second-order IVP

It can be shown (the technical proof is presented in Appendix K) that the only *homogeneous* solution of a second-order IVP, $L[f] = F$, for which f and $\mathrm{d}_t f$ *both* vanish at the initial point $t = a$, is the trivial solution $f_\mathrm{h} = 0$. Consequently, if there exist two solutions f and g such that $f(a) - g(a) = 0$ and $\dot{f}\big|_a - \dot{g}\big|_a = 0$, then $f = g$ everywhere. $\mathcal{D}^\mathrm{h} = \{0\}$, and $L[\mathcal{D}^\mathrm{h}]$ is indeed invertible. We conclude that the general solution to a second-order IVP always exists and is unique. With the inhomogeneous solution derived in Appendix J, we find:

$$f(t) = c_1\, f_1(t) + c_2\, f_2(t) + \int_a^\infty \theta(t-t')\left(\frac{f_1(t')\, f_2(t) - f_2(t')\, f_1(t)}{\alpha(t')\, W(t')}\right) F(t')\,\mathrm{d}t' \tag{4.9}$$

where $\theta(t-t')$ is the step-function which vanishes for $t < t'$ and equals 1 when $t > t'$, and:

$$c_1 = \frac{\dot{f}_2(a)\, f(a) - f_2(a)\, \dot{f}(a)}{W(a)}, \qquad c_2 = -\frac{\dot{f}_1(a)\, f(a) - f_1(a)\, \dot{f}(a)}{W(a)}$$

4.1.6 Second-order BVP

The Dirichlet problem for a 2^nd-order differential operator is also addressed in Appendix J, and we will re-visit it at length in the context of Green functions. The Neumann problem is fiddlier, and we will say much less about it. For a start, there may be a constraint on the B.C., or on the source. For instance, integrating $\mathrm{d}_x^2 f = F$ over $[a, b]$ immediately yields $\mathrm{d}_x f\big|_a^b = \int F(x)\mathrm{d}x$. For homogeneous Neumann B.C., this translates into a constraint on the source.

4.2 Solving One-dimensional Second-order Equations with Green Functions (BF 7.3)

4.2.1 Solutions in terms of Green Functions

We shall now investigate the conditions that allow the existence and uniqueness of f_inh formally written as $f_\mathrm{inh}(t) = [L^{-1}F](t)$, where L^{-1} is an integral operator whose action on $F(t)$ is:

$$\big[L^{-1}F\big](t) = \int G(t,\, t')\, F(t')\,\mathrm{d}t' \tag{4.10}$$

Assuming that F is square-integrable over some interval, we want L^{-1} to return a square-integrable result $\big[L^{-1}F\big](t)$, ie., $f_\mathrm{inh}(t)$; this is the case if the two-point function $G(t,\, t')$ is itself square-integrable over the interval (see the end of section BF7.1 for more details).

Now, acting on the above equation with L gives: $[Lf](t) = F(t) = \int [L_t G](t,\, t')\, F(t')\,\mathrm{d}t'$. This is satisfied provided $G(t,\, t')$ obeys:

$$\big[L_t G\big](t, t') = \delta(t - t') \tag{4.11}$$

where $\delta(t-t')$ is the Dirac delta-function. Eq. (4.11) will be the defining equation for a **Green function** $G(t,t')$ of L. We expect that any **indefinite** solution of eq. (4.11) must be supplemented with B.C. related to those on $f(t)$.

For the Green function to exist, L must be invertible, which we have seen requires that there be no non-trivial homogeneous solution with homogeneous B.C.

The link between the existence of the Green function and the criterion for invertibility of L can be made more tangible if L has a complete set of orthonormal eigenfunctions ϕ_j of L, with associated eigenvalues λ_j, on the interval. A version of the spectral theorem of operator theory asserts that such a set exists if L is in self-adjoint form, ie., if $\beta = \dot{\alpha}$ in eq. (4.5). Even if it isn't, it can always be put in such a form by multiplying it by a function w and imposing $\mathrm{d}_t(w\alpha) = w\beta$, which determines w up to a constant.

Then $f_{\rm inh}$ can be expanded over the subset that satisfies homogeneous B.C. (with unknown coefficients a_j), and so can F, with *known* coefficients $b_j = \int \phi_j^*(t')F(t')\,\mathrm{d}t'$. 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, assuming that integral and summation signs can be interchanged, there comes (EXERCISE) the inhomogeneous solution:

$$f_{\rm inh}(t) \;=\; \sum_j \frac{\phi_j(t)}{\lambda_j}\left[\int \phi_j^*(t')\,F(t')\,\mathrm{d}t'\right] \;=\; \int\left[\sum_j \phi_j(t)\,\phi_j^*(t')/\lambda_j\right]F(t')\,\mathrm{d}t'$$

We can write this solution as $f(t) = \int G(t,t')F(t')\,\mathrm{d}t'$, so long as the Green function:

$$G(t,t') \;=\; \sum_j \frac{\phi_j(t)\,\phi_j^*(t')}{\lambda_j} \tag{4.12}$$

exists, ie., only if there is *no non-trivial* ϕ_j satisfying homogeneous B.C. such that $L[\phi_j]=0$. Note, however, that even if $G(t,t')$ $\big($defined as obeying equation (4.11)$\big)$ does not exist, the *solution* $f(t)$ might still exist, provided that any ϕ_j associated with $\lambda_j = 0$ satisfies $b_j = 0$. But such a solution would be far from unique, because any multiple of ϕ_j could be added to it (see Appendix L for more details).

4.2.2 1-dim Green Functions without boundary conditions

What restrictions does eq. (4.11) impose on $G(t,t')$? Two, 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 is discontinuous, eg., $L = -\mathrm{id}_t$ has as Green function the step-function $\mathrm{i}\,\theta(t-t')$.

(b) $\dot{G}$ must have a discontinuity at $t=t'$. To see this, integrate eq. (4.11) from $t=t'-\epsilon$ to $t=t'+\epsilon$. Since the coefficients in L are continuous, they hardly vary when the interval is arbitrarily small ($\epsilon \to 0$). In that limit, the integrals of G and $\dot{G}$ both vanish because G is continuous, and only the integral of $\ddot{G}$ contributes:

$$\lim_{\epsilon\to 0}\dot{G}(t,t')\Big|_{t=t'-\epsilon}^{t=t'+\epsilon} \;=\; \frac{1}{\alpha(t')}$$

Because of the discontinuity in its derivative at $t=t'$, G should be different on either side while satisfying $[LG](t,t')=0$, so that it 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 discontinuity in $\dot{G}$ at $t=t'$ then yield the matrix equation at t':

$$\begin{pmatrix} f_1(t') & f_2(t') \\ \dot{f}_1(t') & \dot{f}_2(t') \end{pmatrix}\begin{pmatrix} a_1-b_1 \\ a_2-b_2 \end{pmatrix} \;=\; \begin{pmatrix} 0 \\ 1/\alpha \end{pmatrix}$$

For the system to have a solution, the determinant of the matrix, ie. the Wronskian, $W \equiv f_1\dot{f}_2 - \dot{f}_1 f_2$, cannot vanish anywhere, or else it would vanish everywhere, and f_1 and f_2 would not be independent as postulated. Then:

$$\begin{pmatrix} a_1 - b_1 \\ a_2 - b_2 \end{pmatrix} = \frac{1}{\alpha(t')\,W(t')} \begin{pmatrix} -f_2(t') \\ f_1(t') \end{pmatrix}$$

Eliminating a_1 and a_2 with this equation, the Green function for L without B.C. must take the general form:

$$G(t,t') = \begin{cases} b_1(t')\,f_1(t) + b_2(t')\,f_2(t) - \dfrac{f_1(t)\,f_2(t') - f_2(t)\,f_1(t')}{[\alpha\,W](t')} & t' < t \\ b_1(t')\,f_1(t) + b_2(t')\,f_2(t) & t' > t \end{cases} \tag{4.13}$$

The term with the Wronskian vanishes at $t = t'$, ensuring 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.

4.3 Green functions for the IVP and the BVP

4.3.1 Initial-value problem

In an IVP, we must impose the following initial conditions on G: $G(a,t') = \mathrm{d}_t G(t,t')\big|_{t=a} = 0$. Since $a < t'$, the relevant expression in the general Green function (4.13) is the one for $t < t'$, ie., the solution of the homogeneous equation which must vanish as seen in section 4.1.5 (or Appendix K), so $b_1 = b_2 = 0$. There comes the unique:

$$G_{\mathrm{ivp}}(t,t') = \theta(t-t')\,\frac{f_2(t)f_1(t') - f_1(t)\,f_2(t')}{[\alpha W](t')} \tag{4.14}$$

The step-function does not make $G_{\mathrm{ivp}}(t,t')$ discontinuous because the rest of the expression vanishes at $t' = t$. Eq. (4.9) can now be written as:

$$f(t) = \frac{\dot{f}_2(a)\,f(a) - f_2(a)\,\dot{f}(a)}{W(a)}\,f_1(t) + \frac{f_1(a)\,\dot{f}(a) - \dot{f}_1(a)\,f(a)}{W(a)}\,f_2(t) + \int_a^\infty G_{\mathrm{ivp}}(t,t')\,F(t')\,\mathrm{d}t' \tag{4.15}$$

with G_{ivp} given by eq. (4.14). A physicist is pleased that the B.C. guarantee causality: $G_{\mathrm{ivp}}(t' > t) = 0$.

4.3.2 Two-point boundary-value problem

Although superficially similar, the two-point boundary-value problem (BVP) requires a little more care. Many treatments enforce homogeneous B.C. on G_{bvp} by first choosing f_1 and f_2 such that they (or their derivatives) satisfy homogeneous B.C. at the end-points. Here, we follow a slightly different approach that, as in the IVP, initially only assumes linear independence of f_1 and f_2, without B.C. on f_1 and f_2.

We focus on the Dirichlet problem, where $f_{\mathrm{h}}(x)$ is specified at $x = a$ and $x = b$, with $a < b$. But we do impose homogeneous B.C. on the Dirichlet Green function G_{D}: $G_{\mathrm{D}}(a,x') = 0$ $(a < x')$ immediately leads to: $b_2(x') = -b_1(x')f_1(a)/f_2(a)$, whereas $G_{\mathrm{D}}(b,x') = 0$ $(x' < b)$ gives:

$$b_1(x') = \frac{f_2(a)}{[\alpha\,W](x')}\,\frac{f_2(b)\,f_1x') - f_1(b)\,f_2(x')}{f_1(a)\,f_2(b) - f_1(b)\,f_2(a)} \quad\Longrightarrow\quad b_2(x') = \frac{f_1(a)}{[\alpha\,W](x')}\,\frac{f_1(b)\,f_2(x') - f_2(b)\,f_1(x')}{f_1(a)\,f_2(b) - f_1(b)\,f_2(a)}$$

The resulting Dirichlet Green function factorises (EXERCISE) in x and x':

$$G_{\mathrm{D}}(x,x') = \frac{1}{[\alpha\,W](x')}\,\frac{\big[f_2(b)\,f_1(x_>) - f_1(b)\,f_2(x_>)\big]\big[f_2(a)\,f_1(x_<) - f_1(a)\,f_2(x_<)\big]}{f_1(a)\,f_2(b) - f_1(b)\,f_2(a)} \tag{4.16}$$

where $x_> := \max(x,x')$ and $x_< := \min(x,x')$. Linear independence of f_1 and f_2 guarantees the non-vanishing of W, but unlike an IVP, G_{D} exists only if $f_1(a)f_2(b) - f_1(b)f_2(a) \neq 0$.

The most simple case occurs when $f_1(a) = f_2(b) = 0$; then $f_1(b)$ and $f_2(a)$ drop out, leaving: $G_{\rm D}(x,x') = f_1(x_<)f_2(x_>)/\alpha(x')W(x')$. But with the BC $f_1(a) = f_1(b) = 0$, or $f_2(a) = f_2(b) = 0$, $G_{\rm D}$ does not exist.

We can now write down the general solution to the Dirichlet problem:

$$f(x) \;=\; \frac{f_2(b)\,f(a) - f_2(a)\,f(b)}{f_1(a)\,f_2(b) - f_1(b)\,f_2(a)}\,f_1(x) + \frac{f_1(a)\,f(b) - f_1(b)\,f(a)}{f_1(a)\,f_2(b) - f_1(b)\,f_2(a)}\,f_2(x) + \int_a^b G_{\rm D}(x,x')\,F(x')\,{\rm d}x' \quad (4.17)$$

The homogeneous Dirichlet B.C. $f(a) = f(b) = 0$ automatically prevent the existence of a non-zero homogeneous solution, as desired, but the extra condition $f_1(a)f_2(b) - f_1(b)f_2(a) \neq 0$ must still be met.

4.3.3 Examples

Example 4.1. A Helmholtz operator

Consider the operator $\mathrm{d}_t^2 + \omega_0^2$ with initial conditions on f and $\dot{f}$ at a single point (IVP). We choose the linearly independent $f_1 = \sin\omega_0 t$ and $f_2 = \cos\omega_0 t$. Also, noting that $\alpha = 1$ and $W = -\omega_0$, eq. (4.14) yields the IVP Green function:

$$G_{\rm ivp}(t,t') \;=\; G_{\rm ivp}(t-t') \;=\; \theta(t-t')\,\frac{\sin\big[\omega_0(t-t')\big]}{\omega_0}$$

Note the dependence of the IVP Green function on the *difference* $t - t'$. Indeed, it can be shown (EXERCISE) that for the second-order linear differential equation: $[Lf](t) = F(t)$ with *constant* coefficients, Green functions for a one-dim 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).

By contrast, for the same L, f_1 and f_2, (with $\omega_0 = k$), but with a Dirichlet problem at $a = 0$ and b, we immediately obtain from eq. (4.16):

$$G_{\rm D}(x,x') \;=\; \frac{1}{k\sin kb}\,\sin\big[k(x_> - b)\big]\,\sin kx_< \quad (4.18)$$

and, provided $kb \neq n\pi$, the unique inhomogeneous part of the solution to $(\mathrm{d}_x^2 + k^2)f(x) = F(x)$ is:

$$f_{\rm inh}(x) \;=\; \frac{\sin\big[k(x-b)\big]}{k\sin kb}\int_0^x \sin(kx')\,F(x')\,{\rm d}x' \;+\; \frac{\sin kx}{k\sin kb}\int_x^b \sin\big[k(x'-b)\big]\,F(x')\,{\rm d}x'$$

If $kb = n\pi$ $(n \in \mathbb{Z})$, ie., if b is an integer multiple of the half-period, the condition for the existence of a Dirichlet Green function, $f_1(a)f_2(b) - f_1(b)\,f_2(a) = -\sin kb \neq 0$, is violated.

Note that the same result would have followed from the initial choice $f_1(0) = f_2(b) = 0$, where $f_1 = \sin kx$ and $f_2 = \sin k(x-b)$, with now $W = -k\sin kb$. If $k = n\pi/b$ for some integer $n \neq 0$, $W = 0$, so that f_1 and f_2 are linearly dependent. $\phi_0(x) = \sin\big[n\pi(b-x)/b\big]$ satisfies homogeneous B.C. (at $x = 0$ and b) and solves the homogeneous equation. Thus, $L = \mathrm{d}_x^2 + (n\pi/b)^2$ is not invertible and the standard Green-function approach fails. As discussed in Appendix L, a modified Green function could still be constructed if $\phi_0(x)F(x)$ integrates to zero over the interval, but the complete solution would not be unique unless an extra normalisation condition is imposed on the homogeneous solutions.

Example 4.2. Eq. (4.13) has no *explicit* dependence on the coefficient of the first–order derivative in L. This reflects the option we know we have to eliminate it from a second-order equation. For instance, invoking eq. (4.8) with constant coefficients transforms the homogeneous equation for a damped harmonic oscillator, $(\mathrm{d}_t^2 + 2\gamma\mathrm{d}_t + \omega_0^2)f(t) = 0$, into $\mathrm{d}_t^2 g(t) + (\omega_0^2 - \gamma^2)g(t) = 0$, with $f(t) = g(t)\mathrm{e}^{-\gamma t}$. Inserting a solution of the form $\mathrm{e}^{\lambda t}$, we find the independent homogeneous solutions:

$f_1(t) = \mathrm{e}^{-\gamma t}\sin\left[\sqrt{\omega_0^2-\gamma^2}\,t\right]$, $f_2(t) = \mathrm{e}^{-\gamma t}\cos\left[\sqrt{\omega_0^2-\gamma^2}\,t\right]$. Now $W = -\sqrt{\omega_0^2-\gamma^2}\,\mathrm{e}^{-2\gamma t'}$, and a straightforward substitution into eq. (4.14) for an IVP gives:

$$G(t,t') \;=\; \theta(t-t')\,\mathrm{e}^{-\gamma(t-t')}\,\frac{\sin\left[\sqrt{\omega_0^2-\gamma^2}(t-t')\right]}{\sqrt{\omega_0^2-\gamma^2}} \tag{4.19}$$

Example 4.3. 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 complex-analysis techniques with Green functions. The idea is to write the equation:

$$\ddot{f}(t) \;+\; 2\gamma\,\dot{f}(t) \;+\omega_0^2\,f(t) \;=\; F(t)$$

in Fourier space, assuming that the driving term 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, writing the Fourier representation of a function and of its drivative:

$$\mathrm{d}_t f(t) \;=\; \frac{1}{\sqrt{2\pi}}\int \mathrm{d}_t\left[f(\omega)\,\mathrm{e}^{\mathrm{i}\omega t}\right]\mathrm{d}\omega \;=\; \frac{\mathrm{i}\,\omega}{\sqrt{2\pi}}\int f(\omega)\,\mathrm{e}^{\mathrm{i}\omega t}\,\mathrm{d}\omega$$

it is easy to see that our differential equation becomes:

$$\frac{1}{\sqrt{2\pi}}\int\left[f(\omega)\left(-\,\omega^2\;+\;2\mathrm{i}\,\gamma\,\omega\;+\;\omega_0^2\right)\;-\;F(\omega)\right]\mathrm{e}^{\mathrm{i}\omega t}\,\mathrm{d}\omega \;=\; 0$$

Then, because the Fourier transform of the zero function vanishes everywhere, the differential equation is turned into the *algebraic* equation:

$$f(\omega) \;=\; \frac{F(\omega)}{-\omega^2\;+\;2\mathrm{i}\,\gamma\,\omega\;+\omega_0^2}$$

Going back to t space, we write a solution to the inhomogeneous equation:

$$\begin{aligned} f_{\mathrm{inh}}(t) \;&=\; \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(\omega)\,\mathrm{e}^{\mathrm{i}\omega t}\,\mathrm{d}\omega \;=\; \int\left[\frac{1}{2\pi}\int\frac{\mathrm{e}^{\mathrm{i}\omega(t-t')}\,\mathrm{d}\omega}{-\omega^2\;+\;2\mathrm{i}\,\gamma\,\omega\;+\omega_0^2}\right]F(t')\,\mathrm{d}t' \\ &=\; \int_{-\infty}^{\infty} G(t,t')\,F(t')\,\mathrm{d}t' \end{aligned}$$

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:

$$\begin{aligned} G(t-t') \;&=\; -\,\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{\mathrm{e}^{\mathrm{i}\omega(t-t')}}{(\omega-\omega_+)(\omega-\omega_-)}\,\mathrm{d}\omega \;=\; 2\pi\mathrm{i}\,\frac{-1}{2\pi}\left(\frac{\mathrm{e}^{\mathrm{i}\omega_+(t-t')}}{\omega_+-\omega_-}\;-\;\frac{\mathrm{e}^{\mathrm{i}\omega_-(t-t')}}{\omega_+-\omega_-}\right) \\ &=\; \mathrm{e}^{-\gamma(t-t')}\,\frac{\sin\left[\sqrt{\omega_0^2-\gamma^2}(t-t')\right]}{\sqrt{\omega_0^2-\gamma^2}} \end{aligned}$$

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. (4.19) for an IVP. Here, however, no knowledge of the homogeneous solutions was needed to find the Green function! As for a BVP, if we can find a particular solution to eq. (4.11), we can enforce, eg., $G_{\rm D} = 0$ at the end-points for a Dirichlet problem, by adding a suitable term $\tilde{G}$ that satisfies $L[\tilde{G}) = 0$.

4.3.4 Green's second 1-dim identity and general solution of a BVP in terms of Green functions

Assume that a second-order linear operator $L_x = \alpha \mathrm{d}_x^2 + \beta \mathrm{d}_x + \gamma$ has been put in self-adjoint form, that is, $\beta = \alpha'$, with $\alpha' = \mathrm{d}_x \alpha$. Then a few manipulations (EXERCISE) lead to **Lagrange's identity**:

$$v\, L_x[u] \;-\; u\, L_x[v] \;=\; \big[\alpha\,(v\,u' \;-\; u\,v')\big]'$$

where $u, v \in \mathcal{D}$ of L. Integrate over an interval $[a,\, b]$ to obtain **Green's second identity** in one dimension:

$$\int_a^b \big(v\, L_x[u] \;-\; u\, L_x[v]\big)\,\mathrm{d}x \;=\; \alpha\,(v\,u' \;-\; u\,v')\Big|_{x=a}^{x=b} \tag{4.20}$$

Thanks to this identity, the homogeneous part, $f_{\rm h}$, of the solution to a BVP for $L[f] = F$ can be expressed in terms of the same Green function that appears in the inhomogeneous solution, and for *any B.C.*, homogeneous or not.

Indeed, suppose that $u = G(x,x')$ and that $v = f(x)$ is the general solution to the inhomogeneous equation. Then one easily shows from Green's identity that for $x' \in [a,\, b]$:

$$f(x') \;=\; \int_a^b G(x,x')\,F(x)\,\mathrm{d}x \;-\; \alpha\,\big[G(x,x')\,\partial_x f \;-\; f\,\partial_x G(x,x')\big]\Big|_{x=a}^{x=b} \tag{4.21}$$

where $G(x,x')$ is a Green function for L_x. We are already familiar with the first (inhomogeneous) term, but the second one warrants careful examination. Obviously, it must be related to the homogeneous solution. But wait—is $f(x')$ 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* f' at both a and b as this would be in general inconsistent. If f is specified at the end-points, then we must *first* find the solution for f in order to know what its derivatives are at the end-points.

For a Dirichlet problem, $G_{\rm D} = 0$ at the end-points. One important property of Dirichlet Green functions can immediately be derived by letting $v = G_{\rm D}(x'',x')$ and $u = G_{\rm D}(x'',x)$ in Green's second 1-dim identity (4.20), which holds for differential operators of the form $L_{x''} = \mathrm{d}_{x''}(\alpha \mathrm{d}_{x''}) + \gamma$. Because $G_{\rm D} = 0$ at the end-points and $\big[L_{x''}G\big](x'',y) = \delta(x''-y)$, we immediately find that $G_{\rm D}$ for such operators is symmetric in its arguments:

$$G_{\rm D}(x,x') \;=\; G_{\rm D}(x',x) \tag{4.22}$$

After interchanging x and x' in eq. (4.21), then using the symmetry of $G_{\rm D}$ in its arguments, and implementing the B.C. on $G_{\rm D}$, there comes the general *solution*:

$$f(x) \;=\; \int_a^b G_{\rm D}(x,x')\,F(x')\,\mathrm{d}x' \;+\; \Big[\alpha\, f\,\partial_{x'} G_{\rm D}\Big]_{x'=a}^{x'=b} \qquad\qquad G_{\rm D}(x,a) \;=\; G_{\rm D}(x,b) \;=\; 0 \tag{4.23}$$

Compare this form of the general solution, which explicitly depends only on $F(x)$ and $G_{\rm D}$, plus $f(a)$ and $f(b)$, to the solution (4.17) in terms of the linearly independent homogeneous solutions. It is a very instructive EXERCISE to show their equivalence. Also, if f happens to obey homogeneous B.C., $f(a) = f(b) = 0$, there is no homogeneous part; of course, $G_{\rm D}$ must exist for this approach to be ussable.

The Neumann problem, with $\mathrm{d}_x f$ specified at a and b, is not so straightforward. Just setting $\partial G_{\rm N}$ to zero at the end-points may be inconsistent, as one can see, eg., for $L = \mathrm{d}_x^2$, by integrating $\partial_x^2 G = \delta(x-x')$ once to get $\partial_x G\big|_a^b = 1$. Instead, since L is assumed to be in self-adjoint form, one can introduce a modified Green function $\mathcal{G}$ with the defining equation $L_x \mathcal{G} = \delta(x-x') - \phi_0(x)\phi_0(x')$, where ϕ_0 is a non-zero solution of the homogeneous equation: $[Lf](x) = 0$, with $\mathrm{d}_x\phi_0 = 0$ at the end-points (see Appendix L for the details).

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.

4.4 Linear Partial Differential Equations (PDE)

Unless you are working on superstrings, it is usually 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. Then. in Carttesian coordinates, the (simplified — no mxed derivatives) form of a second-order linear PDE that we shall use is:

$$L[f(\mathbf{x})] = \left[\alpha^\mu(\mathbf{x})\,\partial_\mu^2 \;+\; \beta^\nu(\mathbf{x})\,\partial_\nu \;+\; \gamma(\mathbf{x})\right] f(\mathbf{x}) \;=\; F(\mathbf{x}) \tag{4.24}$$

where $\mathbf{x}$ is the generalised position. As before, the coefficients are assumed to be continuous in $\mathbf{x}$.

We follow Hadamard (1923) and classify $L[f]$ according to the coefficients of the second-order derivatives:

Definition 4.1.
- If at least one of the α^μ vanishes at some point, the operator (and corresponding homogeneous PDE will be said to be **parabolic** at that point (eg. heat equation, Schrödinger equation, in which there is no second-order time-derivative).
- If the coefficients α^μ are not zero but one of them has a sign different from all others at some point, we say that L is **hyperbolic** at that point (eg., in Minkowski spacetime, the wave equation).
- If all α^μ coefficients, all non-zero, have the same sign at some point (as in a Euclidean space), L is **elliptic** at that point (eg. Laplace and Helmholtz operators — static 3-dim problems).

4.5 Separation of Variables in Elliptic Problems

Since the Laplacian operator occurs in almost all of these 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.

4.5.1 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:

$$\begin{aligned}\mathbf{L}\cdot\mathbf{L} \;&=\; -\,\epsilon_{ijk}\epsilon^{imn}\,x^j\partial^k\,(x_m\partial_n)\\ &=\; -\,x^j\left(\partial_j \;+\; x_j\,\partial^k\partial_k \;-\; 3\partial_j \;-\; x_k\partial^k\partial_j\right) \;=\; -\,x^jx_j\,\partial^k\partial_k \;+\; x^j\,\partial_j \;+\; x^j\partial_j(x^k\partial_k)\end{aligned}$$

Extracting the Laplacian and reverting to coordinate-free notation, there comes:

$$\nabla^2 \;=\; -\frac{\mathbf{L}^2}{r^2} \;+\; \frac{1}{r}\left[\partial_r \;+\; \partial_r(r\,\partial_r)\right] \;=\; -\frac{\mathbf{L}^2}{r^2} \;+\; \frac{1}{r^2}\,\partial_r(r^2\,\partial_r) \tag{4.25}$$

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 corresponds to the direction along $\mathbf{x}$; the other two coordinates are *angular*.

[†] Anyway, it is straightforward to generalise our discussion to any number of spatial dimensions plus one time dimension.

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:

$$\mathbf{L}^2 = -\left[\frac{1}{\sin\theta}\,\partial_\theta\,(\sin\theta\,\partial_\theta) + \frac{1}{\sin^2\theta}\,\partial_\phi^2\right] \tag{4.26}$$

So $\mathbf{L}^2$ depends only on the angular coordinates. Eq. (4.25) makes it obvious that the **commutator** $[\nabla^2,\ \mathbf{L}^2] = 0$.

Also, one readily shows (see section 3.2.4) that the following important relations hold:

$$[L_x,\ L_y] = \mathrm{i}\,L_z, \qquad [L_y,\ L_z] = \mathrm{i}\,L_x, \qquad [L_z,\ L_x] = \mathrm{i}\,L_y \tag{4.27}$$

By symmetry, we have immediately that $[\mathbf{L}^2, \mathbf{L}] = 0$.

4.5.2 Eigenvalues of $\mathbf{L}^2$ and L_z

Operators $\mathbf{J}$ that satisfy eq. (4.27) were studied earlier in section 3.6.2, where we found that the eigenvalues of $\mathbf{J}^2$ are $j(j+1)$, where l is a positive integer or half-integer. We also found that the eigenvalues m of J_z take $2j+1$ values, from $-j$ to j.

Since $[\mathbf{J}^2,\ \mathbf{J}] = 0$, $\mathbf{J}^2$ and J_z have a common set of eigenfunctions f_{jm}. The action of the raising and lowering operators $J_\pm = J_x \pm \mathrm{i}\,J_y$ on these eigenfunctions is given (up to normalisation) by eq. 3.31:

$$J_\pm f_{jm} = \sqrt{j(j+1) - m(m\pm 1)}\,f_{j,m\pm 1} \tag{4.28}$$

4.5.3 Eigenfunctions of $\mathbf{L}^2$ and $\mathbf{L}_z$

The eigenfunctions of $L_z = -\mathrm{i}\,\partial_\phi$ are readily obtained by solving the differential equation: $-\mathrm{i}\,\partial_\phi f(\theta,\phi) = m\,f(\theta,\phi)$. With a separation ansatz: $f(\theta,\phi) = F(\theta)H(\phi)$, the solution for H is:

$$H(\phi) = \mathrm{e}^{\mathrm{i}m\phi} \tag{4.29}$$

Now we require that H (and f) be single-valued, that is, $H(\phi + 2\pi) = H(\phi)$. 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, $l := m_{\max}$ must also be an integer. This is what rules out the possibility of half-integer eigenvalues allowed for a self-adjoint operator $\mathbf{J}$ that just satisfies the canonical commutation relations: $[J_i,\ J_j] = \mathrm{i}\epsilon_{ijk}J^k$.

The θ dependence of the eigenfunctions must be derived from the eigenvalue equation for $\mathbf{L}^2$. Call $f(\theta,\phi) = Y_l^m(\theta,\phi) = F(\theta)H(\phi)$; these must satisfy:

$$L^2\,Y_l^m(\theta,\phi) = -\left[\frac{1}{\sin\theta}\,\partial_\theta\,(\sin\theta\,\partial_\theta) + \frac{1}{\sin^2\theta}\,\partial_\phi^2\right] Y_l^m(\theta,\phi) = l(l+1)\,Y_l^m(\theta,\phi)$$

as well as $L_z\,Y_l^m(\theta,\phi) = mY_l^m(\theta,\phi)$, ie., $L_z\,H(\phi) = mH(\phi)$. Then $Y_l^m(\theta,\phi) = F(\theta)\mathrm{e}^{\mathrm{i}\,m\phi}$, and:

$$-\left[\frac{1}{\sin\theta}\,\mathrm{d}_\theta\,(\sin\theta\,\mathrm{d}_\theta) - \frac{m^2}{\sin^2\theta}\right] F(\theta) = l(l+1)\,F(\theta)$$

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**:

$$Y_l^m(\theta,\phi) \;=\; \frac{(-1)^m}{2^l l!}\sqrt{\frac{2l+1}{4\pi}\,\frac{(l-m)!}{(l+m)!}}\,(1-x^2)^{m/2}\left[\mathrm{d}_x^{l+m}(x^2-1)^l\right]\mathrm{e}^{\mathrm{i}m\phi} \qquad x = \cos\theta \tag{4.30}$$

4.5.4 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. (4.25), we have:

$$\begin{aligned}\nabla^2\Psi \;+\; \gamma(\mathbf{x})\Psi \;&=\; -\,\frac{\mathbf{L}^2\Psi}{r^2} \;+\; \frac{1}{r}\left[\partial_r\Psi \;+\; \partial_r(r\,\partial_r\Psi)\right] \;+\; \gamma(\mathbf{x})\,\Psi \\ &= -\,R(r)\frac{\mathbf{L}^2F(\theta,\phi)}{r^2} \;+\; \frac{F(\theta,\phi)}{r}\left[\mathrm{d}_rR(r) \;+\; \mathrm{d}_r(r\,\mathrm{d}_rR(r))\right] \;+\; \gamma(\mathbf{x})\,R(r)F(\theta,\phi)\end{aligned}$$

Multiplying the second line by $r^2/(R(r)F(\theta,\phi))$, 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:

$$\frac{1}{r^2}\,\mathrm{d}_r\left(r^2\,\mathrm{d}_rR_l(r)\right) \;+\; \left(\gamma(r) \;-\; \frac{l(l+1)}{r^2}\right)R_l(r) \;=\; 0$$

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:

$$\Psi(r,\theta,\phi) \;=\; \sum_{l=0}^{\infty}\sum_{m=-l}^{l}\left(A_{lm}\,r^l \;+\; \frac{B_{lm}}{r^{l+1}}\right)Y_l^m(\theta,\phi) \tag{4.31}$$

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:

$$\mathrm{d}_r^2 R_l(r) \,+\, \frac{2}{r}\,\mathrm{d}_r R_l(r) \,+\, \left(k^2 \,-\, \frac{l(l+1)}{r^2}\right) R_l(r) \;=\; 0$$

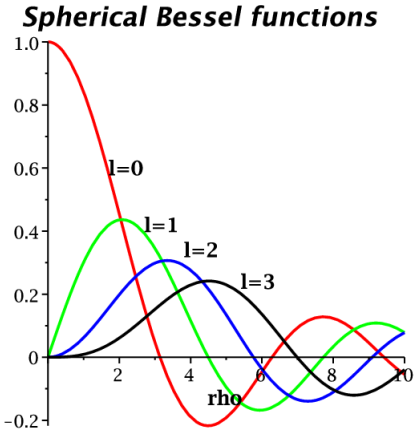
Defining dimensionless $x = kr$ readily transforms it into a form of the **Bessel equation** whose solutions are the **spherical Bessel functions** of the first and second (Neumann) kind, usually written as (see also Jackson's *Classical Electrodynamics*, section 9.6):

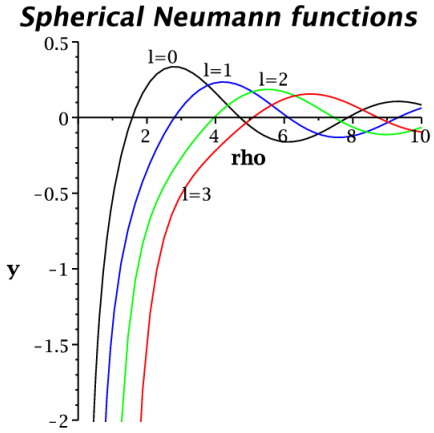
$$j_l(x) \;=\; (-x)^l \left(\frac{1}{x}\frac{\mathrm{d}}{\mathrm{d}x}\right)^l \left(\frac{\sin x}{x}\right) \qquad \sim \begin{cases} x^l & x \ll (1,l) \\ \dfrac{1}{x}\sin(x - l\pi/2) & x \gg l \end{cases} \tag{4.32}$$

$$n_l(x) \;=\; -\,(-x)^l \left(\frac{1}{x}\frac{\mathrm{d}}{\mathrm{d}x}\right)^l \left(\frac{\cos x}{x}\right) \qquad \sim \begin{cases} -\dfrac{1}{x^{l+1}} & x \ll (1,l) \\ -\dfrac{1}{x}\cos(x - l\pi/2) & x \gg 1 \end{cases} \tag{4.33}$$

The general solution of the Helmholtz equation is a linear combination of the j_l and n_l.

Here are a few spherical Bessel and Neumann functions as plotted on *Maple*, with $\rho = x$:





The n_l diverge at the origin and thus are excluded from any solution regular at the origin.

Spherical Bessel functions $h_l^{(1,2)}(x) = j_l(x) \pm \mathrm{i}\, n_l(x)$, aka **Hankel** functions of the first and second kind, can come in handy. One can express the general solution of the Helmholtz equation in terms of the $h_l^{(1,2)}$.

4.6 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 the self-adjoint form: $L[f] = \partial^i\big(\alpha(\mathbf{x})\partial_i f\big) + \gamma(\mathbf{x})f$.

Write the divergence theorem for $\boldsymbol{\nabla}\cdot(\alpha f\boldsymbol{\nabla}g)$ over a connected volume, and expand the divergence to get:

$$\int_V \Big[f\,\boldsymbol{\nabla}\cdot(\alpha\,\boldsymbol{\nabla}g) + \alpha\,\boldsymbol{\nabla}f\cdot\boldsymbol{\nabla}g\Big]\,\mathrm{d}^3x = \oint_{\partial V}\alpha\,f\,\boldsymbol{\nabla}g\cdot\mathrm{d}\mathbf{S} = \oint_{\partial V}\alpha\,f\,\boldsymbol{\nabla}g\cdot\hat{\mathbf{n}}\,\mathrm{d}S \tag{4.34}$$

where ∂V is the closed boundary of the volume V of integration, and the unit vector $\hat{\mathbf{n}}$ normal to ∂V, by convention, always points *outward* from the volume. This is **Green's first identity** in three dimensions; when α is a constant, and introducing the **normal derivative** $\partial_n = \hat{\mathbf{n}}\cdot\boldsymbol{\nabla}$, it reduces to the more familiar form:

$$\int_V \Big[f\,\nabla^2 g + \boldsymbol{\nabla}f\cdot\boldsymbol{\nabla}g\Big]\,\mathrm{d}^3x = \oint_{\partial V} f\,\boldsymbol{\nabla}g\cdot\mathrm{d}\mathbf{S} = \oint_{\partial V} f\,\partial_n g\,\mathrm{d}S \tag{4.35}$$

Interchanging f and g in the first identity (4.34) and subtracting, adding and subtracting γfg in the volume integral yields the **second Green identity** in three dimensions—compare with one-dim eq. (4.20):

$$\int_V \big(f\,L[g] - g\,L[f]\big)\mathrm{d}^3x = \oint_{\partial V}\alpha\,\big(f\,\boldsymbol{\nabla}g - g\,\boldsymbol{\nabla}f\big)\cdot\mathrm{d}\mathbf{S} = \oint_{\partial V}\alpha\,\big(f\,\partial_n g - g\,\partial_n f\big)\,\mathrm{d}S \tag{4.36}$$

With α a constant, this becomes the well-known Green theorem:

$$\int_V \big(f\,\nabla^2 g - g\,\nabla^2 f\big)\mathrm{d}^3x = \oint_{\partial V}\big(f\,\boldsymbol{\nabla}g - g\,\boldsymbol{\nabla}f\big)\cdot\mathrm{d}\mathbf{S} \tag{4.37}$$

Example 4.4. Uniqueness and existence of solutions for the Poisson equation with B.C.

The Poisson (inhomogeneous Laplace) equation is of the form $\nabla^2\Psi(\mathbf{x}) = F(\mathbf{x})$. We also specify B.C. for *either* Ψ *or* $\partial_n\Psi$ on ∂V. With $f = g = \Psi_3$ and α constant, eq. (4.35) becomes:

$$\int_V [\Psi_3\nabla^2\Psi_3 + (\boldsymbol{\nabla}\Psi_3)^2]\,\mathrm{d}^3x = \oint_{\partial V}\Psi_3\,\partial_n\Psi_3\,\mathrm{d}S$$

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 $\Psi_3 \coloneqq \Psi_2 - \Psi_1$. Then $\nabla^2\Psi_3 = 0$ inside the volume. The surface integral is zero because either $\Psi_3 = 0$ or $\partial_n\Psi_3 = 0$ 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, so that Ψ_3 is a constant. It follows immediately that if $\Psi_3 = 0$ on ∂V, $\Psi_1 = \Psi_2$ everywhere. On the other hand, when $\partial\Psi_3/\partial n = 0$ on ∂V, 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 overstated: any function that satisfies the inhomogeneous Laplace 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 solution.

The B.C. determine the solution, but only if it exists. Further conditions must be met for this to happen. Indeed, integrate $\nabla^2\Psi(\mathbf{x}) = F(\mathbf{x})$ over (connected!) V; the divergence theorem yields the condition:

$$\int_V F(\mathbf{x})\,\mathrm{d}^3x = \int_{\partial V}\partial_n\Psi(\mathbf{x})\,\mathrm{d}S \tag{4.38}$$

Another condition for the existence of a solution is that the enclosing boundary be "reasonably" smooth (eg. no spikes ...) if we wish to specify $\partial_n\Psi$ on ∂V.

Finally, if $\nabla^2\phi_n = \lambda_n\phi_n$, and taking $f = \phi_n^*$ and $g = \phi_n$ in eq. (4.35), one shows (EXERCISE) that the eigenvalues of the Laplacian are always negative.

4.7 3-dim Boundary Value (Elliptic) Problems with Green Functions

Consider $L_{\mathbf{x}}$ in self-adjoint form. Introduce Green functions that satisfy $\left[L_{\mathbf{x}}G\right](\mathbf{x},\mathbf{x}')\big] = \delta(\mathbf{x}-\mathbf{x}')$ (some authors multiply the right-hand side by $\pm 4\pi$) in regions with closed boundaries. If we are a little careful, we will find that for some B.C. this kind of problem can admit unique Green functions. Just as in one dimension, this requires that there exist no non-trivial solution to $L_{\mathbf{x}} f(\mathbf{x}) = 0$ with homogeneous B.C.

4.7.1 Dirichlet and Neumann Boundary Conditions for an Elliptic Problem

Suppose that $\Psi(\mathbf{x})$ satisfies[†] $L_{\mathbf{x}}\Psi(\mathbf{x}) = F(\mathbf{x})$. Proceeding exactly like in the 1-dim case, take $f = \Psi$ and $g = G$ in Green's second identity (eq. (4.36)):

$$\int_V \left(\Psi\, L_{\mathbf{x}}[G] - G\, L_{\mathbf{x}}[\Psi]\right) \mathrm{d}^3x \;=\; \oint_{\partial V} \alpha\, (\Psi\, \partial_n G - G\, \partial_n \Psi)\, \mathrm{d}S$$

We obtain:

$$\int_V \left[\Psi(\mathbf{x})\, \delta(\mathbf{x}-\mathbf{x}') \;-\; F(\mathbf{x})\, G(\mathbf{x},\mathbf{x}')\right] \mathrm{d}^3x \;=\; \oint_{\partial V} \alpha\, (\Psi\, \partial_n G \;-\; G\, \partial_n \Psi)\, \mathrm{d}S$$

With $\mathbf{x}'$ inside the volume, re-arranging then yields:

$$\Psi(\mathbf{x}') \;=\; \int_V F(\mathbf{x})\, G(\mathbf{x},\mathbf{x}')\, \mathrm{d}^3x \;+\; \oint_{\partial V} \alpha\, \left(\Psi\, \partial_n G - G\, \partial_n \Psi\right) \mathrm{d}S \tag{4.39}$$

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 any point on ∂V as such values will in general be inconsistent.

As before, specifying Ψ on ∂V gives **Dirichlet B.C.**, whereas specifying $\partial_n\Psi$ gives **Neumann B.C.**

For a Dirichlet problem we demand that $G_{\mathrm{D}}(\mathbf{x},\mathbf{x}') = 0\ \forall\, \mathbf{x} \in \partial V$. With Green's 2^{nd} identity, it is then quite easy to prove (EXERCISE) that $G_{\mathrm{D}}(\mathbf{x},\mathbf{x}')$ is symmetric in its arguments. After interchanging $\mathbf{x}$ and $\mathbf{x}'$ in eq. (4.39) and implementing the symmetry of G_{D}, the *solution* for Ψ is:

$$\Psi(\mathbf{x}) \;=\; \int_V F(\mathbf{x}')\, G_{\mathrm{D}}(\mathbf{x},\mathbf{x}')\, \mathrm{d}^3x' \;+\; \oint_{\partial V} \alpha\, \Psi(\mathbf{x}')\, \partial_{n'} G_{\mathrm{D}}(\mathbf{x},\mathbf{x}')\, \mathrm{d}S' \tag{4.40}$$

The Dirichlet solution is uniquely determined by the B.C. on Ψ via G_{D}. Note that the total surface ∂V enclosing the volume may be disjoint, as for instance with the volume between two concentric spheres.

If we have managed to find G_{D} for a particular type of boundary, the source-free solution ($F(\mathbf{x}') = 0$) is just the surface integral; but if it happens that $\Psi = 0$ on ∂V, only the volume integral contributes. Many boundary-value problems in electrostatics, for which the B.C. are reasonably simple, can be solved this way.

Similar considerations apply to Neumann B.C., ie. when $\partial\Psi/\partial n$ rather than Ψ is known on the boundary. But we must be a little careful about the B.C. on $\partial_n G_{\mathrm{N}}$: we cannot always put this equal to 0 in eq. (4.39). Indeed, take for instance $L = \nabla^2$; then, from the divergence theorem and the defining equation $L[G_{\mathrm{N}}] = \delta(\mathbf{x}-\mathbf{x}')$:

$$\int \boldsymbol{\nabla}\cdot\boldsymbol{\nabla} G_{\mathrm{N}}\, \mathrm{d}^3x \;=\; \oint_{\partial V} \partial_n G_{\mathrm{N}}\, \mathrm{d}S \;=\; 1$$

A consistent B.C. is $\partial_n G_{\mathrm{N}}\big|_{\partial V} = 1/S$, and the (non-unique) solution to the Neumann problem for $L = \nabla^2$ reads:

$$\Psi(\mathbf{x}) \;=< \Psi >_{\partial V} \;+\; \int_V F(\mathbf{x}')\, G_{\mathrm{N}}(\mathbf{x},\mathbf{x}')\, \mathrm{d}^3x' \;-\; \oint_{\partial V} G_{\mathrm{N}}(\mathbf{x},\mathbf{x}')\, \partial_{n'}\Psi(\mathbf{x}')\, \mathrm{d}S' \tag{4.41}$$

unique up to the a priori unknown average of Ψ over the surface. Often (but not always!) the volume is bounded by two surfaces, one closed and finite and the other at infinity, in which case $\partial_n G_{\mathrm{N}}$ can be set to zero on the entire boundary, and $<\ \Psi\ >_{\partial V}$ vanishes, still leaving an arbitrary additive constant in the solution. Also, $G_{\mathrm{N}}(\mathbf{x},\mathbf{x}')$ itself is determined only up to an additive function $h(\mathbf{x})$, with $L_{\mathbf{x}} h = 0$. When $L = \nabla^2$, however, condition (4.38) prevents h from contributing to the solution, and h can be used to make G_{N} symmetric in $\mathbf{x}$ and $\mathbf{x}'$.

[†] Although we call it "inhomogeneous", nothing in what we will do here prevents $F(\mathbf{x})$ from depending on $\Psi(\mathbf{x})$.

4.7.2 Green 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. 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 4.3), an "inhomogeneous" solution is:

   $$\Psi(\mathbf{x}) \;=\; -\frac{1}{(2\pi)^{3/2}} \int \frac{F(\mathbf{k})}{k^2 \,+\, \kappa^2}\, \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{x}}\, \mathrm{d}^3k \;=\; -\frac{1}{(2\pi)^3} \iint \mathrm{d}^3x'\, \mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{x}'}\, \frac{F(\mathbf{x}')}{k^2 \,+\, \kappa^2}\, \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{x}}\, \mathrm{d}^3k$$

   Compare with the Green-function form of the inhomogeneous solution, $\int_V F(\mathbf{x}')G(\mathbf{x},\mathbf{x}')\,\mathrm{d}^3x'$, and inegrate over the angles in k-space (EXERCISE):

   $$G(\mathbf{x},\mathbf{x}') \;=\; -\frac{1}{(2\pi)^3} \int \frac{\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}')}}{k^2 \,+\, \kappa^2}\, \mathrm{d}^3k \;=\; \frac{\mathrm{i}}{(2\pi)^2\,|\mathbf{x}-\mathbf{x}'|} \int_{-\infty}^{\infty} \frac{k\,\mathrm{e}^{\mathrm{i}k|\mathbf{x}-\mathbf{x}'|}}{k^2 \,+\, \kappa^2}\, \mathrm{d}k$$

   This last integral is easily evaluated as part of a contour integral around a semi-circle at infinity in the upper complex k half-plane. As in the one-dimension example, the contribution of the semi-circle 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:

   $$G(\mathbf{x},\mathbf{x}') \;=\; -\frac{1}{4\pi}\, \frac{\mathrm{e}^{-\kappa|\mathbf{x}-\mathbf{x}'|}}{|\mathbf{x}-\mathbf{x}'|} \tag{4.42}$$

   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 charge density $\rho(\mathbf{x})$ that is either localised, or at least vanishes at infinity faster than $|\mathbf{x}-\mathbf{x}'|^2$.

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:

   $$G_q^{(\pm)}(\mathbf{x},\mathbf{x}') \;=\; -\frac{1}{(2\pi)^3} \lim_{\epsilon\to 0} \int \frac{\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}')}}{k^2 \,-\, (q \pm \mathrm{i}\epsilon)^2}\, \mathrm{d}^3k \;=\; -\frac{1}{4\pi}\, \frac{\mathrm{e}^{\pm\mathrm{i}q|\mathbf{x}-\mathbf{x}'|}}{|\mathbf{x}-\mathbf{x}'|} \tag{4.43}$$

   For details of the calculation, see pp. BF415–416.

Do check that these Green functions satisfy $(\nabla^2 + \lambda)G(\mathbf{x},\mathbf{x}') = \delta(\mathbf{x}-\mathbf{x}')$. But 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:

$$\Psi_q^{(\pm)}(\mathbf{x}) \;=\; -\frac{1}{4\pi} \int_V \frac{F(\mathbf{x}')\,\mathrm{e}^{\pm\mathrm{i}q|\mathbf{x}-\mathbf{x}'|}}{|\mathbf{x}-\mathbf{x}'|}\, \mathrm{d}^3x' \tag{4.44}$$

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:

$$\Psi_q^{(\pm)}(\mathbf{x}) \;=\; A\,\mathrm{e}^{\mathrm{i}\mathbf{q}\cdot\mathbf{x}} \;-\; \frac{1}{4\pi} \int_V \frac{F(\mathbf{x}')\,\mathrm{e}^{\pm\mathrm{i}q|\mathbf{x}-\mathbf{x}'|}}{|\mathbf{x}-\mathbf{x}'|}\, \mathrm{d}^3x' \tag{4.45}$$

Note that these expressions for Green's functions assume no boundary surfaces (except at infinity)!

4.7.3 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. (4.42), 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 to force the combined Green function to vanish on the boundary. In the case of the Laplacian, we take:

$$G_\mathrm{D}(\mathbf{x},\mathbf{x}') \;=\; -\frac{1}{4\pi}\left(\frac{1}{|\mathbf{x}-\mathbf{x}'|} + \frac{g}{|\mathbf{x}-\mathbf{x}''|}\right)$$

which means that if the second term is to satisfy the Laplace equation $\forall\,\mathbf{x}$ *inside* the volume where we are looking for a solution, $\mathbf{x}''$ must lie *outside* the volume containing $\mathbf{x}$ and $\mathbf{x}'$.

Example 4.5. 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 of G_D dictates that $\mathbf{x}''$ and $\mathbf{x}'$ be collinear, which means that, at $|\mathbf{x}| = r = a$, we can write:

$$G_\mathrm{D}(a\hat{\mathbf{n}},\mathbf{x}') \;=\; -\frac{1}{4\pi}\left(\frac{1}{a\left|\hat{\mathbf{n}} - \frac{r'}{a}\hat{\mathbf{n}}'\right|} + \frac{g}{r''\left|\frac{a}{r''}\hat{\mathbf{n}} - \hat{\mathbf{n}}'\right|}\right)$$

where $r\hat{\mathbf{n}} = \mathbf{x}$, etc. By inspection, we see that if $G_\mathrm{D}(a\hat{\mathbf{n}},\mathbf{x}')$ is to vanish for $\hat{\mathbf{n}}$ in an arbitrary direction, we must have: $1/a = -g/r''$ and $r'/a = a/r''$. Then:

$$g \;=\; -a/r', \qquad\qquad r'\,r'' \;=\; a^2 \tag{4.46}$$

Thus, $\mathbf{x}''$ does lie *outside* the sphere if $\mathbf{x}'$ is inside, and *vice-versa*. Replacing $a\hat{\mathbf{n}}$ by $r\hat{\mathbf{n}} = \mathbf{x}$ yields:

$$\begin{aligned} G_\mathrm{D}(\mathbf{x},\mathbf{x}') \;&=\; -\frac{1}{4\pi}\left[\frac{1}{|\mathbf{x}-\mathbf{x}'|} - \frac{1}{\left|(r'/a)\mathbf{x} - (a/r')\mathbf{x}'\right|}\right] \\ &=\; -\frac{1}{4\pi}\left[\frac{1}{\sqrt{r^2+r'^2-2rr'\cos\gamma}} - \frac{1}{\sqrt{r^2r'^2/a^2+a^2-2rr'\cos\gamma}}\right] \end{aligned} \tag{4.47}$$

The second form makes is most easy to see that not only $G_\mathrm{D}(\mathbf{x},a\hat{\mathbf{n}}') = 0$, but also $G_\mathrm{D}(a\hat{\mathbf{n}},\mathbf{x}') = 0$, as desired. In spherical coordinates centered on the sphere, the angle γ between $\mathbf{x}$ amd $\mathbf{x}'$ is, from spherical trigonometry: $\cos\gamma = \cos\theta\,\cos\theta' + \sin\theta\,\sin\theta'\,\cos(\phi-\phi')$. The Dirichlet Green function we have found is valid for *any* ball since it does not care about which particular B.C. is specified for $\Psi(\mathbf{x})$ on its spherical boundary.

When $\Psi(r' = a) = 0$, the surface integral in eq. (4.40) vanishes; the volume integral remains the same since it is independent of the B.C. for Ψ. If $\Psi(r' = a) \neq 0$, we must evaluate $\partial_{n'}G_\mathrm{D}$ on the sphere. In spherical coordinates, this is:

$$\left.\frac{\partial G_\mathrm{D}}{\partial n'}\right|_{\partial V} \;=\; \pm\left.\frac{\partial G_\mathrm{D}}{\partial r'}\right|_{r'=a} \;=\; \pm\frac{1}{4\pi a^2}\,\frac{a(a^2-r^2)}{(r^2+a^2-2ar\cos\gamma)^{3/2}}$$

depending on whether $\mathrm{d}\mathbf{S}'$, the normal to the surface which always points *out* of the volume, is in the direction of $\mathbf{x}'$ or in the opposite direction. Then the general solution of the Poisson equation with B.C. specified on the surface $r = a$ for Ψ is:

$$\Psi(\mathbf{x}) \;=\; \frac{1}{4\pi}\int F(\mathbf{x}')\left[\frac{1}{\sqrt{r^2r'^2/a^2+a^2-2rr'\cos\gamma}} - \frac{1}{\sqrt{r^2+r'^2-2rr'\cos\gamma}}\right]\mathrm{d}^3x' \\ \pm\;\frac{1}{4\pi}\oint \Psi(r'=a)\,\frac{a^2-r^2}{a\,(r^2+a^2-2ar\cos\gamma)^{3/2}}\,\mathrm{d}S' \tag{4.48}$$

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, $\Psi\partial_{n'}G_{\mathrm{D}}$, 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 the Laplace equation $\nabla^2\Psi = 0$, with solution:

$$\Psi(\mathbf{x}) \;=\; \pm\oint \Psi(a,\theta',\phi')\left[\frac{1}{4\pi a^2}\,\frac{a(a^2-r^2)}{(r^2+a^2-2ar\,\cos\gamma)^{3/2}}\right]\mathrm{d}S' \tag{4.49}$$

Clearly also, if $\Psi(a,\theta',\phi') \neq 0$ and $F(\mathbf{x}) = 0$ for $r > a$, then $F(\mathbf{x}) \neq 0$ somewhere in the region $r < a$, and vice-versa. One can also show that for a ball of radius a and surface Ω_{n-1} in $\mathbb{R}^n$:

$$G_{\mathrm{D}}(\mathbf{x},\mathbf{x}') \;=\; \begin{cases} \frac{1}{2\pi}\left(\ln|\mathbf{x}-\mathbf{x}'| - \ln\left|\frac{r'}{a}\,\mathbf{x} - \frac{a}{r'}\mathbf{x}'\right|\right) & (n=2) \\ -\,\dfrac{1}{(n-2)\,\Omega_{n-1}}\left(\dfrac{1}{|\mathbf{x}-\mathbf{x}'|^{n-2}} - \dfrac{1}{|(r'/a)\mathbf{x}-(a/r')\mathbf{x}'|^{n-2}}\right) & (n>2) \end{cases} \tag{4.50}$$

which leads to a unified expression, valid for $n \geq 2$, for the normal derivative of G_{D} on the sphere:

$$\partial_{n'}G_{\mathrm{D}}\Big|_{r'=a} \;=\; \pm\,\frac{1}{\Omega_{n-1}}\,a^{n-2}\,\frac{a^2-r^2}{|\mathbf{x}-\mathbf{x}'|^n}\Bigg|_{r'=a} \tag{4.51}$$

4.7.4 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:

$$\begin{aligned}\nabla^2_{\mathbf{x}}G(\mathbf{x},\mathbf{x}') \;&=\; \delta(\mathbf{x}-\mathbf{x}') \\ &=\; \frac{1}{r^2}\delta(r-r')\sum_{l=0}^{\infty}\sum_{m=-l}^{l} Y^*_{lm}(\theta',\phi')\,Y_{lm}(\theta,\phi)\end{aligned} \tag{4.52}$$

where the completeness relation for spherical harmonics has been invoked:

$$\sum_{l=0}^{\infty}\sum_{m=-l}^{l} Y^*_{lm}(\theta',\phi')Y_{lm}(\theta,\phi) \;=\; \delta(x-x')\delta(\phi-\phi') \qquad\qquad (x=\cos\theta) \tag{4.53}$$

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. (4.52), we immediately find with eq. (4.25) that $g_l(r,r')$ must satisfy the radial equation:

$$r^2\,\nabla_r^2\,g_l(r,r') \;=\; \mathrm{d}_r\big[r^2\,\mathrm{d}_r\,g_l(r,r')\big] \;-\; l(l+1)\,g_l(r,r') \;=\delta(r-r')$$

We now find ourselves in the familiar territory of 1-dim Green-function problems for self-adjoint operators, and we can connect with eq. (4.13) for a 1-dim Dirichlet Green function. We have $\alpha(r')=r'^2$ and, with $f_1=r^l$ and $f_2=r^{-(l+1)}$, $W(r')=-(2l+1)/r'^2$, so that $\alpha W=-(2l+1)$.

At this point, we must specify the boundary, and we take two concentric spheres of radius a and b, with $b>a$. Then a straightforward computation using eq. (4.16) leads to (EXERCISE):

$$G_{\mathrm{D}}(\mathbf{x},\mathbf{x}') \;=\; \sum_{l=0}^{\infty}\sum_{m=-l}^{l}\frac{Y^*_{lm}(\theta',\phi')\,Y_{lm}(\theta,\phi)}{(2l+1)\big[1-(a/b)^{2l+1}\big]}\left(r_<^l\;-\;\frac{a^{2l+1}}{r_<^{l+1}}\right)\left(\frac{r_>^l}{b^{2l+1}}\;-\;\frac{1}{r_>^{l+1}}\right) \tag{4.54}$$

where, as before, $r_<=\min(r,r')$ and $r_>=\max(r,r')$.

Inspection of the radial factors confirms 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:

$$G_{\mathrm{D}}(\mathbf{x},\mathbf{x}') \;=\; \sum_{l=0}^{\infty}\sum_{m=-l}^{l}\frac{Y^*_{lm}(\theta',\phi')Y_{lm}(\theta,\phi)}{(2l+1)}\,r_<^l\left(\frac{r_>^l}{b^{2l+1}}\;-\;\frac{1}{r_>^{l+1}}\right)\qquad(a=0) \tag{4.55}$$

$$G_{\mathrm{D}}(\mathbf{x},\mathbf{x}') \;=\; \sum_{l=0}^{\infty}\sum_{m=-l}^{l}\frac{Y^*_{lm}(\theta',\phi')\,Y_{lm}(\theta,\phi)}{(2l+1)}\frac{1}{r_>^{l+1}}\left(\frac{a^{2l+1}}{r_<^{l+1}}\;-\;r_<^l\right)\qquad(b\to\infty) \tag{4.56}$$

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:

$$G(\mathbf{x},\mathbf{x}') \;=\; -\sum_{l=0}^{\infty}\sum_{m=-l}^{l}\frac{1}{2l+1}\,\frac{r_<^l}{r_>^{l+1}}\,Y^*_{lm}(\theta',\phi')\,Y_{lm}(\theta,\phi) \tag{4.57}$$

This 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 rewrite (EXERCISE) the surface integral in eq. (4.40) 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 immediately determines the A_{lm} coefficients in the general solution (4.31). It 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.

4.7.5 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:

$$(\nabla^2 + \lambda)\,\psi(\mathbf{x}) = \frac{2m}{\hbar^2}\,V(\mathbf{x})\,\psi(\mathbf{x}) \tag{4.58}$$

where $\lambda = 2mE/\hbar^2$. Although the right-hand side is not inhomogeneous, our previous results still hold.

For bound states ($E < 0$) of an attractive potential, $\lambda = -\kappa^2 < 0$, and we have the integral equation:

$$\psi(\mathbf{x}) = -\frac{m}{2\pi\hbar^2}\int \frac{\mathrm{e}^{-\kappa|\mathbf{x}-\mathbf{x}'|}}{|\mathbf{x}-\mathbf{x}'|}\,V(\mathbf{x}')\,\psi(\mathbf{x}')\,\mathrm{d}^3x'$$

For unbound states ($E > 0$), $\lambda > 0$, and eq. (4.45) immediately becomes one form of the **Lippmann-Schwinger equation**:

$$\psi^{(\pm)}_{\mathbf{q}}(\mathbf{x}) = \frac{A}{(2\pi)^{3/2}}\,\mathrm{e}^{\mathrm{i}\mathbf{q}\cdot\mathbf{x}} - \frac{m}{2\pi\hbar^2}\int \frac{\mathrm{e}^{\pm\mathrm{i}q|\mathbf{x}-\mathbf{x}'|}}{|\mathbf{x}-\mathbf{x}'|}\,V(\mathbf{x}')\,\psi^{(\pm)}_{q}(\mathbf{x}')\,\mathrm{d}^3x' \tag{4.59}$$

with $q = \sqrt{2mE/\hbar^2}$.

We can make contact with more usual forms by introducing the convolution $[V * \psi](\mathbf{q})$:

$$[V * \psi](\mathbf{q}) \coloneqq \frac{1}{(2\pi)^{3/2}}\int V(\mathbf{q}-\mathbf{q}')\psi(\mathbf{q}')\,\mathrm{d}^3q'$$

According to the convolution theorem, the Fourier transform of $[V * \psi](\mathbf{q})$ is just $V(\mathbf{x})\psi(\mathbf{x})$ in eq. (4.58). Then the Fourier representation of eq. (4.58) can be written as:

$$-\int (k^2-\kappa^2)\psi(\mathbf{k})\,\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{x}}\,\mathrm{d}^3k = \frac{2m}{\hbar^2}\int\left[\frac{1}{(2\pi)^{3/2}}\int V(\mathbf{q}-\mathbf{q}')\psi(\mathbf{q}')\,\mathrm{d}^3q'\right]\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{x}}\,\mathrm{d}^3k$$

and there comes the better-known expression in Fourier space:

$$\psi^{(\pm)}_{\mathbf{q}}(\mathbf{k}) = A\,\delta(\mathbf{q}-\mathbf{k}) - \frac{2m}{(2\pi)^{3/2}\hbar^2}\int \frac{V(\mathbf{q}-\mathbf{q}')\psi(\mathbf{q}')}{k^2-(q\pm\mathrm{i}\,\epsilon)^2}\,\mathrm{d}^3q'$$

The asymptotic form of eq. (4.59) 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:

$$\begin{aligned}\psi^{(\pm)}_{\mathbf{q}}(\mathbf{x}) &\underset{r\to\infty}{=} \frac{A}{(2\pi)^{3/2}}\,\mathrm{e}^{\mathrm{i}\mathbf{q}\cdot\mathbf{x}} - \frac{m}{2\pi\hbar^2}\frac{\mathrm{e}^{\pm\mathrm{i}qr}}{r}\int \mathrm{e}^{\mp\mathrm{i}q\hat{\mathbf{n}}\cdot\mathbf{x}'}\,V(\mathbf{x}')\,\psi^{(\pm)}_{\mathbf{q}}(\mathbf{x}')\,\mathrm{d}^3x' \\ &= \frac{A}{(2\pi)^{3/2}}\left[\mathrm{e}^{\mathrm{i}\mathbf{q}\cdot\mathbf{x}} + f_{\pm}(\mathbf{q})\,\frac{\mathrm{e}^{\pm\mathrm{i}qr}}{r}\right]\end{aligned}$$

This expression represents the spatial dependence of a superposition of a plane wave and a **scattered** spherical wave propagating inward or outward from the origin. The function $f_{\pm}(\mathbf{q})$ is called the **scattering amplitude**; it also obeys an integral equation in $\mathbf{q}$ (momentum) space, eq. BF7.75, and its square modulus is directly related to experimental data. See BF p. 414–420 for more details and an application to the Yukawa potential.

4.8 A Hyperbolic Problem: the d'Alembertian Operator

With the Fourier integral representation (note the different normalisation and sign in the exponentials!):

$$\begin{aligned}\Psi(\mathbf{x},t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\Psi(\mathbf{x},\omega)\,\mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{d}\omega \\ \Psi(\mathbf{x},\omega) &= \int_{-\infty}^{\infty}\Psi(\mathbf{x},t)\,\mathrm{e}^{\mathrm{i}\omega t}\,\mathrm{d}t\end{aligned} \tag{4.60}$$

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:

$$(\nabla^2 \;+\; k^2)\Psi(\mathbf{x},\omega) \;=\; F(\mathbf{x},\omega) \qquad\qquad k^2 \equiv (\omega/c)^2 \tag{4.61}$$

Just as for the Laplacian, there exist Green functions for $\nabla^2 + k^2$; we have found them earlier in eq. (4.43):

$$G^{(\pm)}(R) \;=\; -\frac{1}{4\pi}\,\frac{\mathrm{e}^{\pm \mathrm{i}kR}}{R} \qquad\qquad R \equiv |\mathbf{x}-\mathbf{x}'| \tag{4.62}$$

Now we are ready to derive the full Green functions for the d'Alembertian operator, which satisfy:

$$\Box_{\mathbf{x}} G(\mathbf{x},t;\mathbf{x}',t') \;=\; \delta(\mathbf{x}-\mathbf{x}')\,\delta(t-t') \tag{4.63}$$

With the important representation of the delta-function:

$$\delta(t-t') \;=\; \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}\,\omega(t-t')}\,\mathrm{d}\omega \tag{4.64}$$

the defining equation becomes, 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 yields: $G^{\pm}(\mathbf{x},\mathbf{x}',\omega,t') = -\mathrm{e}^{\mathrm{i}(\pm kR+\omega t')}/4\pi R$, after using (4.61). Then, transforming back to the time domain, and replacing k by ω/c, we arrive at the Green functions:

$$G^{(\pm)}(\mathbf{x},t;\mathbf{x}',t') \;=\; -\frac{1}{8\pi^2 R}\int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}\omega[\pm R/c+(t'-t)]}\,\mathrm{d}\omega \;=\; -\frac{1}{4\pi}\,\frac{\delta\big(t' \;-\; [t \mp R/c]\big)}{R} \tag{4.65}$$

Thus, in non-elliptic problems, Green functions can contain δ-functions and so may not be actual *functions*!

Using eq. (4.63), we also recognise that:

$$\Box_{\mathbf{x}} \int_{\substack{\text{all}\\ \text{space}}} \mathrm{d}^3x' \int_{-\infty}^{\infty} G^{(\pm)}(\mathbf{x},t;\mathbf{x}',t')\,F(\mathbf{x}',t')\,\mathrm{d}t' \;=\; \int \mathrm{d}^3x' \int_{-\infty}^{\infty} F(\mathbf{x}',t')\,\Box_{\mathbf{x}} G^{(\pm)}(\mathbf{x},t;\mathbf{x}',t')\mathrm{d}t' \;=\; F(\mathbf{x},t)$$

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** or **advanced** solutions:

$$\begin{aligned}\Psi_{\left\{\substack{\text{ret}\\ \text{adv}}\right\}}(\mathbf{x},t) \;&=\; \Psi_{\left\{\substack{\text{in}\\ \text{out}}\right\}}(\mathbf{x},t) \;+\; \int\int_{-\infty}^{\infty} G^{(\pm)}(\mathbf{x},t;\mathbf{x}',t')\,F(\mathbf{x}',t')\,\mathrm{d}^3x'\,\mathrm{d}t' \\ &=\; \Psi_{\left\{\substack{\text{in}\\ \text{out}}\right\}}(\mathbf{x},t) \;-\; \frac{1}{4\pi}\int \frac{F(\mathbf{x}',t\mp R/c)}{\left|\mathbf{x}-\mathbf{x}'_{\left\{\substack{\text{ret}\\ \text{adv}}\right\}}\right|}\,\mathrm{d}^3x'\end{aligned} \tag{4.66}$$

where in the integral the position $\mathbf{x}'$ must be evaluated at the **retarded time** $t - R/c$, or at the **advanced time** $t+R/c$. The retarded case ensures the proper causal behaviour of the solutions, in the sense that, eg., the solution at time t only depends on the behaviour of the source point $\mathbf{x}'$ at time $t - R/c$. Ψ_{in} and Ψ_{out} are possible plane-wave solutions of the *homogeneous* wave equation for Ψ. Often they can be dropped.

4.9 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 set 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 Cauchy problem can be solved, the dynamical behaviour of the set of variables can be uniquely predicted from its initial data.

Most often, the equations of "motion" take the form of a set of equations 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 derivative at t_0. When there are inherent, built-in **constraints** on the initial data, however, these constraint equations must be discovered and solved. Also, we must find which initial data we are allowed to specify freely. In **gauge theories**, such constraints mean that the time evolution of some functions is *arbitrary*, ie., their Cauchy problem cannot be solved. Thus, they cannot be proper, physical dynamical variables, or observables.

We study in some depth a very important example: Maxwell's theory. In linear, unpolarised and unmagnetised media, the component form of Maxwell's equations in Gaussian units is $\partial_\mu F^{\mu\nu} = -4\pi J^\mu/c$ and $\partial_\mu \star F^{\mu\nu} = 0$. In the so-called $3+1$ formalism, this translates to:

$$\begin{aligned} \boldsymbol{\nabla}\cdot\mathbf{E} &= 4\pi\,\rho & \boldsymbol{\nabla}\times\mathbf{B} - \partial_{ct}\mathbf{E} &= 4\pi\,\mathbf{J}/c \\ \boldsymbol{\nabla}\cdot\mathbf{B} &= 0 & \boldsymbol{\nabla}\times\mathbf{E} + \partial_{ct}\mathbf{B} &= 0 \end{aligned} \tag{4.67}$$

where c is the speed of light and $J^\mu = (\rho c, \mathbf{J})$. The source term satisfies a continuity equation: $\partial_\mu J^\mu = 0$, or $\partial_t \rho = \frac{1}{4\pi}\boldsymbol{\nabla}\cdot\,\partial_t\mathbf{E} = -\,\boldsymbol{\nabla}\cdot\mathbf{J}$.

The two homogeneous equations are equivalent to:

$$\mathbf{E} = -\,\partial_{ct}\mathbf{A} - \boldsymbol{\nabla}\Phi \qquad\qquad \mathbf{B} = \boldsymbol{\nabla}\times\mathbf{A} \tag{4.68}$$

If we perform the **gauge transformations** $\Phi \to \Phi - \partial_{ct} f$ and $\mathbf{A} \to \mathbf{A} + \boldsymbol{\nabla} f$, where $f(\mathbf{x}, t)$ is an arbitrary real differentiable function, neither $\mathbf{E}$ nor $\mathbf{B}$ change! We say that Maxwell's theory is **gauge-invariant**.

The inhomogeneous Maxwell equations (4.67) become *second-order* equations for Φ and $\mathbf{A}$:

$$\begin{aligned} \nabla^2\Phi + \partial_{ct}(\boldsymbol{\nabla}\cdot\mathbf{A}) &= -4\pi\,\rho \\ \Box\mathbf{A} - \boldsymbol{\nabla}\left(\boldsymbol{\nabla}\cdot\mathbf{A} + \partial_{ct}\Phi\right) &= -\,4\pi\,\mathbf{J}/c \end{aligned} \tag{4.69}$$

.

4.9.1 Second-order Cauchy problem using transverse/longitudinal projections

While eq. (4.69) 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. (4.69) since we can always perform an arbitrary gauge transformation on them at some arbitrary later time t, as often as we wish. This serious issue must be 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** (see section 1.6.2) 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:

$$\mathbf{A} = \underbrace{\boldsymbol{\nabla} u}_{\mathbf{A}_\mathrm{L}} + \underbrace{\boldsymbol{\nabla}\times\mathbf{w}}_{\mathbf{A}_\mathrm{T}}$$

$\mathbf{A}_\mathrm{L} = \boldsymbol{\nabla} u$, whose curl vanishes *identically*, is the **longitudinal** part (or projection) of $\mathbf{A}$; $\mathbf{A}_\mathrm{T} = \boldsymbol{\nabla}\times\mathbf{w}$, whose divergence vanishes *identically*, is the **transverse** projection of $\mathbf{A}$. This allows us to decompose Maxwell's equations for the fields and the potential into longitudinal and tranverse parts, which are perpendicular to each other.

Project the second equation (4.69). The transverse projection immediately gives:

$$\Box \mathbf{A}_{\rm T} \;=\; -4\pi\, \mathbf{J}_{\rm T}/c \tag{4.70}$$

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 (4.69):

$$\begin{aligned}\boldsymbol{\nabla}\cdot\Big[\Box\mathbf{A}_{\rm L} \;-\; \boldsymbol{\nabla}\left(\boldsymbol{\nabla}\cdot\mathbf{A}_{\rm L} \;+\; \partial_{ct}\Phi\right) \;+\; 4\pi\,\mathbf{J}_{\rm L}/c\Big] \;&=\; \Box(\boldsymbol{\nabla}\cdot\mathbf{A}_{\rm L}) \;-\; \nabla^2(\boldsymbol{\nabla}\cdot\mathbf{A}_{\rm L}) \;-\; \partial_{ct}\nabla^2\Phi \;+\; 4\pi\;\boldsymbol{\nabla}\cdot\mathbf{J}_{\rm L}/c \\ &=\; -\,\partial_{ct}\Big[\partial_{ct}(\boldsymbol{\nabla}\cdot\mathbf{A}_{\rm L}) \;+\; \nabla^2\Phi \;+\; 4\pi\,\rho\Big]\end{aligned}$$

where the continuity equation has been invoked in the second line. But the terms in the square bracket on that line are just the first of equations (4.69). Therefore, 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**.

4.9.2 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 (4.67), we arrive at:

$$\begin{aligned}\Box\mathbf{E} \;&=\; 4\pi\,\boldsymbol{\nabla}\rho \;+\; 4\pi\,\partial_{ct}\mathbf{J} \\ \Box\mathbf{B} \;&=\; -\,4\pi\,\boldsymbol{\nabla}\times\mathbf{J}/c\end{aligned} \tag{4.71}$$

These look like wave equations for six quantities. *But only those of their solutions which also satisfy the first-order field equations* (4.67)*, including at initial time* 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 transverse ones. 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 equations (4.67) which contain time derivatives. Suppose we specify $\mathbf{E}$ and $\partial_t\mathbf{E}$ at $t = t_0$, so as to solve the $2^{\rm nd}$-order equations, eq. (4.71). Then the two transverse components of $\mathbf{B}$ are determined by $\boldsymbol{\nabla}\times\mathbf{B} = 4\pi\mathbf{J}/c + \partial_{ct}\mathbf{E}$; $\partial_{ct}\mathbf{B}$ 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, the first-order equations take over and determine the others at $t = t_0$. Alternatively, we could have specified the two transverse components of $\mathbf{B}$ and their time derivatives at $t = t_0$ to constrain all the other field components and time derivatives.

You can also use (EXERCISE) the transverse/longitudinal projections of the first-order equations (4.67) 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, This Cauchy analysis does not rely on some particular solution, but is valid for *any* electromagnetic field and potential.

Since Maxwell's theory contains no information about $\boldsymbol{\nabla}\cdot\mathbf{A}$, this must be supplied by a so-called **gauge condition**. One that is frequently used is the *Lorenz condition*: $\boldsymbol{\nabla}\cdot\mathbf{A} = -\partial_{ct}\Phi$,. Inserting it into Maxwell's equation (4.69) for $\mathbf{A}$ 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. In Appendix M, moreover, we show that $\mathbf{A}_{\rm L}$ can be made to disappear without affecting Maxwell's equations for the fields *and* the potentials.

Appendices

J Solving an Inhomogeneous Equation in Terms of Homogeneous Solutions

Let $f_1$and f_2 be independent solutions to the homogeneous differential equation (4.5). We use them to derive a *particular* solution $f_{\rm inh}(t)$ to the inhomogeneous equation. The key step is to insert $f_{\rm inh}(t) = f_1(t)g(t)$ to obtain a first-order equation for $\dot g$: $\ddot g + \big(\mathrm{d}_t(\ln f_1^2) + \beta/\alpha\big)\dot g = F/\alpha f_1$. Then the general first-order solution (4.4), together with Abel's formula (4.6) and $W(x)/f_1^2 = \mathrm{d}_t(f_2/f_1)$, yields:

$$\begin{aligned}
\dot g(t) &= \mathrm{d}_t\left(\frac{f_2}{f_1}\right)\left(B + \int_a^t \frac{f_1(t')\,F(t')}{[\alpha\,W](t')}\,\mathrm{d}t'\right) \\
&= \mathrm{d}_t\left[\frac{f_2}{f_1}\left(B + \int_a^t \frac{f_1(t')\,F(t')}{[\alpha\,W](t')}\,\mathrm{d}t'\right)\right] - \frac{f_2}{f_1}\,\mathrm{d}_t\left(\int_a^t \frac{f_1(t')\,F(t')}{[\alpha\,W](t')}\,\mathrm{d}t'\right) \\
&= \mathrm{d}_t\left[\frac{f_2}{f_1}\left(B + \int_a^t \frac{f_1(t')\,F(t')}{[\alpha\,W](t')}\,\mathrm{d}t'\right)\right] - \frac{f_2(t)\,F(t)}{[\alpha\,W](t)}
\end{aligned}$$

where B is an arbitrary constant. A final integration leads to:

$$f_{\rm inh}(t) \;=\; f_1\,g \;=\; \int_a^t \left(\frac{f_1(t')\,f_2(t) \,-\, f_2(t')\,f_1(t)}{[\alpha\,W](t')}\right) F(t')\,\mathrm{d}t' \;+\; Af_1(t) \;+\; Bf_2(t) \tag{J.1}$$

Because we have not implemented homogeneous boundary conditions (B.C.) on this inhomogeneous solution, we must include the terms $Af_1 + Bf_2$, even though they look like belonging to the homogeneous solution.

We know that $f_{\rm inh}(t)$ must satisfy homogeneous (B.C.). We consider the two most important cases.

With one-point B.C. (IVP), $f_{\rm inh}(t)$ and its derivative must vanish at $t = a$. The integral term and its derivative are automatically zero at $t = a$. The other contribution also vanishes because the IVP has no non-zero homogeneous solution for homogeneous B.C.. In other words, the integral term satisfies the B.C. without any help from the adjustable constants A and B. Therefore, the inhomogeneous solution to an IVP Is:

$$f_{\rm inh}(t) \;=\; \int_a^t \left(\frac{f_1(t')\,f_2(t) \,-\, f_2(t')\,f_1(t)}{[\alpha\,W](t')}\right) F(t')\,\mathrm{d}t' \;=\; \int_a^\infty \theta(t-t')\,\left(\frac{f_1(t')\,f_2(t) \,-\, f_2(t')\,f_1(t)}{[\alpha\,W](t')}\right) F(t')\,\mathrm{d}t' \tag{J.2}$$

and it is the general solution when the boundary conditions on the general solution are homogeneous. When they are not, we must add to $f_{\rm inh}(t)$ the homogeneous solution with appropriate non-zero constants A and B. Of course, α should not vanish for $t > a$, and neither can the Wronskian, but the latter is guaranteed by our assumption that f_1 and f_2 are linearly independent. We conclude that a unique solution to the IVP always exists, provided that $\alpha \neq 0$ and that the source term, $F(t)$, is piecewise continuous for $t > a$.

We should check that our solution (J.2) satisfies the inhomogeneous equation. A surprise awaits us: because $L[f_1] = L[f_2] = 0$, the integrand does not contiribute to $L[f_{\rm inh}]$ for $a \le t' < t$; the sole contribution must come from the point $t' = t$. This suggests that the Direc delta-function must somehow be involved, and this is indeed what happens if we we use the second expression with the step-function, whose derivative is the delta-function.

The other case we wish to address is the Dirichlet problem, ie. the boundary-value problem (BVP) with f specified at the two end-points. While the integral term in eq. .(J.1) satisfies a homogeneous B.C. at $t = a$, it does not at the other end of the interval, at $t = b$. Enforcing $f_{\rm inh}(b) = 0$ requires adjusting the constants A and B. $f_{\rm inh}(a) = 0$ immediately leads to $B = -Af_1(a)/f_2(a)$. Then $f_{\rm inh}(b) = 0$ determines A, and we arrive at:

$$\begin{aligned}
f_{\rm inh}(t) \;=\; &\int_a^t \left(\frac{f_1(t')\,f_2(t) \,-\, f_2(t')\,f_1(t)}{\alpha(t')\,W(t')}\right) F(t')\,\mathrm{d}t' \\
&+ \frac{f_2(a)f_1(t) - f_1(a)f_2(t)}{f_2(a)f_1(b) - f_1(a)f_2(b)} \int_a^b \left(\frac{f_1(t')\,f_2(b) \,-\, f_2(t')\,f_1(b)}{[\alpha\,W](t')}\right) F(t')\,\mathrm{d}t'
\end{aligned}$$

This expression looks more symmetric if we combine the two integrals from a to t. Some tedious algebra yields:

$$f_{\rm inh}(t) \;=\; \frac{f_2(b)f_1(t)-f_1(b)f_2(t)}{f_1(a)f_2(b)-f_2(a)f_1(b)}\int_a^t\left(\frac{f_1(t')\,f_2(a)\;-f_2(t')\,f_1(a)}{[\alpha\,W](t')}\right)F(t')\,{\rm d}t' \\ +\;\frac{f_2(a)f_1(t)-f_1(a)f_2(t)}{f_1(a)f_2(b)-f_2(a)f_1(b)}\int_b^t\left(\frac{f_1(t')\,f_2(b)\;-f_2(t')\,f_1(b)}{[\alpha\,W](t')}\right)F(t')\,{\rm d}t'$$

which can be written in the compact form:

$$f_{\rm inh}(t) \;=\; \int_a^b\left[\frac{\big[f_2(b)f_1(t_>)-f_1(b)f_2(t_>)\big]\big[f_2(a)\,f_1(t_<)-f_1(a)\,f_2(t_<)\big]}{\alpha(t')\,W(t')\,\big[f_1(a)f_2(b)-f_2(a)f_1(b)\big]}\right]F(t')\,{\rm d}t' \tag{J.3}$$

where $t_> \coloneqq \max(t,t')$ and $t_< \coloneqq \min(t,t')$. The existence of the inhomogeneous solution depends on the denominator of the integrand not vanishing, as well as piecewise continuity of $F(t)$.

As with the IVP, checking the validity of this solution by calculating $L[f_{\rm inh}]$ reveals the same behaviour: only the point $t'=t$ contributes. The presence of the delta-function is easier to see if we write the expression in terms of step-functions which split the integral over two intervals.

The Green-function formalism introduced in the main body of these notes will shed more light on these results.

K Solution of a Homogeneous IVP with Homogeneous B.C.

We show that the only eigenfunction of the operator $L=\mathrm{d}_t^2+\beta(t)\mathrm{d}_t+\gamma(t)$ with eigenvalue zero, that satisfies homogeneous one-point boundary conditions, is the zero function. The relevant eigenvalue problem is the homogeneous equation: $L[f(t)]=0$ over $[a,\,b]$, with $f(a)=\dot f\big|_a=0$. The proof is adapted from section 13.3 in Hassani's textbook. The strategy relies on the following, easy to show, fact:

If there exists a constant, $c>0$, and a differentiable function, $h(t)$, such that $\dot h(t)\le ch(t),\ \forall\, t\in[a,\,b]$, then: $h(t)\le h(a)\mathrm{e}^{c(t-a)}$.

Indeed, starting from $\dot h\le ch$ and recalling that an exponential is never negative, one has: $\dot h\,\mathrm{e}^{-ct}\le ch\,\mathrm{e}^{-ct}$, or: $\mathrm{d}_t(h\,\mathrm{e}^{-ct})\le 0$. Integrating yields: $h(t)\,\mathrm{e}^{-ct}-h(a)\,\mathrm{e}^{-ca}\le 0$, which is the result sought.

Therefore, if we can find a function h of f and $\dot f$ that satisfies $\dot h\le ch$ for some positive constant c and that vanishes at $t=a$ by virtue of the initial conditions on f, then it will be bounded from above by zero; and if this same function happens to be a linear combination of *positive* functions of f and $\dot f$, we can conclude that the function is exactly zero, and so therefore are f and $\dot f$.

We make a prescient choice that does satisfy $h(a)=0$: $h=f^2+\dot f^2\ge 0$, with appropriate constants (dropped here to lower clutter) to make units consistent if t is not dimensionless.

Now comes the somewhat fiddly part. First, note that $(f\pm\dot f)^2=f^2+\dot f^2\pm 2f\dot f$, so that $|2f\dot f|\le f^2+\dot f^2$. Now differentiate h, and use the homogeneous equation to eliminate $\ddot f$, plus the properties of absolute values:

$$|\dot h|\;\le\;(1+|\gamma|)\,\big|2f\dot f\big|\;+\;2|\beta|\dot f^2\;\le\;(1+|\gamma|)\,(f^2\;+\;\dot f^2)+2|\beta|\dot f^2\;=\;(1+|\gamma|)\,f^2\;+\;\big[1+|\gamma|\;+\;2|\beta|\big]\dot f^2$$

Since the coefficient of f^2 on the right-hand side is always smaller than the coefficient of $\dot f^2$, we can take the maximum of the latter over $[a,\,b]$ as our constant c, and we have proved that our choice $h=f^2+\dot f^2\ge 0$ satisfies $\dot h\le|\dot h|\le c\,h$, which with $h(a)=0$ is the condition to have $h\le h(a)\mathrm{e}^{c(t-a)}=0$. The only possibility is then $h=0$, and thus $f=\dot f=0$ over the interval.

A shorter, niftier argument relies on expressing the general homogeneous solution in a basis $\{f_1,\,f_2\}$: $f_{\rm h}=Af_1+Bf_2$. Differentiating and setting $f(a)=\dot f\big|_a=0$ yields the matrix equation:

$$\begin{pmatrix} f_1(a) & f_2(a)\\ \dot f_1\big|_a & \dot f_2\big|_a\end{pmatrix}\begin{pmatrix}A\\B\end{pmatrix}\;=\;\begin{pmatrix}0\\0\end{pmatrix}$$

For at least one of A or B to be non-zero, the determinant of the matrix, ie., the Wronskian of f_1 and f_2, must vanish at $t=a$. But then, these functions being themselves solutions of the homogeneous equation, their Wronskian vanishes everywhere in $[a,\,b]$ if it vanishes anywhere because of Abel's formula (4.6), and the functions are linearly dependent, a contradiction. we conclude that A and B, and therefore $f_{\rm h}$, are zero.

L Modified Green Functions for the One-dim Boundary-value Problem

In sections 4.1.1 and 4.1.2 we saw that the one-dim BVP has no solution in terms of Green functions unless no homogeneous solution $f_\mathrm{h} \neq 0$ exists that satisfies the homogeneous boundary conditions (B.C.) at $x = a$ and b. There is an escape clause, however.

For simplicity's sake, let us assume that there is only one eigenfunction $\phi_0(x)$ corresponding to eigenvalue $\lambda_0 = 0$ that obeys $[L\phi_0](x) = 0$ with homogeneous B.C. If a solution f to $L[f] = F$ exists, it can be expanded over the complete set of orthonormal eigenfunctions of (self-adjoint!) L: $f(x) = \sum_{j\neq 0} a^j \phi_j(x)$. Then $[Lf](x) = \sum_{j\neq 0} a^j \lambda_j \phi_j(x) = F(x)$. Orthogonality of ϕ_0 with the other ϕ_j immediately gives:

$$\int_a^b \phi_0^*(x)\, F(x)\, \mathrm{d}x \;=\; 0 \tag{L.1}$$

Thus, a solution exists only if the driving term is itself orthogonal to ϕ_0 over the interval. To discover what form that solution takes, consider the **modified Green function**:

$$\mathcal{G}(x,x') \;\coloneqq\; \sum_{j\neq 0} \frac{\phi_j(x)\,\phi_j^*(x')}{\lambda_j} \qquad\qquad \lambda_j \neq 0 \tag{L.2}$$

While superficially identical to eq. (4.12), this expression specifically omits the *now non-zero* $\phi_0(x)\,\phi_0^*(x')$ term which simply did not exist in eq. (4.12).

Not surprisingly, and although they satisfy the same homogeneous B.C., $\mathcal{G}(x,x')$ and $G(x,x')$ solve different defining equations: $[L\mathcal{G}](x,x') = \sum_{\mathrm{all}\,j} \phi_j(x)\,\phi_j^*(x') - \phi_0(x)\,\phi_0^*(x')$, that is:

$$\big[L\mathcal{G}\big](x,x') \;=\; \delta(x-x') \;-\; \phi_0(x)\,\phi_0^*(x') \tag{L.3}$$

because $\{\phi_j\}$ with ϕ_0 included is a complete set. Then one quickly shows that if condition (L.1) holds, the form:

$$f(x) \;=\; C\,\phi_0(x) \;+\; \int_a^b \mathcal{G}(x,x')\,F(x')\,\mathrm{d}x' \tag{L.4}$$

is the solution to $[Lf[(x) = F(x)$. But since C is an arbitrary constant, we have lost unicity—in fact the number of solutions is infinite. Of course, if f must satisfy non-homogeneous B.C., we must also add the homogeneous solution that satisfies them.

When solving eq. (L.3) to find a modified Green function, we can proceed as in sections 4.2.2 and 4.3. Adding the particular solution of $[L\mathcal{G}](x,x') = -\phi_0(x)\,\phi_0^*(x')$ to its homogeneous solution does not change the conditions on $\mathcal{G}(x,x')$ at $x = x'$, but that particular solution must be added to (4.13), which will alter the b_1 and b_2 coefficients calculated by imposing the relevant homogeneous B.C. on $\mathcal{G}$. Unfortunately, without an explicit ϕ_0, it becomes impossible to write a general result for the modified Green function.

> **Example L.1.** The one-dim Laplace equation, $\mathrm{d}_x^2 f(x) = 0$, has for general solution $f_\mathrm{h}(x) = Ax + B$, with A and B constants. The Dirichlet B.C., $f_\mathrm{h}(a) = f_\mathrm{h}(b) = 0$, lead to $f_\mathrm{h}(x) = 0$ everywhere, and the Green function always exists. The homogeneous Neumann B.C., $\mathrm{d}_x f_\mathrm{h}\big|_a = \mathrm{d}_x f_\mathrm{h}\big|_b = 0$, however, do not determine B, and the homogeneous equation is solved by the non-trivial $\phi_0(x) = B$. Integrating eq. (L.3) once, the same Neumann B.C. on the modified Green function $\mathcal{G}_\mathrm{N}$ are consistent only if $B = 1/\sqrt{L}$, with $L \coloneqq b - a$. The equation $\mathrm{d}_x^2 \mathcal{G}_\mathrm{N} = -1/L$ for $x \neq x'$ is then solved by: $\mathcal{G}_\mathrm{N}(x,x') = -x^2/2L + b_1(x')x + b_2(x')$. Implementing the homogeneous Neumann B.C. leads to: $\mathcal{G}_\mathrm{N}(x,x') = -x^2/2L + ax/L + \theta(x-x')(x-x') + b_2(x')$, with $b_2(x')$ arbitrary.
>
> Insert this into eq. (4.21), and interchange x and x' to obtain $f(x)$; then choose $b_2(x) = -x^2/2L + ax/L$, which makes $\mathcal{G}_\mathrm{N}(x,x')$ symmetric, but does not affect $f(x)$ (why?). Implementing the symmetry, there comes the (unique up to a constant C) Neumann solution:
>
> $$f(x) \;=\; C \;+\; \int_a^b \mathcal{G}_\mathrm{N}(x,x')\,F(x')\,\mathrm{d}x' \;-\; \Big[\mathcal{G}_\mathrm{N}\,\mathrm{d}_{x'} f\Big]_{x'=a}^{x'=b} \qquad\qquad \partial\mathcal{G}_\mathrm{N}(x,a) \;=\; \partial\mathcal{G}_\mathrm{N}(x,b) \;=\; 0 \tag{L.5}$$

Provided the source in $\mathrm{d}_x^2 f = F$ integrates to zero over $[a, b]$, as required by eq. (L.1), a general solution to this Neumann problem (with homogeneous B.C.!) is given by eq. (L.5), but it is unique only up to an arbitrary constant.

M Counting Electromagnetic Degrees of Freedom in the Lorenz Gauge

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 $-\partial_{ct}\Phi$ to obtain:

$$\boldsymbol{\nabla}\cdot(\mathbf{A} + \boldsymbol{\nabla} f) = -\partial_{ct}(\Phi - \partial_{ct} f) \tag{M.1}$$

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:

$$\boldsymbol{\nabla}\cdot\mathbf{A} = \partial_x A_x = \mathrm{i}kA_{0x}\mathrm{e}^{\mathrm{i}(kx-\omega t)}, \qquad \partial_{ct}\Phi = -\mathrm{i}\,\frac{\omega}{c}\,\Phi_0 \mathrm{e}^{\mathrm{i}(kx-\omega t)}$$

Inserting into the Lorenz condition with $\omega/c = k$ yields, as expected, a relation between the longitudinal component A_x and Φ: $A_{0x} = \Phi_0$.

Now fold in $f = f_0 \mathrm{e}^{\mathrm{i}(kx-\omega t)}$ into eq. (M.1) for the gauge-transformed potentials, to get:

$$\mathrm{i}k\,(A_{0x} + \mathrm{i}k\,f_0)\mathrm{e}^{\mathrm{i}(kx-\omega t)} = \mathrm{i}\,\frac{\omega}{c}\,(\Phi_0 + \mathrm{i}\,\frac{\omega}{c}\,f_0)\,\mathrm{e}^{\mathrm{i}(kx-\omega t)}$$

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 only the transverse components of $\mathbf{A}$!

Tte conclusion is the same as that of the analysis of the field equations: only the two transverse components of $\mathbf{A}$ propagate, in the sense that they carry energy to infinity.